Cell. Mol. Life Sci. 64 (2007) 386–400
1420-682X/07/040386-15
DOI 10.1007/s00018-006-6295-0

Cellular and Molecular Life Sciences

Review

The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome

N. D. Heintzman [b] and B. Ren [a, *]

[a] Ludwig Institute for Cancer Research, Department of Cellular and Molecular Medicine, University of California, San Diego School of Medicine, 9500 Gilman Drive, La Jolla, California 92093 (USA), e-mail: biren@ucsd.edu
[b] Ludwig Institute for Cancer Research, Department of Cellular and Molecular Medicine, University of California, Biomedical Sciences Graduate Program, 9500 Gilman Drive, La Jolla, California 92093 (USA)

Received 27 June 2006; received after revision 24 August 2006; accepted 7 November 2006
Online First 14 December 2006

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

Keywords. Promoter, transcriptional regulation, ChIP-chip, epigenetics, regulatory networks.

Introduction

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

Eukaryotic promoter structure and regulation of expression for protein-coding genes have been extensively reviewed elsewhere [1, 2, 6], so we will briefly define and summarize key features and events involved in regulating the initiation of transcription (see Fig. 1). A eukaryotic promoter is located at the 5′-end of its transcribed sequence and serves as the point of transcriptional initiation. Typically, the term ‘promoter’ refers to the ‘core promoter’ and its adjacent sequences. The core promoter immediately surrounds the transcription start site (TSS) and comprises 70–80 bp that contain canonical sequence features (described below) sufficient for recognition by

* Corresponding author.

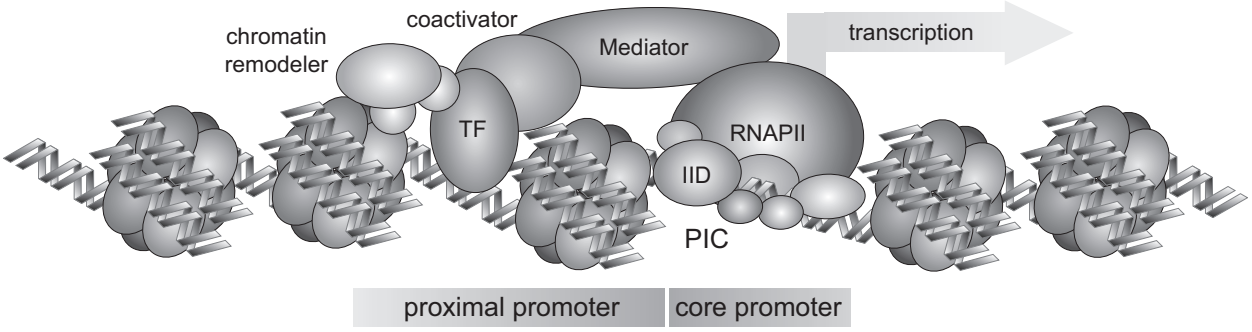

Figure 1. Typical structure of an active eukaryotic promoter. The promoter consists of a core promoter region immediately surrounding the transcription start site, adjacent to a more extended proximal promoter region. RNA polymerase II (RNAPII) and various general transcription factors (for example, transcription factor IID) form the pre-initiation complex (PIC) around the transcriptional start site. Other transcriptional regulatory proteins, including Mediator, chromatin remodelers, coactivators and sequence-specific transcription factors (TF), are involved in regulating transcription at the promoter. All of these events occur in the context of chromatin, made up of DNA wrapped around octamers of histone proteins.

the basal transcriptional machinery and initiation of transcription. The 'proximal promoter' includes the region extending upstream of the core promoter (generally ~250 bp from the TSS [7], though this limit can be somewhat subjective). Proximal promoters contain other sequence features critical to transcriptional regulation, for instance binding sites for tissue-specific transcription factors, and may in fact encompass transcriptional enhancers (which by their nature impart additional regulatory specificity to expression of the target gene). But due to their close proximity to the core promoter and our evolving understanding of promoter structure, we will refer to these regions collectively as the promoter unless otherwise noted.

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

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

ing of transcriptional regulation at promoters, in light of recent genome-scale investigations.

Large-scale promoter discovery in eukaryotic genomes

Identification of transcriptional promoters throughout the genome is critical to increasing our understanding of their contributions to gene regulation. Because much can be learned from comparing multiple examples of these regulatory elements, efforts to curate our knowledge of promoter regions have been ongoing for over 20 years [14]. Rapidly improving high-throughput and bioinformatics approaches have accelerated the discovery and location of promoters and have enhanced the quality of characterization and annotation of these elements, but the goal has remained the same: to understand the mechanisms by which promoters regulate transcription. Numerous resources and techniques now contribute to the large-scale study and analysis of an ever-expanding library of eukaryotic promoters.

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

In addition to transcript-based promoter identification, the maturation of technologies like ChIP (chromatin immunoprecipitation)-chip [24] has allowed the biochemical determination of promoters based on the protein-DNA interactions between the transcriptional machinery and the promoter sequences (see Fig. 2). By examining genome-wide binding patterns of components of the PIC in human fibroblast cells, one study located over 10 000 active promoter sites and almost 1200 novel promoters for previously unannotated transcriptional loci [25]. In addition to promoters for protein-coding genes, some of these novel promoters correspond to microRNA genes, whose transcripts were not amenable for identification by conventional cDNA cloning methods [26]. Therefore, the ChIP-chip approach complements the cDNA library-based method for promoter mapping.

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

Signatures of promoters

The diverse approaches to promoter identification described above have provided unprecedented resources for large-scale promoter characterization. Recent advances in high-throughput experimental methods and computational analysis strategies have provided significant insight into the physical and functional features of promoters. One goal of such investigations is to define the ‘signature’ of a promoter, that is, the sequence elements and chromatin features that dictate the promoter’s regulatory properties (see Fig. 3).

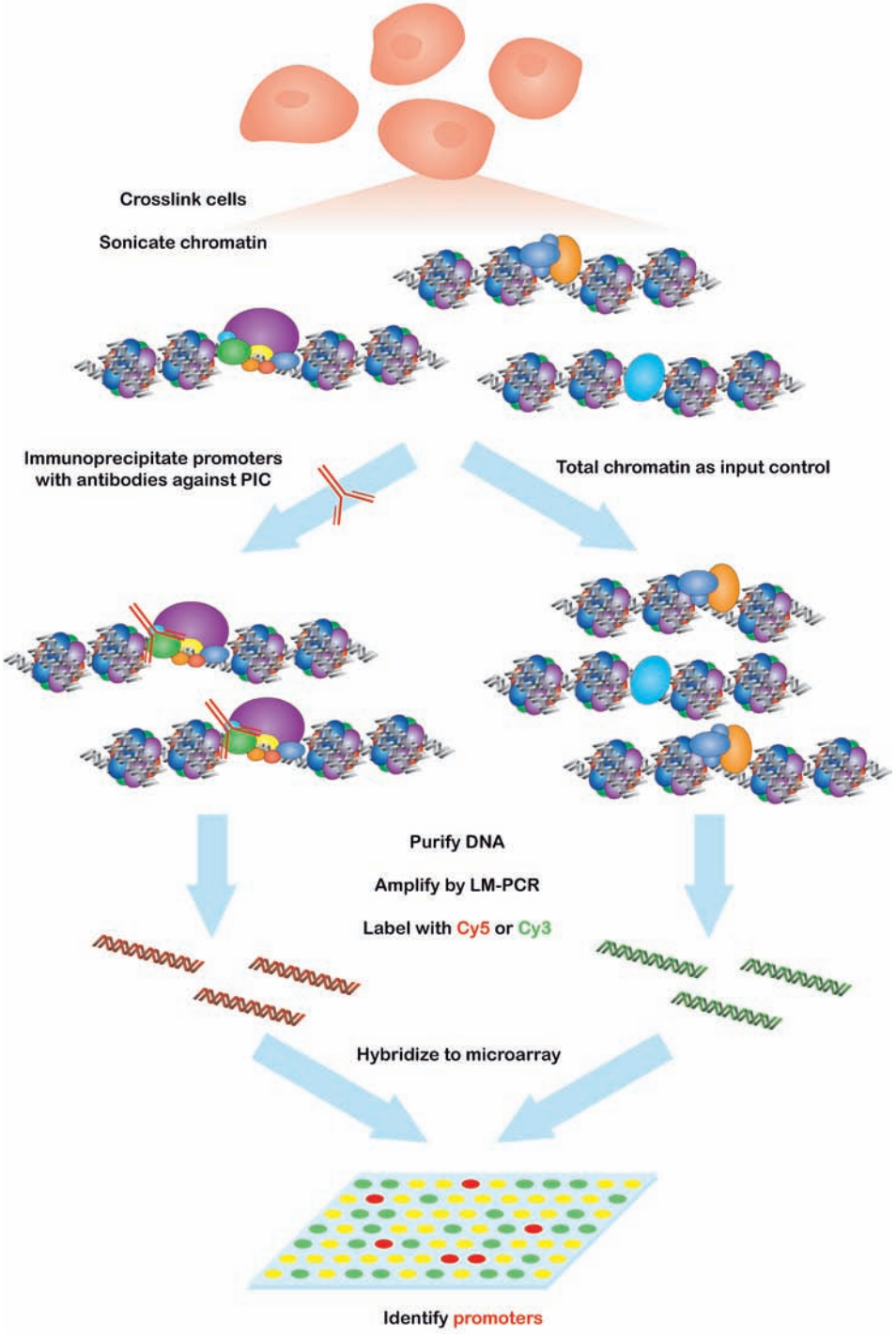

Figure 2. Promoter discovery using ChIP (chromatin immunoprecipitation)-chip. Cells are treated with formaldehyde to chemically cross-link DNA and interacting proteins. Chromatin is isolated and sheared to small pieces by sonication, then subjected to immunoprecipitation with antibodies specific to components of PIC. Promoter fragments bound by PIC will be enriched in the IP sample relative to a total chromatin control sample. DNA from both samples is purified, amplified and labeled with fluorescent dye, then hybridized to a microarray covering large continuous stretches of the human genome. Promoters are identified on the basis of their enrichment in the IP sample, visualized as a red spot on the microarray. LM-PCR, ligation-mediated polymerase chain reaction.

Sequence signatures

As noted above, the promoter consists of a core region immediately surrounding the TSS, and additional proximal promoter regions extending further upstream of the core promoter. Because the core promoter is the minimum region required for docking of the transcriptional machinery and initiation of basal transcription, extensive research in a variety of organisms has been devoted to uncovering

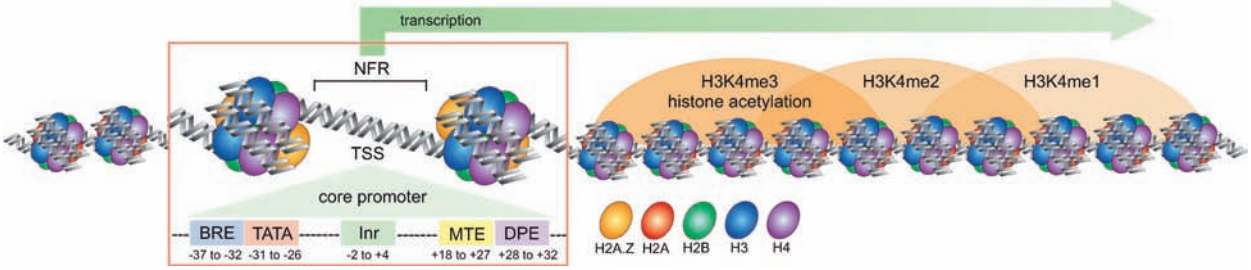

Figure 3. Signatures of active promoters. A nucleosome free region (NFR) surrounds the transcriptional start site (TSS) in the core promoter, which may contain core promoter elements, including BRE, TATA, Inr, MTE, DPE and others (positions are relative to the +1 TSS within the Inr; please see detailed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

the sequence motifs responsible for this critical step in gene regulation, revealing a collection of short regulatory DNA sequence elements conserved across species. While the first core promoter element has been known for almost 30 years, additional novel sequence elements have been discovered recently, emphasizing the importance of continued research of these regulatory sequences. Most of the canonical core promoter elements have been thoroughly reviewed elsewhere [2], but it is useful to describe their general features here (see Table 1) in light of recent genome-wide analyses of these elements. Note that there are no 'universal' core promoter elements; the sequences described below are found in only a subset of promoters, and the origins and functional consequences of the resulting core promoter diversity are a topic of current study.

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

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

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

The TFIIB recognition element (BRE; SSRCGCC) consists of the 7 bp immediately upstream of the TATA-box, and as its name suggests, it is bound by transcription factor IIB [44]. The BRE has been shown to both stimulate and repress transcriptional activity [45].

The motif ten element (MTE; CSARCSSAACGS) was identified in a computational survey of *Drosophila* promoters [46], located +18 to +29 downstream of the TSS and overlapping slightly with the 5′-end of the DPE. The MTE requires Inr and functions synergistically with the

Table 1. Summary of sequence and frequency of core promoter elements.

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

* The TSS is assigned to position +1.
** Degenerate nucleotides represented using IUPAC codes.

TATA-box or DPE, but can also function in a TATA- and DPE-independent manner and can compensate for mutations in either of these other elements [47]. It appears that the MTE contributes to interaction with TFIID.

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

Several recent large-scale analyses have confirmed the lack of universal core promoter elements, demonstrating that each element is found in subsets of promoters, with differing relative representation among species (see Table 1). For example, the TATA-box was once presumed to be a general feature of promoters, but genomic analyses clearly indicate that its presence is variable between species and actually atypical. A consensus TATA-box is present in only 33–43% of promoters in *Drosophila* [40, 46], and in only about 10–16% of mouse and human promoters [25, 52–54]. Furthermore, while 69% of *Drosophila* promoters contain the Inr [40, 46], only about 55% of human promoters possess this element [25]. In contrast, the DPE appears to be more abundant in human promoters (about 48%) [53] than in *Drosophila* (about 40%) [40]. CpG islands seem to be the most highly represented class of promoter element, with recent estimates of 79–88% of human promoters and 71% of mouse promoters [25, 53] containing this feature, much higher than earlier estimates of about half of promoters [55]. Of course, these elements may be present in various combinations. For example, in *Drosophila* the TATA-box and DPE occur together in 14% of promoters [40], and 12% of TATA-containing vertebrate promoters also contain a BRE [44]. Further research is necessary to exhaustively catalog these and other core promoter elements and sequence variants throughout entire genomes. But the variety in core promoter structure within and between species suggests a significant role for core promoter diversity in transcriptional regulation, contrary to early single-gene studies that implied a universal promoter sequence.

Epigenetic signatures

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

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

as revealed in recent genomic investigations in multiple organisms.

One key component of chromatin is, in fact, absent from active promoters: the nucleosome. Different experimental approaches in yeast have demonstrated depletion of nucleosomes at transcriptionally active promoters. ChIP-chip studies examining the enrichment patterns of core histones revealed a markedly reduced density of these proteins at the promoters of active genes genome-wide [58–61], indicating nucleosome depletion at these sites. High-resolution nucleosome mapping in yeast confirmed this observation, revealing a nucleosome-free region (NFR) of ~150 bp in size located ~200 bp upstream of the start codon [62]. The nucleosomes flanking this NFR contain the histone variant H2A.Z [63–65], implicating H2A.Z in NFR formation or maintenance, though differences in experimental techniques make it unclear how H2A.Z enrichment relates to transcriptional activity. The significant structural differences between normal H2A and H2A.Z provide distinct protein interaction domains unique to this variant; these features may contribute to a role for H2A.Z in antagonizing gene silencing [66]. Interestingly, a short DNA sequence element was demonstrated to be responsible for NFR formation [64], consistent with the observation of sequence-dependent DNA-histone interactions in yeast promoter regions [61]; these findings further emphasize the connection between DNA sequence and chromatin structure. Similar patterns of nucleosome depletion at active promoters were observed in *Drosophila* [67] and humans [N. Heintzman and B. Ren, unpublished data], contrary to an earlier study in mammalian cells [68] that found no change in nucleosome density at promoters. These recent findings are consistent with numerous reports demonstrating increased chromatin accessibility (as assayed by nuclease sensitivity) at promoters and other regulatory elements [69], and indicate that nucleosome depletion is an evolutionarily conserved mechanism of transcriptional regulation. An additional histone variant, H3.3, was found to be enriched at active promoters in *Drosophila* [67], further emphasizing the intimate relationship between nucleosome composition and transcriptional regulation. While the structure of H3.3 (and other H3 variants) is quite similar to that of normal H3, the recent 'H3 barcode hypothesis' [70] proposes that subtle changes in nucleosome stability resulting from incorporation of H3 variants can influence protein interaction, nuclear localization and post-translational modification, with profound impacts on gene regulation, epigenetic memory and chromatin structuring.

The discovery that the histone proteins within nucleosomes could be covalently modified led to the proposal of the histone code hypothesis [71], wherein distinct functional and regulatory information is encoded in patterns of histone acetylation and methylation, among other possible modifications [5, 56]. The field of epigenetics has exploded in recent years, and it would be impossible to thoroughly cover it in this review, so we will focus on relevant global studies of histone modifications associated with gene activation (using current nomenclature [72]). As genomics technology has rapidly evolved over the past few years, so has the coverage, resolution and specificity of the data gained from genome-wide epigenetic analyses. We will primarily discuss the most comprehensive current findings, acknowledging that they often confirm the results of many previous smaller-scale experiments. New and unexpected insights into promoter epigenetics have also been gained by the genome-wide expansion of previous single-gene findings.

Histone acetylation has long been found associated with active genetic regions, and many lysine residues within the various histone tails are subject to this modification [73]. Acetylation of histone lysines is a reversible modification controlled by two antagonistic protein families, the histone acetyltransferases (HATs) and histone deacetylases (HDACs). A genome-wide, high-resolution (~266 bp) assessment of histone acetylation in yeast revealed that acetylation of histone H3 lysines 9 (H3K9ac) and l4 (H3K14ac) and general acetylation of histone H4 (H4ac) are localized predominantly to promoters in a manner associated with transcriptional activity [60]. These modifications peak slightly downstream of the TSS. Similar ChIP-chip experiments in fly [74] and mammalian systems [25, 68, 75] demonstrated that acetylation of H3 and H4 is a conserved feature of transcriptionally active promoters. It is worth noting, however, that H3ac and H4ac have also been associated with some distal regulatory elements such as enhancers [75] [N. Heintzman and B. Ren, unpublished data]. A single-nucleosome resolution study of residue-specific histone acetylation patterns in 500 kbp of the yeast genome offered additional insight, including the observation that specific lysines (H2AK7, H3K9, H3K14, H3K18, H4K5, H4K12) are hyperacetylated on nucleosomes at the 5′-ends of active genes, adjacent to a hypoacetylated region surrounding the active promoter; intriguingly, principal component analysis revealed that the 12 histone modifications examined actually sort into two main classes (either promoter proximal or as a continuum through coding regions), rather than exhibiting independent distribution patterns [76]. Another study using histone-lysine mutants combined with global expression analysis suggested significant functional redundancy of residue-specific acetylation in histone H4, as only mutation of H4K16 caused specific changes in gene expression patterns [77]. These findings challenge the original hypothesis of a histone code with great combinatorial complexity conferred by distinct modifications, suggesting instead a simpler system in which multiple modifications play redundant roles in gene regulation, similar to the signaling network model of chromatin [78].

As with acetylation, histone lysines can be modified by methylation. Histone methylation seems more complex, however, as distinct histone methyltransferases (HMTs) can modify lysine residues by the addition of one, two or three methyl groups, each of which appear to have distinct localization patterns and regulatory potential. Further, methylation is associated with both activation and repression of transcription, depending on the modified residue. Though lysine methylation was long thought to be irreversible, the recent discovery of histone demethylases [79] suggests that this modification may be as dynamic as acetylation. The aforementioned studies in yeast [60, 76] revealed a gradient of methylation of H3K4 from 5′ to 3′ within actively transcribed genes, with trimethylation of this residue (H3K4me3) peaking at the 5′-end of the gene and giving way to di- and then monomethylation (H3K4me2, H3K4me1) with increasing distance from the promoter. Like acetylation, these methylation patterns correlate with transcriptional activity, a relationship also generally observed in investigations of H3K4me3 and H3K4me2 in fly [74] and mammalian systems [25, 68]. Another recent high-resolution study confirmed the H3K4me3-me2-me1 gradient at active human promoters [N. Heintzman and B. Ren, unpublished data]. H3K4me3 appears to mark active promoters exclusively, while H3K4me2 and H3K4me1 are also found elsewhere in the genome at other putative regulatory elements [68] [N. Heintzman and B. Ren, unpublished data].

The chromatin features of inactive promoters are less well characterized, but the above studies demonstrated that inactive promoters generally lack the histone modifications associated with promoter activity, including acetylation of H3 and H4 and methylation of H3K4. Trimethylation of H3K27 appears to be localized to promoters of repressed genes genome-wide [80, 81]. Also, repressed genes are frequently located in heterochromatin [82], where the condensed structure ostensibly prevents transcription factor access to regulatory DNA sequences, though some characteristic features of open, active chromatin have been noted at inactive promoters in yeast [61, 63, 64].

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

Promoter function and regulation

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

Regulatory networks

Sequence-specific transcription factors (TFs) play a critical role in regulating transcription by recruiting coactivators and promoting the formation of the PIC [9, 83]. Consequently, many investigations have focused on the discovery of direct targets of TF binding. TF consensus binding motifs are often somewhat degenerate, causing sequence-based computational methods to predict many thousands of binding sites for a given TF, only a fraction of which may be biologically relevant. Indeed, even binding sites for which the cognate TF has a very high affinity *in vitro* are not necessarily bound *in vivo*, consistent with our understanding of mechanisms underlying tissue-specific programs of gene expression. Conversely, TF targets may not contain consensus binding motifs [84, 85], suggesting that the TFs are binding to uncharacterized motifs or through cooperation with additional factors. Thus, any apparent connection between expression data and promoter DNA sequence is, at best, circumstantial evidence of TF binding.

The development of technologies like ChIP-chip enabled the rapid and direct biochemical purification of DNA sequences bound by TFs in the genome *in vivo* and the subsequent generation of target maps and transcriptional regulatory networks. The first global studies of TF binding in yeast revealed that, in spite of the presence of con-

sensus binding motifs for Gal4 and Ste12 throughout the yeast genome, these factors localize to the promoters of functionally related genes to form distinct regulatory modules [86]. Similar patterns were observed for the TF Rap1 [87], suggesting that additional features such as chromatin architecture are involved in the selective binding of these TFs to promoters. Extension of this assay to over 100 yeast TFs revealed that many yeast promoters are bound by multiple TFs, echoing the combinatorial complexity postulated for higher eukaryotes [88]. This study also introduced the integration of network motifs (such as autoregulation, feedforward and multi-input) with expression data to construct regulatory networks for processes like metabolism and the cell cycle. Such strategies were expanded to include over 200 yeast TFs [89], resulting in the discovery of novel regulatory DNA sequences, insights into promoter structure, and a system of TF classification based on functional binding data. These experiments provided the first broad view of promoter topography on a genomic scale.

Such investigations are more complex in higher eukaryotes. Metazoans are composed of many different cell types, requiring a much larger arsenal of TFs to regulate elaborate patterns of differentiation, homeostasis and environmental response, not to mention the corresponding increase in the size and complexity of the genome. Given the larger size of mammalian genomes, initial location analyses in mouse and human systems examined patterns of TF binding using microarrays representing thousands of promoter regions, which at the time were the only regulatory elements that could be effectively located. Even examining these small fractions of the genome in tissue-specific contexts proved enlightening. For example, an investigation of TCF4 target genes revealed that the EDN1 oncogene is a direct regulatory target of β-catenin in colon cancer, providing important insight into the activation of this growth factor in colon and other cancers [90]. An examination of several myogenic TFs at promoters in proliferating and differentiating mouse myoblasts uncovered a complex, dynamic network governing skeletal myogenesis as well as unexpected involvement in stress response and regeneration [91]. Studying the binding patterns of HNFs in human liver and pancreatic cells revealed distinct and common regulatory targets between tissues and provided mechanistic insight into the potential of HNF4α misregulation to contribute to type II diabetes [92]. Similar experiments with c-Myc and its binding partner Max in Burkitt's lymphoma cells revealed that over 15% of the promoters studied are bound by both factors [84], comparable to observations in HL60 cells [93]. The surprisingly large number of targets for these TFs suggested a general role for c-Myc in global transcriptional regulation, a model supported by additional experiments analyzing Myc targets and gene expression in *Drosophila* [94].

While these studies provoked new ideas about transcriptional regulation at mammalian promoters, the coverage and resolution of the microarray platforms used in these experiments limited the insight that could be gained. Improved genome sequence annotation and technological advances in microarray synthesis and analysis led to the development of 'tiling' arrays, wherein short oligonucleotide probes provide continuous coverage along large regions of the genome, in contrast to previous arrays that sampled isolated chunks of promoters or other genomic sites. A more advanced promoter microarray platform was developed that covered 10-kb regions tiling almost 18 000 human promoters with 60-mer oligos, and used to identify targets of Oct4, Sox2 and Nanog in human embryonic stem cells (hESCs) [95]. These experiments yielded precise binding sites of these TFs within their target promoters and revealed a large number of targets common to all three factors, forming coordinated feedforward and auto-regulatory loops with intriguing implications in hESC pluripotency and self-renewal.

In addition to their utility in finding regulatory targets of TFs, tiling arrays have enabled the unbiased discovery of regulatory regions through analysis of genomic binding patterns of TF and other proteins (most of the chromatin architecture discussed above was determined using tiling arrays). Some TFs are found primarily at promoters, like YY1 [K. Wang and B. Ren, unpublished data] and E2F1 [96], and their binding with RNAPII at many promoters (>20% for E2F1) suggests a general role for these TFs in transcriptional regulation. Interestingly, however, a growing number of experiments show that many TFs bind to distal sites throughout the genome, far from any annotated genes. Tiling arrays covering human chromosomes 21 and 22 revealed that only a small fraction of p53 binding occurred near known promoters [97], and similar patterns have been observed for estrogen receptor (ER) in the same regions [98]; transcription factors NF-κB, CREB and STAT1/2 on chromosome 22 [99–101]; and p53 throughout the entire human genome [102]. The widespread binding patterns of these TFs are reminiscent of the genomic distribution of distal regulatory elements like enhancers, and several lines of experimental evidence support a physiological enhancer function for the distal ER binding sites [98]. Another explanation proposed for promoter-distal binding involves regulation of non-coding RNAs [97]. Distal binding sites aside, these experiments identified many novel target genes for these TFs and provided insight into the requirement for and sequence of consensus binding motifs. Additionally, the overlap of TF binding at promoters observed within the experiments above lends support to theories of a combinatorial code in transcriptional regulation in higher eukaryotes, wherein the coordinated action of several TFs at a given promoter is required for precise regulation of expression. Further assessment of binding patterns of ad-

ditional TFs in multiple tissues will hopefully lead to the development of complete human transcriptional regulatory networks that address the complex genetic mechanisms underlying development and disease.

Regulatory mechanisms

In addition to identifying targets of specific TFs, location analysis of components of the basal transcriptional machinery has provided some insight into general mechanisms of gene regulation. The majority of active promoters in human fibroblasts are bound by the general TF TAF1 [25], consistent with the critical role of this protein in PIC assembly. It has also been demonstrated that hypophosphorylated RNAPII is localized primarily at promoters in humans, while total RNAPII is found enriched throughout genes, primarily at exons [25, 103]. These findings are consistent with existing models of transcriptional initiation control through regulated phosphorylation of RNAPII [6], and support coordinated mechanisms for transcriptional elongation and mRNA processing events. About 75% of promoters occupied by the PIC appeared to be transcriptionally active, indicating that TAF1 and RNAPII occupancy are a general feature of active promoters, even considering the diversity of core promoter elements found in these promoters [25]. Promoters marked by a PIC but with no evidence of transcription could reflect the competent promoters mentioned earlier, awaiting further activating signals. It is important to remember, however, that the basal transcriptional machinery is not always composed of the same subunits [104], so further large-scale experiments are needed to determine the precise constitution of the PIC at diverse promoters. Additional TAF1 and RNAPII binding distal to known promoters may signify the presence of novel promoters or other putative regulatory elements, providing some insight into mechanisms of interaction between promoters and distal elements like enhancers. Comparison of these sites to high-resolution maps of histone modifications and TF binding should prove informative.

Owing to the diversity of sequence-specific transcription factors in eukaryotic genomes and the coactivators through which they mediate transcriptional regulation [105] and considering the tissue-specificity of many gene expression patterns, promoter activation is difficult to generalize at the level of the sequence-specific TF. Some common patterns of coregulator localization, however, have recently begun to emerge. Most active promoters in yeast are occupied by HAT enzymes like Gcn5 and Esa1 [60, 106], consistent with models linking gene activation to acetylation of histones by these enzymes and with the acetylation patterns observed at active promoters as discussed above. Similarly, the HAT p300 has been observed at many active promoters in human cells [N. Heintzman and B. Ren, unpublished data], supporting a conserved role for such factors in positively regulating transcription. The precise purpose of histone acetylation at promoters is not yet known, but several lines of thought address the mechanistic significance of this modification. Many transcriptional regulatory proteins (including TAF1) possess bromodomains capable of recognizing acetylated lysines, which would serve to initiate and stabilize interactions between these proteins and the promoter region [107]. Histone acetylation also appears to influence binding of sequence specific transcription factors to DNA by revealing some consensus binding motifs and occluding others [J. Lanier and E. Turner, personal communication], similar to the formation of the NFR that presumably facilitates binding of the transcriptional machinery. Furthermore, histone deacetylation has been linked to transcriptional elongation [12], so it is possible that the relatively hyperacetylated histones at promoters serve to distinguish physically adjacent yet functionally discrete components of a genetic unit.

As with the various HATs, the HMT responsible for catalyzing the trimethylation of H3K4 in yeast, Set1, has also been demonstrated to associate with the promoter regions of active genes [108], and similar patterns were observed with the human Set1 homolog, MLL1 [109]. Again, the functional significance of this modification has yet to be entirely deciphered, but as with acetylation, methlyated lysines can be recognized by numerous regulatory proteins that contain chromodomains [107, 110]. Additional evidence suggests that H3K4me3 may be involved in regulating HAT and HDAC activity in the rapid turnover of acetylation at active promoters [111]. The hyperacetylation could then be preferentially maintained at promoters while H3K4me2, H3K4me1, and/or other distinct methylated histone residues facilitate the aforementioned deacetylation that occurs in coding regions, again creating a functional compartmentalization mediated and marked by a methylation gradient. Intriguingly, several recent reports identify PHD-finger-containing proteins as novel recognizers of H3K4me3 with implications in both maintenance and repression of gene expression [112–115], suggesting that H3K4me3 is a multi-purpose marker for active promoters, recognized in specific contexts by activator or repressor proteins in response to cellular signaling pathways. Further experiments are required to more finely resolve these regulatory mechanisms, but the presence of various HATs and HMTs at the majority of active promoters is consistent with a general role for these factors in transcriptional regulation.

One attribute common to histone features and transcription factor binding at promoters is their association with maintaining patterns of gene activity through mitosis [116], even when these promoters are not transcriptionally active. This ‘gene bookmarking’ supports the concept of a cellular memory, in which epigenetic features associated with gene activity persist through transcriptional

inactivation to mark these genes for potential subsequent reactivation, protecting them from permanent silencing through incorporation into heterochromatin. Additional genome-scale studies will be useful in elucidating the connections between transcriptional activation and maintenance of promoter competence and activity.

Connecting sequence to regulation

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

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

Additional evidence links CpG islands to bidirectional promoters, which represent over 10% of human promoters; intriguingly, 77% of bidirectional promoters are located within CpG islands, while only 8% of these promoters contain a TATA-box [117]. This study also found conservation of these bidirectional promoter structures in mouse, and uncovered interesting relationships between promoter bidirectionality and gene function and regulation of expression. While this investigation showed that a significant proportion of genes appear to share promoter sequences, other recent studies have revealed widespread usage of alternative promoters throughout mammalian genomes by examining binding of the transcriptional machinery to multiple sites at gene 5′-ends [25], transcript-based identification of adjacent but distinct TSS [23, 118], and functional analysis of putative promoters [54]. In addition to demonstrating the tissue specificity of many promoters even without the influence of distal regulatory elements, this functional study also found distinct regions of the proximal promoter that are related to transcriptional activity, including the intriguing general presence of positive regulatory regions 40–350 bp upstream of the TSS and negative regulatory regions 350–1000 bp upstream of the TSS [118]. The mechanisms of regulation by these regions have yet to be determined, but such findings clearly highlight the importance of considering the proximal promoter when studying transcriptional activation and repression. In addition to providing insight into the general functional properties of promoters, such large-scale functional assays also form the basis for investigating the contributions of DNA sequence and chromatin structure to tissue-specific gene expression and promoter usage.

Conclusion and perspectives

Constantly evolving computational and experimental methodologies will continue to make significant contributions to our knowledge of promoter signatures at the DNA sequence and epigenetic levels. Genomic sequencing of additional organisms and advances in sequence alignment strategies will provide expanded resources for comparative promoter analyses, potentially revealing novel promoter sequence elements with transcriptional regulatory properties. Furthermore, only a small fraction of the >100 known histone modifications have currently been mapped on a large scale. Future studies will investigate these modifications in other systems and will expand to include additional modifications and histone variants, contributing to a more complete understanding of the chromatin architecture at promoters and other transcriptional regulatory elements. Another current focus of epigenetic research is examining the patterns of DNA methylation, wherein methylation of cytosine (usually within CpG islands) represses gene expression by inducing heterochromatin formation or by interfering with transcription factor binding [119]. The recent development of a large-scale DNA methylation profiling assay enabled the generation of a DNA methylation map of the entire human genome [120], revealing surprising results related to the role of DNA methylation in heterochromatin formation,

X-chromosome silencing and development of malignant cancer. A similar study examining a large collection of human promoters uncovered evidence for a targeted instructive mechanism for DNA methylation of promoters in cancer cells [121]. Additional experiments are needed to resolve the mechanisms underlying DNA methylation during development and oncogenesis and the impact of this modification on transcriptional regulation.

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

Acknowledgements. This work was supported by funding from LICR, NHGRI, and NCI (BR), and an NIH graduate training grant (NH).

1 Orphanides, G. and Reinberg, D. (2002) A unified theory of gene expression. Cell 108, 439–451.
2 Smale, S. T. and Kadonaga, J. T. (2003) The RNA polymerase II core promoter. Annu. Rev. Biochem. 72, 449–479.
3 West, A. G. and Fraser, P. (2005) Remote control of gene transcription. Hum. Mol. Genet. 14 Spec. No. 1, R101–R111.
4 Mellor, J. (2005) The dynamics of chromatin remodeling at promoters. Mol. Cell 19, 147–157.
5 Nightingale, K. P., O'Neill L, P. and Turner, B. M. (2006) Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. Curr. Opin. Genet. Dev. 16, 125–136.
6 Hahn, S. (2004) Structure and mechanism of the RNA polymerase II transcription machinery. Nat. Struct. Mol. Biol. 11, 394–403.
7 Butler, J. E. and Kadonaga, J. T. (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev. 16, 2583–2592.
8 Glass, C. K. and Rosenfeld, M. G. (2000) The coregulator exchange in transcriptional functions of nuclear receptors. Genes Dev. 14, 121–141.
9 Lemon, B. and Tjian, R. (2000) Orchestrated response: a symphony of transcription factors for gene control. Genes Dev. 14, 2551–2569.
10 Lewis, B. A. and Reinberg, D. (2003) The mediator coactivator complex: functional and physical roles in transcriptional regulation. J. Cell Sci. 116, 3667–3675.
11 Ju, B. G., Lunyak, V. V., Perissi, V., Garcia-Bassets, I., Rose, D. W., Glass, C. K. and Rosenfeld, M. G. (2006) A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription. Science 312, 1798–1802.
12 Lieb, J. D. and Clarke, N. D. (2005) Control of transcription through intragenic patterns of nucleosome composition. Cell 123, 1187–1190.
13 (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640.
14 Schmid, C. D., Perier, R., Praz, V. and Bucher, P. (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. Nucleic Acids Res. 34, D82–D85.
15 Maruyama, K. and Sugano, S. (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene 138, 171–174.
16 Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A. and Sugano, S. (1997) Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. Gene 200, 149–156.
17 Yamashita, R., Suzuki, Y., Wakaguri, H., Tsuritani, K., Nakai, K. and Sugano, S. (2006) DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. Nucleic Acids Res. 34, D86–D89.
18 Ng, P., Wei, C. L., Sung, W. K., Chiu, K. P., Lipovich, L., Ang, C. C., Gupta, S., Shahab, A., Ridwan, A., Wong, C. H., Liu, E. T. and Ruan, Y. (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat. Methods 2, 105–111.
19 Hashimoto, S., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S. and Matsushima, K. (2004) 5′-end SAGE for the analysis of transcriptional start sites. Nat. Biotechnol. 22, 1146–1149.
20 Kasai, Y., Hashimoto, S., Yamada, T., Sese, J., Sugano, S., Matsushima, K. and Morishita, S. (2005) 5′SAGE: 5′-end Serial Analysis of Gene Expression database. Nucleic Acids Res. 33, D550–D552.
21 Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P. and Hayashizaki, Y. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. USA 100, 15776–15781.
22 Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B. et al. (2005) The transcriptional landscape of the mammalian genome. Science 309, 1559–1563.
23 Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engstrom, P. G., Frith, M. C., Forrest, A. R., Alkema, W. B., Tan, S. L. et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. 38, 626–635.
24 Kim, T. H. and Ren, B. (2006) Genome-wide analysis of protein-DNA interactions. Ann. Rev. Genomics Hum. Genet. 24 May [Epub ahead of print].
25 Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D. and Ren, B. (2005) A high-resolution map of active promoters in the human genome. Nature 436, 876–880.
26 Carthew, R. W. (2006) Gene regulation by microRNAs. Curr. Opin. Genet. Dev. 16, 203–208.
27 Bajic, V. B., Tan, S. L., Suzuki, Y. and Sugano, S. (2004) Promoter prediction analysis on the whole human genome. Nat. Biotechnol. 22, 1467–1473.
28 Xuan, Z., Zhao, F., Wang, J., Chen, G. and Zhang, M. Q. (2005) Genome-wide promoter extraction and analysis in human, mouse, and rat. Genome Biol. 6, R72.
29 Bucher, P. and Trifonov, E. N. (1986) Compilation and analysis of eukaryotic POL II promoter sequences. Nucleic Acids Res. 14, 10009–10026.

30 Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002) The human genome browser at UCSC. Genome Res. 12, 996–1006.
31 Breathnach, R. and Chambon, P. (1981) Organization and expression of eucaryotic split genes coding for proteins. Annu. Rev. Biochem. 50, 349–383.
32 Struhl, K. (1995) Yeast transcriptional regulatory mechanisms. Annu. Rev. Genet. 29, 651–674.
33 Burley, S. K. and Roeder, R. G. (1996) Biochemistry and structural biology of transcription factor IID (TFIID). Annu. Rev. Biochem. 65, 769–799.
34 Berk, A. J. (2000) TBP-like factors come into focus. Cell 103, 5–8.
35 Smale, S. T. and Baltimore, D. (1989) The 'initiator' as a transcription control element. Cell 57, 103–113.
36 Smale, S. T., Schmidt, M. C., Berk, A. J. and Baltimore, D. (1990) Transcriptional activation by Sp1 as directed through TATA or initiator: specific requirement for mammalian transcription factor IID. Proc. Natl. Acad. Sci. USA 87, 4509–4513.
37 Chalkley, G. E. and Verrijzer, C. P. (1999) DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator. EMBO J. 18, 4835–4845.
38 Kadonaga, J. T. (2002) The DPE, a core promoter element for transcription by RNA polymerase II. Exp. Mol. Med. 34, 259–264.
39 Burke, T. W. and Kadonaga, J. T. (1996) *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes Dev. 10, 711–724.
40 Kutach, A. K. and Kadonaga, J. T. (2000) The downstream promoter element DPE appears to be as widely used as the TATA-box in *Drosophila* core promoters. Mol. Cell. Biol. 20, 4754–4764.
41 Burke, T. W. and Kadonaga, J. T. (1997) The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by TAFII60 of *Drosophila.* Genes Dev. 11, 3020–3031.
42 Butler, J. E. and Kadonaga, J. T. (2001) Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. Genes Dev. 15, 2515–2519.
43 Willy, P. J., Kobayashi, R. and Kadonaga, J. T. (2000) A basal transcription factor that activates or represses transcription. Science 290, 982–985.
44 Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D. and Ebright, R. H. (1998) New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. Genes Dev. 12, 34–44.
45 Evans, R., Fairley, J. A. and Roberts, S. G. (2001) Activator-mediated disruption of sequence-specific DNA contacts by the general transcription factor TFIIB. Genes Dev. 15, 2945–2949.
46 Ohler, U., Liao, G. C., Niemann, H. and Rubin, G. M. (2002) Computational analysis of core promoters in the *Drosophila* genome. Genome Biol. 3, RESEARCH0087.
47 Lim, C. Y., Santoso, B., Boulay, T., Dong, E., Ohler, U. and Kadonaga, J. T. (2004) The MTE, a new core promoter element for transcription by RNA polymerase II. Genes Dev. 18, 1606–1617.
48 Lewis, B. A., Kim, T. K. and Orkin, S. H. (2000) A downstream element in the human beta-globin promoter: evidence of extended sequence-specific transcription factor IID contacts. Proc. Natl. Acad. Sci. USA 97, 7172–7177.
49 Ince, T. A. and Scotto, K. W. (1995) Differential utilization of multiple transcription start points accompanies the overexpression of the P-glycoprotein-encoding gene in Chinese hamster lung cells. Gene 156, 287–290.
50 Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S. and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 434, 338–345.
51 Blake, M. C., Jambou, R. C., Swick, A. G., Kahn, J. W. and Azizkhan, J. C. (1990) Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter. Mol. Cell. Biol. 10, 6632–6641.
52 Bajic, V. B., Choudhary, V. and Hock, C. K. (2004) Content analysis of the core promoter region of human genes. In Silico Biol. 4, 109–125.
53 Bajic, V. B., Tan, S. L., Christoffels, A., Schonbach, C., Lipovich, L., Yang, L., Hofmann, O., Kruger, A., Hide, W., Kai, C., Kawai, J., Hume, D. A., Carninci, P. et al. (2006) Mice and men: their promoter properties. PLoS Genet. 2, e54.
54 Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L. and Myers, R. M. (2006) Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome Res. 16, 1–10.
55 Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y., Suyama, A., Sakaki, Y., Morishita, S. et al. (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res. 11, 677–684.
56 Margueron, R., Trojer, P. and Reinberg, D. (2005) The key to development: interpreting the histone code? Curr. Opin. Genet. Dev. 15, 163–176.
57 Sarma, K. and Reinberg, D. (2005) Histone variants meet their match. Nat. Rev. Mol. Cell. Biol. 6, 139–149.
58 Bernstein, B. E., Liu, C. L., Humphrey, E. L., Perlstein, E. O. and Schreiber, S. L. (2004) Global nucleosome occupancy in yeast. Genome Biol. 5, R62.
59 Lee, C. K., Shibata, Y., Rao, B., Strahl, B. D. and Lieb, J. D. (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat. Genet. 36, 900–905.
60 Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., Bell, G. W., Walker, K., Rolfe, P. A., Herbolsheimer, E., Zeitlinger, J., Lewitter, F., Gifford, D. K. et al. (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 122, 517–527.
61 Sekinger, E. A., Moqtaderi, Z. and Struhl, K. (2005) Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. Mol. Cell 18, 735–748.
62 Yuan, G. C., Liu, Y. J., Dion, M. F., Slack, M. D., Wu, L. F., Altschuler, S. J. and Rando, O. J. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. Science 309, 626–630.
63 Guillemette, B., Bataille, A. R., Gevry, N., Adam, M., Blanchette, M., Robert, F. and Gaudreau, L. (2005) Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. PLoS Biol. 3, e384.
64 Raisner, R. M., Hartley, P. D., Meneghini, M. D., Bao, M. Z., Liu, C. L., Schreiber, S. L., Rando, O. J. and Madhani, H. D. (2005) Histone variant H2A.Z marks the 5′ ends of both active and inactive genes in euchromatin. Cell 123, 233–248.
65 Zhang, H., Roberts, D. N. and Cairns, B. R. (2005) Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss. Cell 123, 219–231.
66 Raisner, R. M. and Madhani, H. D. (2006) Patterning chromatin: form and function for H2A.Z variant nucleosomes. Curr. Opin. Genet. Dev. 16, 119–124.
67 Mito, Y., Henikoff, J. G. and Henikoff, S. (2005) Genome-scale profiling of histone H3.3 replacement patterns. Nat. Genet. 37, 1090–1097.

68 Bernstein, B. E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D. K., Huebert, D. J., McMahon, S., Karlsson, E. K., Kulbokas, E. J., 3rd, Gingeras, T. R., Schreiber, S. L. and Lander, E. S. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120, 169–181.

69 Felsenfeld, G. (1996) Chromatin unfolds. Cell 86, 13–19.

70 Hake, S. B. and Allis, C. D. (2006) Histone H3 variants and their potential role in indexing mammalian genomes: the 'H3 barcode hypothesis'. Proc. Natl. Acad. Sci. USA 103, 6428–6435.

71 Strahl, B. D. and Allis, C. D. (2000) The language of covalent histone modifications. Nature 403, 41–45.

72 Turner, B. M. (2005) Reading signals on the nucleosome with a new nomenclature for modified histones. Nat. Struct. Mol. Biol. 12, 110–112.

73 Turner, B. M. (1993) Decoding the nucleosome. Cell 75, 5–8.

74 Schubeler, D., MacAlpine, D. M., Scalzo, D., Wirbelauer, C., Kooperberg, C., van Leeuwen, F., Gottschling, D. E., O'Neill, L. P., Turner, B. M., Delrow, J., Bell, S. P. and Groudine, M. (2004) The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev. 18, 1263–1271.

75 Roh, T. Y., Cuddapah, S. and Zhao, K. (2005) Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes Dev. 19, 542–552.

76 Liu, C. L., Kaplan, T., Kim, M., Buratowski, S., Schreiber, S. L., Friedman, N. and Rando, O. J. (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. PLoS Biol. 3, e328.

77 Dion, M. F., Altschuler, S. J., Wu, L. F. and Rando, O. J. (2005) Genomic characterization reveals a simple histone H4 acetylation code. Proc. Natl. Acad. Sci. USA 102, 5501–5506.

78 Schreiber, S. L. and Bernstein, B. E. (2002) Signaling network model of chromatin. Cell 111, 771–778.

79 Bannister, A. J. and Kouzarides, T. (2005) Reversing histone methylation. Nature 436, 1103–1106.

80 Boyer, L. A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L. A., Lee, T. I., Levine, S. S., Wernig, M., Tajonar, A., Ray, M. K., Bell, G. W., Otte, A. P., Vidal, M., Gifford, D. K., Young, R. A. and Jaenisch, R. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature 441, 349–353.

81 Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., Chevalier, B., Johnstone, S. E., Cole, M. F., Isono, K., Koseki, H., Fuchikami, T., Abe, K., Murray, H. L. et al. (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 125, 301–313.

82 Craig, J. M. (2005) Heterochromatin – many flavours, common themes. Bioessays 27, 17–28.

83 Kadonaga, J. T. (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell 116, 247–257.

84 Li, Z., Van Calcar, S., Qu, C., Cavenee, W. K., Zhang, M. Q. and Ren, B. (2003) A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. Proc. Natl. Acad. Sci. USA 100, 8164–8169.

85 Zhang, X., Odom, D. T., Koo, S. H., Conkright, M. D., Canettieri, G., Best, J., Chen, H., Jenner, R., Herbolsheimer, E., Jacobsen, E., Kadam, S., Ecker, J. R., Emerson, B., Hogenesch, J. B., Unterman, T., Young, R. A. and Montminy, M. (2005) Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues. Proc. Natl. Acad. Sci. USA 102, 4459–4464.

86 Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P. and Young, R. A. (2000) Genome-wide location and function of DNA binding proteins. Science 290, 2306–2309.

87 Lieb, J. D., Liu, X., Botstein, D. and Brown, P. O. (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat. Genet. 28, 327–334.

88 Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K. and Young, R. A. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298, 799–804.

89 Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E. and Young, R. A. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431, 99–104.

90 Kim, T. H., Xiong, H., Zhang, Z. and Ren, B. (2005) beta-Catenin activates the growth factor endothelin-1 in colon cancer cells. Oncogene 24, 597–604.

91 Blais, A., Tsikitis, M., Acosta-Alvear, D., Sharan, R., Kluger, Y. and Dynlacht, B. D. (2005) An initial blueprint for myogenic differentiation. Genes Dev. 19, 553–569.

92 Odom, D. T., Zizlsperger, N., Gordon, D. B., Bell, G. W., Rinaldi, N. J., Murray, H. L., Volkert, T. L., Schreiber, J., Rolfe, P. A., Gifford, D. K., Fraenkel, E., Bell, G. I. and Young, R. A. (2004) Control of pancreas and liver gene expression by HNF transcription factors. Science 303, 1378–1381.

93 Mao, D. Y., Watson, J. D., Yan, P. S., Barsyte-Lovejoy, D., Khosravi, F., Wong, W. W., Farnham, P. J., Huang, T. H. and Penn, L. Z. (2003) Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. Curr. Biol. 13, 882–886.

94 Orian, A., van Steensel, B., Delrow, J., Bussemaker, H. J., Li, L., Sawado, T., Williams, E., Loo, L. W., Cowley, S. M., Yost, C., Pierce, S., Edgar, B. A., Parkhurst, S. M. and Eisenman, R. N. (2003) Genomic binding by the *Drosophila* Myc, Max, Mad/Mnt transcription factor network. Genes Dev. 17, 1101–1114.

95 Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R. and Young, R. A. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122, 947–956.

96 Bieda, M., Xu, X., Singer, M. A., Green, R. and Farnham, P. J. (2006) Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. Genome Res. 16, 595–605.

97 Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K. and Gingeras, T. R. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116, 499–509.

98 Carroll, J. S., Liu, X. S., Brodsky, A. S., Li, W., Meyer, C. A., Szary, A. J., Eeckhoute, J., Shao, W., Hestermann, E. V., Geistlinger, T. R., Fox, E. A., Silver, P. A. and Brown, M. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell 122, 33–43.

99 Martone, R., Euskirchen, G., Bertone, P., Hartman, S., Royce, T. E., Luscombe, N. M., Rinn, J. L., Nelson, F. K., Miller, P., Gerstein, M., Weissman, S. and Snyder, M. (2003) Distribu-

tion of NF-kappaB-binding sites across human chromosome 22. Proc. Natl. Acad. Sci. USA 100, 12247–12252.
100 Euskirchen, G., Royce, T. E., Bertone, P., Martone, R., Rinn, J. L., Nelson, F. K., Sayward, F., Luscombe, N. M., Miller, P., Gerstein, M., Weissman, S. and Snyder, M. (2004) CREB binds to multiple loci on human chromosome 22. Mol. Cell. Biol. 24, 3804–3814.
101 Hartman, S. E., Bertone, P., Nath, A. K., Royce, T. E., Gerstein, M., Weissman, S. and Snyder, M. (2005) Global changes in STAT target selection and transcription regulation upon interferon treatments. Genes Dev. 19, 2953–2968.
102 Wei, C. L., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T., Shahab, A., Yong, H. C., Fu, Y., Weng, Z., Liu, J., Zhao, X. D., Chew, J. L., Lee, Y. L., Kuznetsov, V. A., Sung, W. K., Miller, L. D., Lim, B., Liu, E. T., Yu, Q., Ng, H. H. and Ruan, Y. (2006) A global map of p53 transcription-factor binding sites in the human genome. Cell 124, 207–219.
103 Brodsky, A. S., Meyer, C. A., Swinburne, I. A., Hall, G., Keenan, B. J., Liu, X. S., Fox, E. A. and Silver, P. A. (2005) Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. Genome Biol. 6, R64.
104 Hochheimer, A. and Tjian, R. (2003) Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. Genes Dev. 17, 1309–1320.
105 Taatjes, D. J., Marr, M. T. and Tjian, R. (2004) Regulatory diversity among metazoan co-activator complexes. Nat. Rev. Mol. Cell. Biol. 5, 403–410.
106 Robert, F., Pokholok, D. K., Hannett, N. M., Rinaldi, N. J., Chandy, M., Rolfe, A., Workman, J. L., Gifford, D. K. and Young, R. A. (2004) Global position and recruitment of HATs and HDACs in the yeast genome. Mol. Cell 16, 199–209.
107 de la Cruz, X., Lois, S., Sanchez-Molina, S. and Martinez-Balbas, M. A. (2005) Do protein motifs read the histone code? Bioessays 27, 164–175.
108 Ng, H. H., Robert, F., Young, R. A. and Struhl, K. (2003) Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Mol. Cell 11, 709–719.
109 Guenther, M. G., Jenner, R. G., Chevalier, B., Nakamura, T., Croce, C. M., Canaani, E. and Young, R. A. (2005) Global and Hox-specific roles for the MLL1 methyltransferase. Proc. Natl. Acad. Sci. USA 102, 8603–8608.
110 Daniel, J. A., Pray-Grant, M. G. and Grant, P. A. (2005) Effector proteins for methylated histones: an expanding family. Cell. Cycle 4, 919–926.
111 Hazzalin, C. A. and Mahadevan, L. C. (2005) Dynamic acetylation of all lysine 4-methylated histone H3 in the mouse nucleus: analysis at c-fos and c-jun. PLoS Biol. 3, e393.
112 Li, H., Ilin, S., Wang, W., Duncan, E. M., Wysocka, J., Allis, C. D. and Patel, D. J. (2006) Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. Nature 442, 91–95.
113 Pena, P. V., Davrazou, F., Shi, X., Walter, K. L., Verkhusha, V. V., Gozani, O., Zhao, R. and Kutateladze, T. G. (2006) Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. Nature 442, 100–103.
114 Shi, X., Hong, T., Walter, K. L., Ewalt, M., Michishita, E., Hung, T., Carney, D., Pena, P., Lan, F., Kaadige, M. R., Lacoste, N., Cayrou, C., Davrazou, F., Saha, A., Cairns, B. R., Ayer, D. E., Kutateladze, T. G., Shi, Y., Cote, J., Chua, K. F. and Gozani, O. (2006) ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. Nature 442, 96–99.
115 Wysocka, J., Swigut, T., Xiao, H., Milne, T. A., Kwon, S. Y., Landry, J., Kauer, M., Tackett, A. J., Chait, B. T., Badenhorst, P., Wu, C. and Allis, C. D. (2006) A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. Nature 442, 86–90.
116 Sarge, K. D. and Park-Sarge, O. K. (2005) Gene bookmarking: keeping the pages open. Trends Biochem. Sci. 30, 605–610.
117 Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otillar, R. P. and Myers, R. M. (2004) An abundance of bidirectional promoters in the human genome. Genome Res. 14, 62–66.
118 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. et al. (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. Genome Res. 16, 55–65.
119 Bird, A. (2002) DNA methylation patterns and epigenetic memory. Genes Dev. 16, 6–21.
120 Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L. and Schubeler, D. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat. Genet. 37, 853–862.
121 Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E., Pikarski, E., Young, R. A., Niveleau, A., Cedar, H. and Simon, I. (2006) Evidence for an instructive mechanism of de novo methylation in cancer cells. Nat. Genet. 38, 149–153.