Transcriptional Regulatory Elements in the Human Genome

Glenn A. Maston, Sara K. Evans, and Michael R. Green

Howard Hughes Medical Institute, Programs in Gene Function and Expression and Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts 01605; email: Glenn.Maston@umassmed.edu, Sara.Evans@umassmed.edu, Michael.Green@umassmed.edu

Key Words

bioinformatics, functional genomics, transcription factors

Abstract

The faithful execution of biological processes requires a precise and carefully orchestrated set of steps that depend on the proper spatial and temporal expression of genes. Here we review the various classes of transcriptional regulatory elements (core promoters, proximal promoters, distal enhancers, silencers, insulators/boundary elements, and locus control regions) and the molecular machinery (general transcription factors, activators, and coactivators) that interacts with the regulatory elements to mediate precisely controlled patterns of gene expression. The biological importance of transcriptional regulation is highlighted by examples of how alterations in these transcriptional components can lead to disease. Finally, we discuss the methods currently used to identify transcriptional regulatory elements, and the ability of these methods to be scaled up for the purpose of annotating the entire human genome.

INTRODUCTION

LCR: locus control region

Combinatorial control: the concerted action of combinations of multiple transcriptional regulatory elements and their cognate transcription factors

The faithful execution of biological processes such as development, proliferation, apoptosis, aging, and differentiation requires a precise and carefully orchestrated set of steps that depend on the proper spatial and temporal expression of genes. As a result, deregulation of gene expression can often lead to disease. The completion of the human genome sequence and its annotation using computational and comparative genomic methods has led to the cataloging of ∼20,000–25,000 protein-coding genes (39). Key questions now relate to understanding how these genes and their products function, as well as how their spatial and temporal expression patterns are established at both the cellular and organismal level.

To understand the molecular mechanisms that govern specific expression patterns on a global scale, it is important to identify the transcriptional regulatory elements associated with each predicted gene. Moreover, the ability to identify such elements is an important step toward understanding how gene expression is altered in pathological conditions. Thus, one of the main emerging challenges for genomics research is to identify all functional elements in the genome, including those that regulate gene expression. The availability of the complete human genome sequence, in combination with genome-wide expression data, will facilitate the comprehensive identification of these transcriptional regulatory elements. In addition, these resources serve as a starting point for studying transcription regulation of human genes on a global scale, and provide information regarding the establishment of spatial and temporal gene expression patterns and the mechanisms required for their establishment.

Here we review the various classes of transcriptional regulatory elements and the current understanding of how they function. We begin with an overview of the eukaryotic transcription process and the molecular machinery that drives it. We then focus on the role of transcriptional regulatory elements in gene expression and highlight diseases that result from their alteration. Finally, we review the methods currently used to identify transcriptional regulatory elements, both experimentally and through bioinformatics approaches.

EUKARYOTIC TRANSCRIPTION: AN OVERVIEW

The expression of eukaryotic protein-coding genes (also called class II or structural genes) can be regulated at several steps, including transcription initiation and elongation, and mRNA processing, transport, translation, and stability. Most regulation, however, is believed to occur at the level of transcription initiation. In eukaryotes, transcription of protein-coding genes is performed by RNA polymerase II. Genes transcribed by RNA polymerase II typically contain two distinct families of *cis*-acting transcriptional regulatory DNA elements: (*a*) a promoter, which is composed of a core promoter and nearby (proximal) regulatory elements, and (*b*) distal regulatory elements, which can be enhancers, silencers, insulators, or locus control regions (LCR) (**Figure 1**). These *cis*-acting transcriptional regulatory elements contain recognition sites for *trans*-acting DNA-binding transcription factors, which function either to enhance or repress transcription.

The structure of human gene promoters can be quite complex, typically consisting of multiple transcriptional regulatory elements. The need for this complexity becomes clear when one considers that although the human genome contains ∼20,000–25,000 genes, each of which may have a unique spatial/temporal expression pattern, it encodes only ∼1850 DNA-binding transcription factors—presumably far less than the number of expression patterns that must be generated (183). The presence of multiple regulatory elements within promoters confers combinatorial control of regulation, which exponentially increases the potential number of unique expression patterns. The challenge now is to

understand how different permutations of the same regulatory elements alter gene expression. An understanding of how the combinatorial organization of a promoter encodes regulatory information first requires an overview of the proteins that constitute the transcriptional machinery.

THE EUKARYOTIC TRANSCRIPTIONAL MACHINERY

Factors involved in the accurate transcription of eukaryotic protein-coding genes by RNA polymerase II can be classified into three groups: general (or basic) transcription factors (GTFs), promoter-specific activator proteins (activators), and coactivators (**Figure 2**). GTFs are necessary and can be sufficient for accurate transcription initiation in vitro (reviewed in 141). Such factors include RNA polymerase II itself and a variety of auxiliary components, including TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. In addition to these "classic" GTFs, it is apparent that in vivo transcription also requires Mediator, a highly conserved, large multisubunit complex that was originally identified in yeast (reviewed in 38, 119).

GTFs assemble on the core promoter in an ordered fashion to form a transcription preinitiation complex (PIC), which directs RNA polymerase II to the transcription start site (TSS). The first step in PIC assembly is binding of TFIID, a multisubunit complex consisting of TATA-box-binding protein (TBP) and a set of tightly bound TBP-associated factors (TAFs). Transcription then proceeds through a series of steps, including promoter melting, clearance, and escape, before a fully functional RNA polymerase II elongation complex is formed. The current model of transcription regulation views this as a cycle, in which complete PIC assembly is stimulated only once. After RNA polymerase II escapes from the promoter, a scaffold structure, composed of TFIID, TFIIE, TFIIH, and Mediator, remains on the core promoter
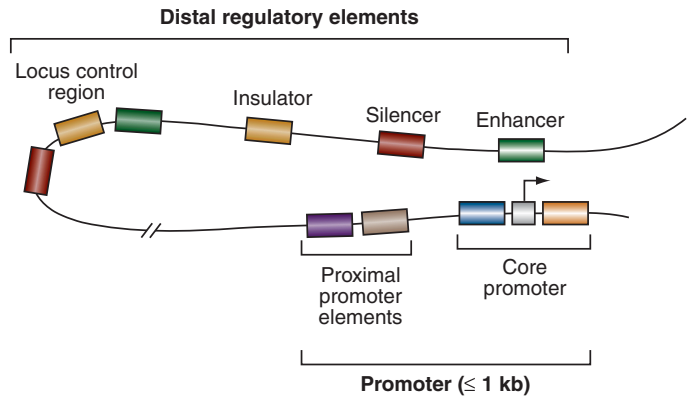

Distal regulatory elements

Proximal promoter elements

Core promoter

Promoter (≤ 1 kb)

Figure 1

Schematic of a typical gene regulatory region. The promoter, which is composed of a core promoter and proximal promoter elements, typically spans less than 1 kb pairs. Distal (upstream) regulatory elements, which can include enhancers, silencers, insulators, and locus control regions, can be located up to 1 Mb pairs from the promoter. These distal elements may contact the core promoter or proximal promoter through a mechanism that involves looping out the intervening DNA.

(73); subsequent reinitiation of transcription then only requires rerecruitment of RNA polymerase II-TFIIF and TFIIB.

The assembly of a PIC on the core promoter is sufficient to direct only low levels of accurately initiated transcription from DNA templates in vitro, a process generally referred to as basal transcription. Transcriptional activity is greatly stimulated by a second class of factors, termed activators. In general, activators are sequence-specific DNA-binding proteins whose recognition sites are usually present in sequences upstream of the core promoter (reviewed in 149). Many classes of activators, discriminated by different DNA-binding domains, have been described, each associating with their own class of specific DNA sequences. Examples of activator families include those containing a cysteine-rich zinc finger, homeobox, helix-loop-helix (HLH), basic leucine zipper (bZIP), forkhead, ETS, or Pit-Oct-Unc (POU) DNA-binding domain (reviewed in 142). In addition to a sequence-specific DNA-binding domain, a typical activator also contains a separable activation domain that is required for the activator to stimulate transcription (149). An

General transcription factor (GTF): a factor that assembles on the core promoter to form a preinitiation complex and is required for transcription of all (or almost all) genes

Coactivators: adaptor proteins that typically lack intrinsic sequence-specific DNA binding but provide a link between activators and the general transcriptional machinery

PIC: preinitiation complex
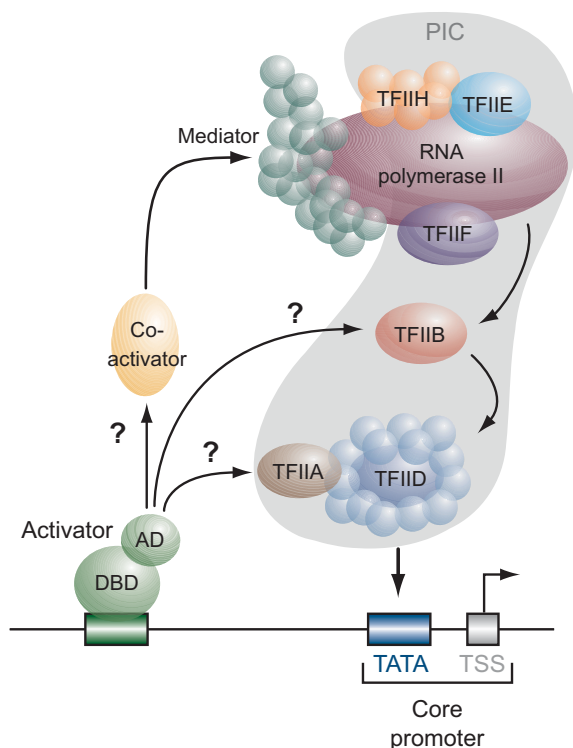
TSS: transcription start site

Figure 2

The eukaryotic transcriptional machinery. Factors involved in eukaryotic transcription by RNA polymerase II can be classified into three groups: general transcription factors (GTFs), activators, and coactivators. GTFs, which include RNA polymerase II itself and TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, assemble on the core promoter in an ordered fashion to form a preinitiation complex (PIC), which directs RNA polymerase II to the transcription start site (TSS). Transcriptional activity is greatly stimulated by activators, which bind to upstream regulatory elements and work, at least in part, by stimulating PIC formation through a mechanism thought to involve direct interactions with one or more components of the transcriptional machinery. Activators consist of a DNA-binding domain (DBD) and a separable activation domain (AD) that is required for the activator to stimulate transcription. The direct targets of activators are largely unknown.

TBP:
TATA-box-binding protein

TAF:
TBP-associated factor

TFBS: transcription factor-binding site

specific activator are typically degenerate, and are therefore described by a consensus sequence in which certain positions are relatively constrained and others are more variable. Many activators form heterodimers and/or homodimers, and thus their binding sites are generally composed of two half-sites. Notably, the precise subunit composition of an activator can also dictate its binding specificity and regulatory action (37).

Although an activator can bind to a wide variety of sequence variants that conform to the consensus, in certain instances the precise sequence of a TFBS can impact the regulatory output. For example, TFBS sequence variations can affect activator binding strength (reviewed in 30), which may be biologically important in situations such as in early development, in which activators are distributed in a concentration gradient (84, 144). TFBS sequence variations may also direct a preference for certain dimerization partners over others (37, 124, 142). Finally, the particular sequence of a TFBS can affect the structure of a bound activator in a way that alters its activity (69, 104, 108, 154, 163). The best-studied examples are nuclear hormone receptors, a large class of ligand-dependent activators. Various studies have shown that the relative orientation of the half-sites, as well as the spacing between them, play a major role in directing the regulatory action of the bound nuclear hormone receptor dimer (37).

Activators work, at least in part, by increasing PIC formation through a mechanism thought to involve direct interactions with one or more components of the transcriptional machinery, termed the "target" (141, 149). Activators may also act by promoting a step in the transcription process subsequent to PIC assembly, such as initiation, elongation, or reinitiation (103). Finally, activators have also been proposed to function by recruiting activities that modify chromatin structure (47, 106). Chromatin often poses a barrier to transcription because it prevents the transcriptional machinery from interacting directly with promoter DNA, and thus can be

extensive discussion of the properties of activators is beyond the scope of this review; readers are referred to several excellent reviews on the subject (87 and references therein).

The DNA-binding sites for activators [also called transcription factor-binding sites (TFBSs)] are generally small, in the range of 6–12 bp, although binding specificity is usually dictated by no more than 4–6 positions within the site. The TFBSs for a

repressive to activator binding and PIC assembly. Chromatin-modifying activities include ATP-dependent remodeling complexes, which use energy to noncovalently modify chromatin structure, and histone-modifying complexes, which add or remove covalent groups (e.g., acetyl groups, methyl groups, and phosphates) from histone tails (103, 137).

The activity of an activator may be modulated by the third group of factors required for eukaryotic transcription: coactivators (reviewed in 115, 168). Typically, coactivators do not exhibit intrinsic sequence-specific DNA binding; instead, they are recruited by protein-protein interactions with one or more DNA-bound activators. Coactivators function in many of the same ways as activators, such as by stimulating PIC assembly or modifying chromatin. The specific set of coactivators present in a cell can play a major role in determining the regulatory response, as they can modify an activator's ability to positively or negatively regulate transcription (106).

A notable property of activators is that they can stimulate transcription synergistically, a phenomenon in which the regulatory effect of multiple factors working together is greater than the sum of the activities driven by each factor individually. This effect can arise from cooperation between multiple copies of the same factor (29), or can be "promiscuous" and result from cooperation between different factors (114) (see also the "Enhanceosomes" sidebar). Significantly, there are limits to the promiscuity of activator cooperativity, and it has been shown that the core promoter can play a role in controlling regulatory signals from upstream elements (132). Transcriptional synergy presumably arises from postbinding interactions, as it can be observed even under conditions of saturated activator binding.

Although the phenomenon of transcriptional synergy has long been recognized, the mechanism underlying it has remained elusive (72). One possibility is that each activator simultaneously interacts with and recruits different GTFs (or cofactors). Another pos-

ENHANCEOSOMES

In some specialized cases, cooperating activators form a tight, stable nucleoprotein complex called an enhanceosome (178). Enhanceosomes appear to act as central processing units, integrating regulatory information from multiple signaling cascades and generating one output to the target promoter. These activators seem to cooperate not in binding, but in activation. In the case of the interferon beta (IFNβ) promoter, multiple activators all present their acidic activation domains together and simultaneously contact the cofactor CBP/p300 (128). Recruitment of the cofactor is most efficient only when all of the activators in the enhanceosome have their activation domains present together. Similar clusters can also interact to repress transcription, and an example of a so-called repressosome has been described (71). Furthermore, it may also be possible that an enhanceosome can switch to a repressosome under different conditions (99). It appears that enhanceosomes tend to form at genes that need to be tightly regulated in medically important pathways, such as wound healing and pathogen defense. Thus, enhanceosome function may be of particular interest for understanding some inherited diseases and how they relate to normal biological processes.

sibility is that different activators may have distinct functions: some may work to modify chromatin structure, whereas others may regulate different steps of transcription, such as promoter escape or elongation. Synergy between identical activators is more difficult to understand; whether each copy of the protein interacts with the same target or different targets remains to be determined.

TRANSCRIPTIONAL REGULATORY ELEMENTS

Core Promoter

The core promoter is the region at the start of a gene that serves as the docking site for the basic transcriptional machinery and PIC assembly, and defines the position of the TSS as well as the direction of transcription (reviewed in 166). The first described core promoter element was the

Transcriptional synergy: the greater-than-additive transcriptional effect resulting from multiple DNA-bound activators
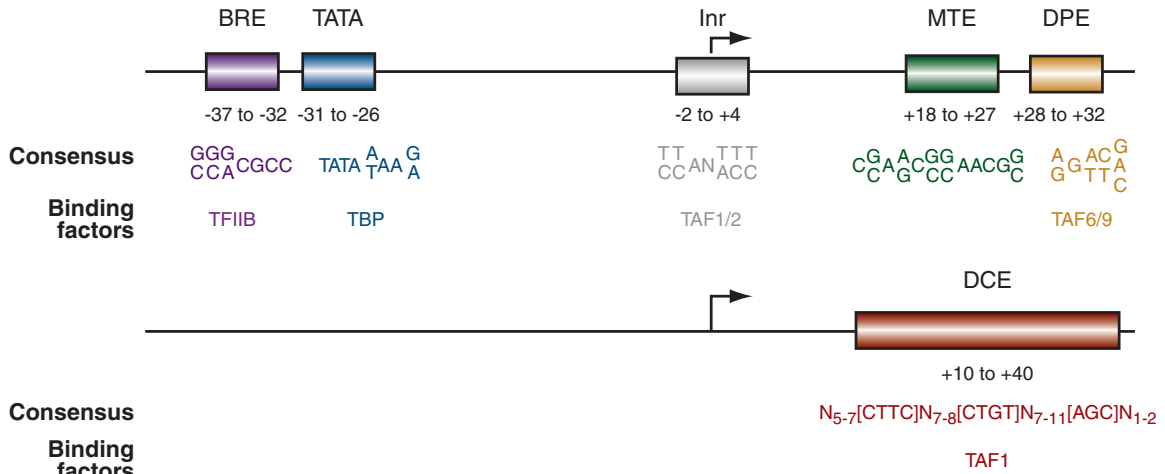
Figure 3

Core promoter elements. Metazoan core promoters are composed of a number of elements that may include a TATA box, an Initiator element (Inr), a Downstream Promoter Element (DPE), a Downstream Core Element (DCE), a TFIIB-Recognition Element (BRE), and a Motif Ten Element (MTE). The human consensus sequence of these elements, their relative positions, and the transcription factors that bind them are shown. The DCE is shown on a separate core promoter for illustration purposes only. Although the DCE can be present in promoters containing a TATA box and/or Inr, it presumably does not occur with a DPE or MTE.

Inr: Initiator

TATA box, the binding site for the TBP subunit of TFIID. In addition to the TATA box, metazoan core promoters can be composed of numerous other elements, including: Initiator element (Inr), Downstream Promoter Element (DPE), Downstream Core Element (DCE), TFIIB-Recognition Element (BRE), and Motif Ten Element (MTE) (113) (**Figure 3**). With the exception of the BRE, which is specifically recognized by TFIIB, all other core promoter elements described to date are TFIID-interaction sites: TAF6 and TAF9 contact the DPE, TAF1 and TAF2 contact the Inr, and TAF1 contacts the DCE (100, 166).

A statistical analysis of ~10,000 predicted human promoters revealed that these known core promoter sequence motifs may not be as universal as previously thought (68). Of the four core promoter elements surveyed (TATA, Inr, DPE, and BRE), the Inr was the most common element, occurring in nearly half of all promoters. By contrast, DPE and BRE were each found in roughly one fourth

of promoters, and TATA boxes were present in only one eighth of promoters. Strikingly, nearly a quarter of all promoters analyzed had none of these four elements, suggesting that either additional core promoter elements or other types of promoter features may yet be discovered. Consistent with this idea, recent reports suggest the existence of more unusual core promoter architectures, such as so-called ATG deserts (102). Moreover, it was recently reported that higher-order structural properties of promoter DNA, which are determined in part by the nucleotide sequence, can be used to identify and classify core promoters (59). Future work may uncover promoter structural properties that are important for GTF-DNA interactions. Indeed, nearly all of the GTFs contact DNA in the core promoter region (reviewed in 73). Although many of those interactions appear to be nonspecific, the efficiency of their function may be affected by structural properties of the promoter DNA, which are affected by the underlying nucleotide content.

Several significant points arise from the observation that core promoters are diverse in their content and organization. First, it is clear that PIC assembly does not depend on a single nucleation point, such as a TATA box; rather, many of the core promoter elements interact with TFIID and stabilize PIC assembly. Second, although it is generally thought that TBP is still required at TATA-less promoters, it also appears that various core promoters may interact preferentially with TFIID complexes having different subunit compositions (36, 133). Such variation may have functional significance, as it has also been observed that different core promoters can limit the upstream regulatory inputs to which they will respond, and thus the core promoter can contribute to the regulatory specificity of a gene (132, 166).

Proximal Promoter Elements

The proximal promoter is defined as the region immediately upstream (up to a few hundred base pairs) from the core promoter, and typically contains multiple binding sites for activators. Historically, vertebrate promoter elements were characterized using a technique called linker-scanning mutagenesis (126). This type of analysis showed that there are multiple functional transcriptional regulatory elements in the region immediately adjacent to the TSS. This early study also showed that regulatory elements acted synergistically, as mutation of any one site caused a significant drop in transcription. As mentioned above, activators are known to work synergistically, but this study of the proximal promoter showed that the synergistic nature of transcriptional regulation is embodied in the promoter structure itself.

An interesting feature of ∼60% of human genes is that their promoter falls near a CpG island (183), a relatively short stretch of DNA, typically 500 bp to 2 kb in length, that has a high G+C nucleotide content and a high frequency of the CpG dinucleotide compared to bulk DNA. Many CpG dinucleotides scat-

tered throughout the genome are methylated at the fifth carbon position of the cytosine base (19); these dinucleotides in CpG islands, however, are normally unmethylated. They are associated with most housekeeping genes as well as many regulated genes (19, 67); in fact, the presence of a CpG island is the most reliable indicator for predicting the presence of a gene (see below) (83). Interestingly, correlations exist between the presence of CpG islands and certain core promoter elements: TATA boxes are more common in promoters that do not have a CpG island nearby, whereas BREs are more common in promoters associated with CpG islands (68).

DNA methylation is associated with transcriptional silencing. Methylation at CpG dinucleotides is believed to repress transcription by blocking the ability of transcription factors to bind their recognition sequences. In addition, methylation-specific binding proteins, such as MeCP2, specifically bind methylated CpG dinucleotides and recruit histone-modifying complexes that establish a repressive chromatin structure (85). The refractory nature of CpG islands to methylation suggests that a role for proximal promoter elements may be to block the local region from being methylated, and therefore inappropriately silenced.

Enhancers

Enhancers were first identified as regions of the SV40 tumor virus genome that could markedly increase the transcription of a heterologous human gene containing a promoter (7, 13, 103). The first human enhancer was found in the immunoglobulin heavy-chain locus (12). Over the past 20 years, the identification of numerous enhancers has shown that they typically regulate transcription in a spatial- or temporal-specific manner, and that they function independent of both the distance from and orientation relative to the promoter. Enhancers are also usually modular, such that a single promoter can be acted upon by distinct enhancer elements at

CpG islands: short stretches of unmethylated DNA that have a high GC content and are associated with the promoters and 5′ ends of most housekeeping genes and many regulated genes

Housekeeping gene: a gene that is involved in basic cell functions, and is constitutively expressed in all (or almost all) cells
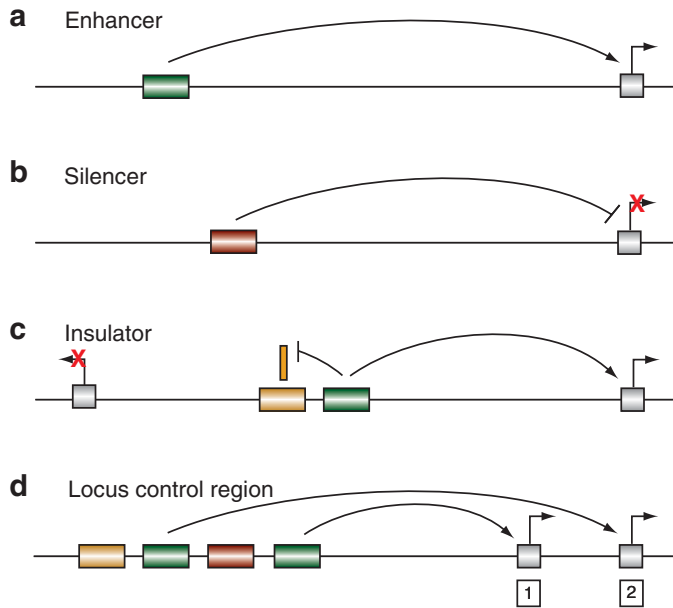
a Enhancer

b Silencer

c Insulator

d Locus control region

Figure 4

Distal transcriptional regulatory elements. (*a*, *b*) Enhancers and silencers function to activate and repress transcription, respectively. (*c*) Insulators function to block genes from being affected by the transcriptional regulatory elements of neighboring genes. (*d*) Locus control regions are typically composed of multiple regulatory elements that function together to confer proper temporal- and/or spatial-specific gene expression to a cluster of nearby genes.

different times or in different tissues, or in response to different stimuli (reviewed in 7). Enhancers are typically composed of a relatively closely grouped cluster of TFBSs that work cooperatively to enhance transcription. The spatial organization and orientation of TFBSs within an enhancer can be critical to its regulatory activity (154, 178); thus, the properties of distance- and orientation independence only apply to the enhancer cluster as a whole.

Enhancers are functionally similar to proximal promoter elements, and the distinction between the two classes is somewhat blurred. In fact, in many cases, the same activators that bind enhancer elements also bind proximal promoter elements in different genes. However, unlike most proximal promoter elements, enhancers are typically long-distance transcriptional control elements that can be

situated quite distally from the core promoter (**Figure 4*a***). For example, enhancers can reside several hundred kilobase pairs upstream of a promoter, downstream of a promoter in an intron, or even beyond the 3′ end of the gene (107 and reviewed in 20).

How do distal elements function over such long physical distances? Data are accumulating in favor of a DNA-looping model, whereby the enhancer and core promoter are brought into close proximity by "looping out" the intervening DNA. A number of recent studies suggest that the DNA-looping model may in fact be a general mechanism by which enhancers function (reviewed in 184). Interestingly, studies have also suggested that PIC formation may begin at a distal enhancer (175), not at the core promoter, as is usually assumed. This would allow for more precise control of the timing of transcription activation, and may be more common in cases in which rapid gene activation is required.

Silencers

Silencers are sequence-specific elements that confer a negative (i.e., silencing or repressing) effect on the transcription of a target gene (**Figure 4*b***). They generally share most of the properties ascribed to enhancers (reviewed in 140). Typically, they function independently of orientation and distance from the promoter, although some position-dependent silencers have been encountered. They can be situated as as part of a proximal promoter, as part of a distal enhancer, or as an independent distal regulatory module; in this regard, silencers can be located far from their target gene, in its intron, or in its 3′-untranslated region. Finally, silencers may cooperate in binding to DNA (74), and they can act synergistically (164).

Silencers are binding sites for negative transcription factors called repressors. Repressor function can require the recruitment of negative cofactors, also called corepressors (148), and in some cases, an activator can switch to a repressor by differential cofactor

recruitment (see, for example, 136, 140, 145). In *Drosophila*, two classes of silencers have been observed: short-range silencers, which generally must reside within ~100 bp of their target gene to have a repressive effect, and long-range silencers, which can repress multiple enhancers or promoters over a span of a few kilobase pairs. It has been suggested that the difference between the two may relate to the recruitment of different cofactors (93).

A number of models have been proposed for repressor function. In some cases, repressors appear to function by blocking the binding of a nearby activator (74), or by directly competing for the same site (see, for example, 110). Alternatively, a repressor may prevent activators and/or GTFs from accessing a promoter by establishing a repressive chromatin structure through the recruitment of histone-modifying activities or chromatin-stabilizing factors (170). Finally, it was recently suggested that a repressor may block transcription activation by inhibiting PIC assembly (35).

For many genes, the "default" transcriptional state is repression, and activation occurs only under specific conditions. One important question is how does a promoter undergo the switch from repression to activation? Recent findings with an interesting class of silencing elements, known as Polycomb group Response Elements (PREs), may shed light on this issue. PREs act as either silencers or antisilencers depending on the protein that is bound, and the switch depends on the presence of noncoding transcription across the PRE element (161). Although the precise mechanism is not understood, the act of transcribing this sequence is thought to induce chromatin modifications that prevent access of repressive complexes to DNA. Noncoding RNAs with no known function have recently been found to be more prevalent than originally anticipated (82), and transcription at silencer elements might represent a novel mechanism by which silencing is counteracted at certain loci.

Insulators

Insulators (also known as boundary elements) function to block genes from being affected by the transcriptional activity of neighboring genes. They thus limit the action of transcriptional regulatory elements to defined domains, and partition the genome into discrete realms of expression (**Figure 4c**). Insulators have two main properties: (*a*) they can block enhancer-promoter communication (i.e., enhancer-blocking activity), and (*b*) they can prevent the spread of repressive chromatin (i.e., heterochromatin-barrier activity). For at least some insulators, these two activities can be separable (152). Typically, insulators are ~0.5–3 kb in length, and function in a position-dependent, orientation-independent manner.

In vertebrates, the most well-characterized insulator element is the chicken β-globin insulator, 5′HS4 (reviewed in 57); a homologous element resides in the human β-globin gene locus (112). Insulator elements have also emerged as a recurrent feature of a number of imprinted loci in the human genome (reviewed in 64); the most well-characterized example is the imprinting control region (ICR) located upstream of the *H19* gene that modulates allele-specific transcription of *H19* and another gene, *Igf2* (11). The number of insulator elements in the human genome is not known. It is now thought, however, that genuine insulator elements may be less common than initially envisaged, and found only in regions with a high density of coding or regulatory information (64).

Although a number of *trans*-acting factors that mediate insulator activity have been identified in *Drosophila* (reviewed in 191), the only known protein to mediate such an activity in vertebrates is CTCF (CCCTC-binding factor). CTCF has been implicated to play a role in many different loci, including chicken globin 5′HS4 (17) and the mammalian *H19/Igf2* ICR (16). The activity of CTCF can be regulated by a number of means, including DNA methylation,

post-translational modification, and interaction with cofactors (reviewed in 190).

The precise mechanism(s) by which insulators carry out their enhancer-blocking and/or heterochromatin-barrier activity is not known. Models proposed to explain insulator function can be broadly classified into two categories (28). The first category posits a link between insulators and the transcriptional regulation machinery; such a model is supported by documented interactions between insulators and transcriptional activators (e.g., see 48). In this model, enhancer-blocking activity is explained by the inability of an insulator-bound activator to interact with its target promoter. Heterochromatin-barrier activity is explained by the recruitment of gene-activating factors or histone-modifying activities, which serve as nucleation sites for a permissive chromatin state that, in turn, blocks the spread of repressive chromatin.

The second category associates insulators with the structural organization of chromatin. Specifically, this model proposes a role for insulators in physically separating chromatin into independent structural domains. This model rests on the assumption that insulators interact with each other and/or with a nuclear attachment substrate, thereby tethering multiple insulator elements to the same foci and resulting in the formation of physically isolated chromatin loops. In this model, positioning an insulator between an enhancer and its target promoter results in enhancer-blocking activity because the physical obstruction between the two elements prevents their communication. Likewise, flanking a gene with insulator elements provides heterochromatin-barrier activity due to the creation of an independent expression domain.

Locus Control Regions

Locus control regions (LCRs) are groups of regulatory elements involved in regulating an entire locus or gene cluster (reviewed in 111) (**Figure 4d**). They are operationally defined as elements that direct tissue-specific, physiological expression of a linked transgene in a position-independent and copy-number-dependent manner. LCRs are typically composed of multiple *cis*-acting elements, including enhancers, silencers, insulators, and nuclear-matrix or chromosome scaffold-attachment regions (MARs or SARs). These elements are bound by transcription factors (both tissue-specific and ubiquitous), coactivators, repressors, and/or chromatin modifiers. Each of the components differentially affects gene expression, and it is their collective activity that functionally defines an LCR and confers proper spatial/temporal gene expression. The most prominent property of LCRs, however, is strong, specific enhancer activity. LCRs are often marked by a cluster of nearby DNase I hypersensitive sites (see below for explanation of DNase I hypersensitivity), and are thought to provide an open-chromatin domain for genes to which they are linked.

The identification of a large number of LCRs has revealed that, like enhancers and silencers, LCRs can regulate gene expression from a distance and that they function in a position-independent manner. Although LCRs are typically located upstream of their target gene(s), they can also be found within an intron of the gene they regulate, exemplified by the human adenosine deaminase LCR (5); downstream of the gene, as in the case of the CD2 (97) or Th2 (101) LCR; or even in the intron of a neighboring gene, as occurs with the CD4 LCR (1).

LCRs have been identified in a broad spectrum of mammalian loci (111). The first LCR identified—and the best-studied one to date—is the mammalian β-globin LCR (reviewed in 34). The human β-globin locus contains five genes that are differentially expressed during development, and are arranged in order of their developmental expression. The β-globin LCR lies ~6–25 kb upstream of the gene cluster, and confers high-level, erythrocyte-specific expression to the genes within the locus. The activity of the β-globin

LCR is orientation-dependent, as inverting the LCR destroys much of its function (177).

How do LCRs accomplish long-range transcriptional control of their target genes? Although a number of models have been proposed (reviewed in 40), a series of recent studies with the β-globin LCR have provided substantial evidence for a "looping" model (reviewed in 15) similar to the enhancer-looping mechanism discussed above. Such long-range physical contacts have been proposed to result in the clustering of sequences into an "active chromatin hub," the formation of which is thought to be crucial for establishing an open-chromatin domain (179). These long-range interactions are only observed when the locus is transcriptionally active, providing support that they play a role in gene activation. The generality of this mechanism for LCR function is supported by the recent observation that similar long-range interactions also occur at the Th2 LCR (169).

TRANSCRIPTIONAL REGULATORY ELEMENTS AND FACTORS IN HUMAN DISEASES

Mutations in transcriptional regulatory elements have been found associated with numerous human disease, an illustrative subset of which are listed in **Table 1**. In many cases, the specific defect is known. For example, mutations in a proximal promoter element of the *GpIb*β gene result in reduced GATA-1 binding and *GpIb*β gene expression, leading to a disease known as Bernard-Soulier Syndrome (117). In other cases, the underlying defect is less well defined. For instance, a 12-mer repeat expansion in the promoter of the *cystatin B* gene has been proposed to cause progressive myoclonus epilepsy, presumably by altering the spacing of elements in the promoter (95).

Similarly, mutations in components of the transcriptional machinery have also been associated with diseases, some of which are listed in **Table 2**. For example, mutations in a subunit of the GTF TFIIH have been associated with the disease xeroderma pigmentosa

(reviewed in 105). Mutations in the activator GATA-1 have been associated with a number of hematopoeitic disorders (reviewed in 27). In addition, mutations in several homeodomain transcription factors (e.g., LMX1B and PHOX2B) are known to cause human diseases (2, 185). Notably, mutations in a number of chromatin-remodeling factors have been associated with cancer. For example, both *BRG1* and *BRM*, mammalian homologs of the SWI/SNF chromatin-remodeling factors, are mutated in numerous cancer cell lines, leading to the altered expression of genes that regulate cell proliferation and metastasis (14). A more extensive compilation of pathologically relevant mutations in regulatory elements and transcription factors is available in the PathoDB database (see link in Related Resources).

A variety of cancers result from chromosomal rearrangements (translocations) involving either regulatory elements or transcription factors. For example, promoter and/or enhancer elements of one gene may become aberrantly linked to a proto-oncogene, thereby causing altered expression of an oncogenic protein. This type of rearrangement is exemplified by fusion of immunoglobulin or T-cell receptor genes to the *cMYC* oncogene, which leads to activation of *cMYC* in Burkitt's lymphoma and acute T-cell leukemia, respectively (reviewed in 146). Chromosomal rearrangements may also lead to the fusion of a transcription factor and another protein, causing the production of a chimeric protein having a new or altered activity. For example, the BCR-ABL fusion associated with chronic myelogenous leukemia brings together the dimerization domain of BCR to the tyrosine kinase ABL, resulting in constitutive kinase activity (reviewed in 157). A fusion event may even involve two transcription factors: for instance, fusion of the transcriptional activation domain of E2A to either PBX-1 or HLF results in pre-B-cell acute lymphoblastic leukemia (reviewed in 98). Interestingly, although recurrent chromosomal rearrangements are characteristic of leukemias

Table 1 Transcriptional regulatory elements involved in human diseases

Regulatory Element	Disease	Mutation (bound factor)	Affected Gene	Reference
Core promoter	β-thalassemia	TATA box, CACCC box, DCE	*β-globin*	(4, 94, 109)
Proximal promoter	Bernard-Soulier Syndrome	133 bp upstream of TSS (GATA-1)	*GpIbβ*	(117)
	Charcot-Marie-Tooth disease	215 bp upstream of TSS	*connexin-32*	(187)
	Congenital erythropoietic porphyria	70, 90 bp upstream of TSS (GATA-1, CP2)	uroporphyrinogen III synthase	(167)
	Familial hypercholesterolemia	43 bp upstream of TSS (Sp1)	*low density lipoprotein receptor*	(92)
	Familial combined hyperlipidemia	39 bp upstream of TSS (Oct-1)	*lipoprotein lipase*	(195)
	Hemophilia	CCAAT box (C/EBP)	*factor IX*	(43)
	Hereditary persistence of fetal hemoglobin	~175 bp upstream of TSS (Oct-1, GATA-1)	*Aγ-globin*	(62)
	Progressive myoclonus epilepsy	Expansion ~70 bp upstream of TSS	*cystatin B*	(96)
	Pyruvate kinase deficient anemia	72 bp upstream of TSS (GATA-1)	*PKLR*	(120)
	β-thalassemia	CACCC box (EKLF)	*β-globin*	(130)
	δ-thalassemia	77 bp upstream of TSS (GATA-1)	*δ-globin*	(125)
	Treacher Collins syndrome	346 bp upstream of TSS (YY1)	*TCOF1*	(123)
Enhancer	Preaxial polydactyly	1 Mb upstream of gene	*SHH*	(107)
	Van Buchem disease	Deletion ~35 kb downstream of gene	*sclerostin*	(116)
	X-linked deafness	Microdeletions 900 kb upstream	*POU3F4*	(46)
Silencer	Asthma and allergies	509 bp upstream of TSS (YY1)	*TFG-β*	(78)
	Fascioscapulohumeral muscular dystrophy	Deletion of D4Z4 repeats	4q35 genes	(66)
Insulator	Beckwith-Wiedemann syndrome	CTCF binding site (CTCF)	*H19/Igf*	(147)
LCR	α-thalassemia	62 kb deletion upstream of gene cluster	*α-globin* genes	(75)
	β-thalassemia	~30 kb deletion removing 5′HS2–5	*β-globin* genes	(52)

and lymphomas, recent evidence indicates they may also be involved in solid tumors. For example, fusions between the androgen-regulated *TMPRSS2* gene and members of the ETS family of transcription factors were recently found to occur in most prostate cancers (180).

A number of recent studies have underscored the possibility of modulating transcription for therapeutic benefit. For instance, insulators have been used to overcome chromatin-dependent repression and to drive high-level, stable expression in gene-therapy applications (reviewed in 153). There

Table 2   Transcriptional machinery components involved in human diseases

Component	Disease	Mutated Factor	Reference
General transcription factors	Xeroderma pigmentosum, Cockayne syndrome, trichothiodystrophy	TFIIH	(105)
Activators	Aniridia	PAX6	(86)
	Campomelic dysplasia	SOX9	(63, 186)
	Congenital central hypoventilation syndrome	PHOX2B	(2)
	Congenital heart disease	Nkx2–5	(162)
	Down syndrome with acute megakaryoblastic leukemia	GATA-1	(77)
	Nail-patella syndrome	LMX1B	(185)
	Prostate cancer	ATBF1	(173)
	X-linked deafness	POU3F4	(45)
	X-linked dyserythropoietic anemia and thrombocytopenia	GATA-1	(138)
	X-linked thrombocytopenia	GATA-1	(65, 127)
Repressors	X linked autoimmunity-allergic dysregulation syndrome	FOXP3	(18)
Coactivators	Parkinson's disease	DJ-1	(23)
	Type II diabetes mellitus	PGC-1	(53)
Chromatin remodeling factors	Cancer	BRG1/BRM	(14)
	Retinal degeneration	ataxin-7	(143)
	Rett syndrome	MeCP2	(3)
	Rubinstein-Taybi syndrome	CREB-binding protein	(135)
	α-thalassemia myelodysplasia syndrome	ATRX	(70)

is also great interest in developing engineered transcriptional activators for use as therapeutic agents in diseases caused by loss of gene expression (reviewed in 91, 151). In addition to the selective reactivation of expression of a specific gene(s), gene expression can also be more generally activated in diseases caused by epigenetic silencing. In particular, many cancers involve the epigenetic inactivation of tumor suppressor genes. DNA-methylation and histone-deacetylation inhibitors can activate epigenetically silenced tumor suppressor genes and are currently under investigation as chemotherapeutic agents (55).

Many human diseases are not caused by a mutation in a single gene, but rather by complex interactions of multiple genes and variants residing therein that may affect, for example, disease susceptibility or progression. Key to understanding the allelic variations that underlie such diseases is categorizing the single-base differences among individuals, known as single-nucleotide polymorphisms

(SNPs). SNPs are the most common type of sequence variants, occurring roughly once in every 1000 bp in the human genome, and are found in both coding and noncoding regions. Thus far, more than four million SNPs in the human genome have been identified and validated (131), and are being used to construct comprehensive variation maps of the human genome (1a). A series of studies analyzing the distribution of SNPs in human promoters found that functional SNPs (i.e., those that result in altered gene expression) occur in 30–60% of human promoters (e.g., see 79, 156) and, moreover, that they tend to cluster in close proximity—within ∼100 bp—of the TSS (25). These data indicate that transcriptional regulatory elements, particularly promoters, may represent a major site where mutations contribute to human disease. Clearly, annotating all functional transcriptional regulatory elements in the human genome will be valuable for future medical studies.

SNP:
single-nucleotide polymorphism

EXPERIMENTAL APPROACHES TO IDENTIFYING TRANSCRIPTIONAL REGULATORY ELEMENTS

Functional Assays that Measure Transcriptional Regulatory Element Activity

One of the more versatile methods for identifying and analyzing transcriptional regulatory element activity is based on the use of a reporter-gene assay. Although traditionally used for directed studies, this method holds the promise of being adapted for use in genome-wide screens. In this assay, the region of DNA to be tested for regulatory activity is cloned into a plasmid upstream of an easily assayable reporter gene, such as the chloramphenicol acetyltransferase (CAT), β-galactosidase, green fluorescent protein (GFP), or luciferase gene. For the purposes of large-scale screens, the genomic segments can be generated randomly either by enzymatic or physical means. The resulting construct is then transfected (either transiently or stably) into cultured cells, and the activity of the reporter is measured to determine if the test segment contains elements that alter reporter gene expression. The precise configuration of the reporter construct depends on the regulatory element to be identified. For instance, if the genomic segment is being tested for core promoter activity, then it is placed immediately upstream of a reporter gene lacking an endogenous promoter (**Figure 5a**). Proximal promoters can be assayed in a similar manner, if they are cloned upstream of a reporter gene driven by a weak heterologous core promoter that allows increases in transcription to be detected (**Figure 5b**). This basic reporter system can also be used to test for enhancers and silencers, if the appropriate strength promoter is used to detect these activities (**Figure 5c,d**). After a genomic segment harboring a regulatory activity is identified, serial deletions, linker-scanning mutagenesis, or site-directed mutagenesis can be employed to more accurately delineate the functional element(s).

Functional assays that measure insulator or LCR activity require more complex reporter constructs and assay systems. Insulator activity can be measured using one of two methods, depending on whether enhancer-blocking or heterochromatin-barrier activity is being assayed (**Figure 5e**). In assays that measure enhancer-blocking activity, the genomic segment containing a putative insulator is positioned between an enhancer and a promoter that are known to interact; if present, an insulator should interfere with enhancer-promoter communication when positioned between the two elements. By contrast, methods that measure heterochromatin-barrier activity require a transgenic reporter assay, in which the reporter gene is stably integrated into the genome. When flanking a transgenic reporter gene, a genomic segment containing an insulator would shield the transgene from position effects, particularly from the repressive effects of heterochromatin, allowing for position-independent reporter gene expression (25a). Similarly, the definitive identification of an LCR requires analyzing the ability of a genomic segment containing an LCR to overcome position effects in a transgenic reporter assay (**Figure 5f**) (72b).

There are several challenges in using functional assays to identify transcriptional regulatory elements. First, regulatory elements can be widely dispersed, and it can be difficult to capture them all in a single reporter construct. Thus, a genomic segment containing only a portion of a promoter element will likely not recapitulate the expression of its corresponding gene. Second, the in vivo activity of a reporter gene may fail to duplicate the expression pattern of its endogenous counterpart due to differences in chromatin context. Third, a given upstream regulatory element may, in reality, only be used in very limited contexts, such as in a specific tissue, developmental stage, or physiological response pathway. If the cell culture system used to assay the reporter gene activity does not match the physiological conditions under which the regulatory element is normally active, then the
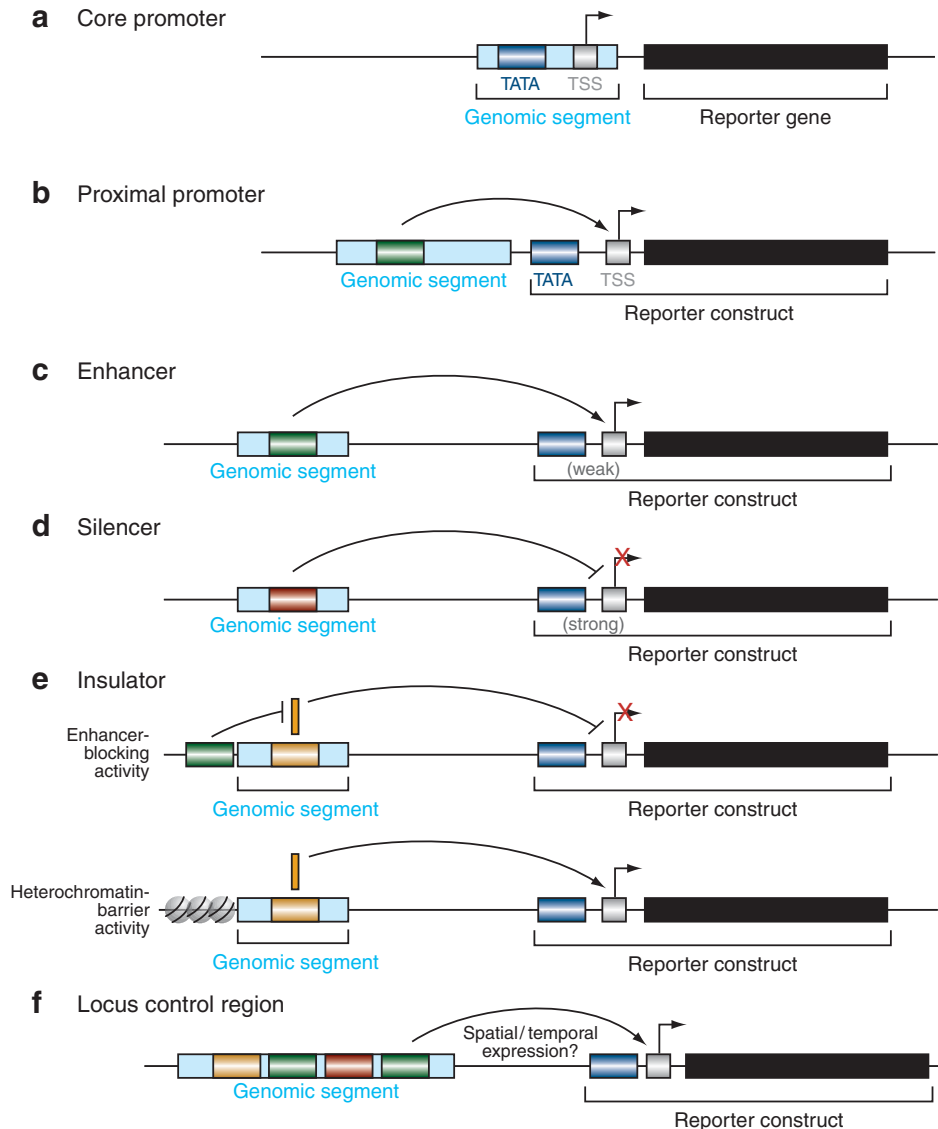
Figure 5

Functional assays that measure transcriptional regulatory element activity. Traditional methods for analyzing the activity of a transcriptional regulatory element are based on the use of plasmid-based or transgenic-reporter gene assays. (*a*) To assay core promoter activity, the genomic segment to be tested (*light blue*) is cloned into a plasmid, immediately upstream of a reporter gene that lacks an endogenous promoter. (*b–d*) Proximal promoters, enhancers, and silencers can be assayed by similar methods, when the genomic segment is cloned upstream of a reporter gene driven by an appropriate promoter. (*e*) Insulator enhancer-blocking activity can be measured using a plasmid-based assay that monitors the ability of a cloned insulator to interfere with enhancer-promoter communication, whereas methods that measure heterochromatin-barrier activity require a transgenic reporter assay to determine the ability of the insulator to shield the transgene from repressive effects of heterochromatin. (*f*) The ability of a locus control region to overcome position effects and confer proper spatial and/or temporal expression is measured by transgenic reporter assay.

element may not be detected. One way to overcome this challenge is by injecting reporter constructs into embryos of model organisms, such as frogs or zebrafish, and following the expression of the reporter gene through development (134, 193). Although these experiments can accurately reveal developmental-specific expression patterns, they are limited by instability and dilution as the embryonic cells multiply; thus, only early developmental events can be reliably assayed in this manner. In addition, the reporter constructs do not become integrated in the host genome, and thus the effects of local chromatin structure on the endogenous gene are not revealed. More sophisticated testing of upstream regulatory elements can be performed by constructing transgenic lines and following reporter gene expression through the entire development of the organism (54, 139). Such a transgenic system overcomes most of the problems associated with simpler reporter gene assays, but is less amenable to large-scale screening. Despite these limitations, however, reporter gene assays remain the most accurate means available to verify the functionality of a transcriptional regulatory element.

Genomic Analysis of Transcription Factor Binding Sites

Several techniques have been developed to identify TFBSs on a genome-wide scale. For example, DNase I hypersensitive site mapping is a technique based on the finding that regions of genomic DNA in which the chromatin state has been perturbed, as can occur due to binding of transcription factors, are more sensitive to DNase I digestion than bulk chromatin. DNase I hypersensitive site mapping has also been used to detect silencers, insulators, and LCRs (72a). Recently, a technique was developed for high-throughput genome-wide detection of DNase I hypersensitive sites (42). Such an approach is powerful in its capacity to detect any regulatory element associated with chromotin perturba-

tion; however, it is limited because the presence of DNase I hypersensitivity at a site implies—but does not demonstrate—an underlying functional transcriptional regulatory element.

Recent experimental analyses of transcription factor binding have taken advantage of the powerful technique of chromatin immunoprecipitation (ChIP), which allows detection and identification of DNA sequences bound by a given protein. DNA purified by ChIP can be either be hybridized to a DNA microarray (ChIP-chip, 155) or cloned to create a "ChIP library" (189) to identify the genomic binding sites of a transcription factor. These methods are powerful because they are unbiased—every TFBS could theoretically be detected. Depending on the protein factor that serves as the immunoprecipitation target, the technique can detect enhancers (24, 80) as well as core promoters (89); it should also be possible to use the technique to identify silencers, insulators, and LCRs. These methodologies, however, have certain limitations. Most notably, ChIP-based methods require a highly specific antibody for each transcription factor of interest. In addition, ChIP-chip experiments are currently limited by the microarray coverage of many genomes of interest. At present, "promoter arrays," such as those that cover ~10-kb regions surrounding the TSSs from ~18,000 known genes (Agilent Technologies), are in use; presumably microarrays covering entire mammalian genomes (e.g., human and mouse) will soon be widely available. By contrast, ChIP cloning is not limited by microarray availability; however, it is more labor-intensive than ChIP-chip, and there is a relatively high background inherent to the cloning procedure that makes it challenging to find bona fide TFBSs.

The data emanating from such large-scale genomic methods must be cautiously interpreted. Although experiments like this show that a transcription factor binds to a certain site in the genome, they do not demonstrate that each and every site is a functional element that regulates transcription of a target gene. In

fact, recent studies suggest that this is highly unlikely. Based on a study of the binding of Sp1, cMyc, and p53 along human chromosomes 21 and 22, an extrapolation to the entire genome predicts a minimum of 12,000 Sp1 binding sites, 25,000 cMyc sites, and 1600 p53 sites (33). Similar results have been obtained for CREB (56) and NF-κB (122). These high numbers are not entirely surprising considering the statistical probability of having a TFBS present by chance; a given 4–6 bp sequence is predicted to occur every ∼250–4000 bp in the human genome. Currently, there is no straightforward method to determine the functional contribution of each candidate TFBS to the regulation of a target gene.

Clearly, one of the challenges in annotating the entire human genome for functional regulatory elements is the sheer magnitude of the task. Indeed, many of the experimental tools that work well for analyzing small regions of DNA are not suitable for high-throughput studies on a genome-wide scale. Toward this end, efforts are under way to adapt existing methods for high-throughput applications, and to develop new methodologies. Much of this is being performed under the auspices of the the ENCODE Project (see sidebar).

COMPUTATIONAL APPROACHES FOR IDENTIFYING TRANSCRIPTIONAL REGULATORY ELEMENTS

Ab Initio Identification of Promoters

As the sequencing of the human genome neared completion, it was clear that computational tools would be required to analyze the enormous amount of newly generated sequence data. Identifying the promoter of a specific gene poses a challenge quite distinct from identifying potential coding regions themselves, as core promoters are often distantly located from the first coding exon due to the presence of 5′-untranslated regions

THE ENCODE PROJECT

In September 2003, the National Human Genome Research Institute (NHGRI) launched the ENCODE (ENCyclopedia of DNA Elements) Project, the goal of which is to annotate the entire human genome for all functional elements. In addition to transcriptional regulatory elements, ENCODE also aims to identify, for example, determinants of chromosome structure and function (such as origins of replication), sequences that affect/control chromosome biology (such as recombination hot spots), and sites of epigenetic changes (such as DNA methylation and chromatin modifications). Initially, ENCODE has focused on a selected 1% (∼30 Mb) of the human genome, and this pilot phase will test and compare a diverse set of new and existing experimental procedures, computational tools, and technologies to identify functional elements. All data generated by ENCODE are being released into public databases. For more information, see **http://www.genome.gov/encode**.

and introns. In addition, because promoters can contain any one of a number of combinations of core promoter elements [and, conversely, many promoters have only one or no such elements (68)], simply searching for the co-occurrence of known core promoter motifs has had only limited success (58). The most successful promoter prediction programs are instead based on the analysis of training data sets (i.e., known core promoters) to look for functionally undefined sequence contexts that are common to all promoters, and then scanning genomic sequences for new occurrences of such sequence contexts. This method has been implemented alone (PromoterInspector; 160), in combination with the modeling of promoter features, such as relation to a CpG island and a potential first exon (FirstEF; 44), and by building a sequence- and positionally constrained promoter model from the training data set (Eponine; 51).

Although much improved over earlier prediction programs, these methods still have limited sensitivity and specificity when applied to genome-scale sequence data (6, 9), primarily resulting from two limitations: first,

the programs depend on the quantity and quality of the available data used for their training; and second, they are limited to finding core promoters that are similar to ones that have already been identified. Toward this end, experimentally verified core promoters and TSSs were recently compiled into high-quality databases [EDP (32) and DbTSS (174)]. Further experimental work aimed at both identifying novel transcripts (31) and testing computational predictions (50) will provide ample data from which to discover novel promoter structures and construct better models of core promoters.

Significantly, there is a major difference in the accurate ab initio identification of promoters with and without an associated CpG island. Recent experiments have confirmed the long-held observation that proximity to a CpG island correlates strongly with a broad, nonspecific pattern of expression, as commonly found with housekeeping genes (194). Consistent with the fact that approximately half of the genes in the human genome fall near CpG islands, a recent critical comparison of promoter-prediction programs found that there is generally good success at predicting this class of promoters (9). Unfortunately, for the other half of genes not associated with CpG islands, whose tissue-specific regulation is arguably more interesting and complex, ab initio promoter predictions are much less reliable.

Ab Initio Identification of Upstream Regulatory Elements

A number of bioinformatics approaches can be used for ab initio identification of previously unidentified upstream transcriptional regulatory elements. Classically, an unannotated sequence can be scanned for sequence motifs that match known TFBSs, which have been experimentally identified from other promoters/regulatory sites. Experimental data regarding the specific binding sites of most well-characterized transcrip-

tion factors have been compiled in databases such as TRANSFAC (192). Multiple examples of experimentally determined TFBSs are then used to build a position-specific scoring matrix for each factor (172). Programs such as MatInspector (150) and, more recently, MATCH (88) compare a genomic sequence input to all the matrices in TRANSFAC, and return a list of potential TFBSs based on a statistical match between a region in the sequence and a site matrix. This analysis is often hampered by the prediction of a large number of sites, a significant fraction of which are likely false positives. This may be due, at least in part, to the quality of the data used to build the TFBS matrices (60). Recently, databases such as JASPAR (158) were developed that use more sophisticated statistical models of TFBSs. In addition to the false-positive problem, the completeness of these databases is also an issue; it is likely that not all DNA-binding transcription factors have been identified, and even for some known factors, their binding specificity has not yet been fully characterized.

Use of a priori expression knowledge. An alternative analysis technique used to overcome the above-mentioned challenges is to amass genes that are suspected to be coregulated (or experimentally determined to be co-expressed, such as from a microarray analysis), and search for common sequence motifs in their upstream regions. This not only allows for the possibility of discovering novel TFBSs, but also for reducing the number of predictions generated. To date, many different programs have become available that implement different algorithms for motif discovery in this setting; AlignACE (81) and MEME (8) are two of the most well known. The plethora of programs available can be overwhelming; to this end, the field is becoming more self-critical and finding ways to evaluate and compare the performance of such programs (181). It is clear that there is room for improvement, especially when analyzing

metazoan sequences, in which transcription factor cooperativity is much more widespread than in yeast and lower eukaryotes. In fact, further improvement in the success of predicting TFBSs has come from algorithms that search for clustered binding sites (182 and citations therein).

Comparative genomics approaches. Another strategy that has become widely exploited to refine searches for TFBSs involves the use of comparative genomics, specifically comparative sequence analysis. In one form of this, known as phylogenetic footprinting (176), genomic sequences from species separated by large evolutionary distances are compared, and those sequences found to be in common (i.e., conserved) are regarded as candidates for being functionally important. This approach is based on the expectation that functional TFBSs will be conserved through evolution, and can thus be detected when orthologous sequences from distantly related species are aligned. A number of programs have been developed to perform such analyses, such as FootPrinter (21) and PhastCons (165). As with the other prediction tools discussed above, a recent analysis of the accuracy of some of these programs suggests that they are acceptable, but imperfect, in correctly identifying known functional sites (90). Two thorough reviews have covered the growing field of comparative genomics (129) and the challenges faced in the statistical implementation of comparative sequence analyses (171). The comments below are thus limited to a broader perspective on the use of comparative genomics for finding functional TFBSs.

Comparative genomics approaches are often complicated by two factors. First, although there is ample evidence that conserved regions do, indeed, often contain functional regulatory motifs (121, 139, 193), this correlation does not always hold (10), and other explanations for observed conservation have been suggested (26). The lack of a precise correlation between conservation and function results, in part, from the presence of a large amount of highly conserved noncoding sequences in the human genome. Genome-wide comparisons have revealed surprising statistics about the frequency of such sequences, some that span >1 kb, which do not follow the pattern expected for any of the known types of transcriptional regulatory elements or clusters of elements (41, 165). It remains to be determined if these conserved regions contain elements relevant to transcriptional regulation, or if they perhaps serve an as-yet defined other role.

The second problem is that not all TFBSs are conserved among species. For example, it has been estimated that roughly one third of TFBSs are not conserved between human and rodents (49). This could be due to a number of reasons. First, due to the degeneracy of TFBSs, perfect sequence conservation of a site is not required; as a result, the same factor may bind to sequence variants of the TFBS that are present in different species. Second, although gene-expression patterns may be conserved across species, a specific regulatory element may not be conserved (61, 118, 188); this can occur because of redundancy of regulatory elements (76, 159) that allows a single element to be gained or lost without affecting the overall expression of the gene.

Finally, some of the most important transcriptional regulatory elements relevant to normal human development and disease may not be highly conserved, but rather might be found only in humans or shared with a small group of our primate relatives. Indeed, it has also been hypothesized that weakly conserved TFBSs may be medically important (171). Detecting these sites by computational methods will likely depend on advances in comparative genomics; this may require new analytical approaches, such as phylogenetic shadowing (22) that analyzes closely related sequences (e.g., those from primates), and increasing the total number of species for which genomic sequence data are available.

Phylogenetic footprinting: multispecies comparative sequence analysis method used to identify highly conserved sequences present in evolutionarily diverse species

Phylogenetic shadowing: an approach for comparative sequence analyses that compares closely related sequences rather than distantly related sequences

CONCLUSIONS

The picture that is emerging suggests that transcriptional regulation is a much more dynamic process than was once perceived. Interplay between the entire suite of core promoters, proximal regulatory elements, and distal regulatory elements, as well as their binding factors and cofactors, contribute to the precise nature of the transcriptional output of a given promoter. Regulatory systems are robust and redundant, and yet highly sensitive as well: Even single-nucleotide differences in a regulatory sequence can have significant effects on gene expression. These results suggest that transcriptional regulation can cover a broad, continuous spectrum of regulatory control, such that it is likely that discrete models of regulatory action may apply to only limited sets of promoters.

Current endeavors aiming to annotate all of the transcriptional regulatory elements in the human genome face considerable challenges. TFBSs are small and degenerate, are often located distantly from the promoter upon which they act, and are not always conserved through evolution. These properties make regulatory elements difficult to identify through computational means alone. Many experimental methods show binding of a transcription factor at a given site, but do not assess the functional significance of that binding. Functional assays that directly assess the regulatory capacity of a site are the best available tools, and the current challenge is to adapt these methods for their high-throughput usage to screen the entire human genome.

SUMMARY POINTS

1. The concerted action of multiple different transcriptional regulatory elements, along with their cognate activators and coactivators, contributes to the overall spatial and temporal regulation of a gene's expression pattern.

2. The modular nature of promoters confers combinatorial control of gene expression; that is, the number of possible gene expression patterns far exceeds the total number of transcription factors.

3. Although an activator can bind to a wide variety of sequence variants within a regulatory element that conform to the consensus, in certain instances, the precise sequence of a TFBS can modulate the activity of an activator.

4. Long-range transcriptional regulatory elements, including enhancers, silencers, insulators, and LCRs, may function through a DNA-looping mechanism that brings regulatory elements into proximity by "looping out" the intervening DNA.

5. Numerous human diseases and disorders have been associated with mutations in both transcriptional regulatory elements and various components of the transcriptional machinery.

6. A major challenge for genomics research is to identify all functional elements in the human genome, including those that regulate gene expression.

7. Both experimental and computational approaches are being developed to identify transcriptional regulatory elements on a genome-wide scale.

8. A predicted TFBS is not necessarily a bona fide binding site, and binding does not necessarily demonstrate a functional role for that site; it is likely that bioinformatics methods will not replace the need for experimental verification of regulatory elements.

FUTURE DIRECTIONS/UNRESOLVED ISSUES

1. Elucidate the precise mechanisms of action of transcriptional activators and repressors.

2. Develop methods to determine the functional contribution of each TFBS to the regulation of its target gene.

3. Determine whether there are rules for the specific combinations of activators that underlie combinatorial control of gene expression.

4. Develop approaches for identifying functional transcriptional regulatory sites on a genome-wide scale.

ACKNOWLEDGMENTS

LITERATURE CITED

1. Adlam M, Siu G. 2003. Hierarchical interactions control CD4 gene expression during thymocyte development. *Immunity* 18:173–84

1a. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, et al. 2005. A haplotype map of the human genome. *Nature* 437:1299–320

2. Amiel J, Laudier B, Attie-Bitach T, Trang H, de Pontual L, et al. 2003. Polyalanine expansion and frameshift mutations of the paired-like homeobox gene *PHOX2B* in congenital central hypoventilation syndrome. *Nat. Genet.* 33:459–61

3. Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY. 1999. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.* 23:185–88

4. Antonarakis SE, Irkin SH, Cheng TC, Scott AF, Sexton JP, et al. 1984. β-Thalassemia in American Blacks: novel mutations in the "TATA" box and an acceptor splice site. *Proc. Natl. Acad. Sci. USA* 81:1154–58

5. Aronow BJ, Silbiger RN, Dusing MR, Stock JL, Yager KL, et al. 1992. Functional analysis of the human adenosine deaminase gene thymic regulatory region and its ability to generate position-independent transgene expression. *Mol. Cell. Biol.* 12:4170–85

6. Ashurst JL, Collins JE. 2003. Gene annotation: prediction and testing. *Annu. Rev. Genomics Hum. Genet.* 4:69–88

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

8. Bailey TL, Elkan C. 1995. The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3:21–29

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

10. **Balhoff JP, Wray GA. 2005. Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites. *Proc. Natl. Acad. Sci. USA* 102:8591–96**

10. Comparative sequence analysis is used to show that transcription factor binding sites in the *endo16* gene promoter are no more conserved than the surrounding (nonfunctional) sequence.

11. Banerjee S, Smallwood A, Lamond S, Campbell S, Nargund G. 2001. Igf2/H19 imprinting control region (ICR): an insulator or a position-dependent silencer? *Sci. World J.* 1:218–24

12. Banerji J, Olson L, Schaffner W. 1983. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* 33:729–40

13. Banerji J, Rusconi S, Schaffner W. 1981. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27:299–308

14. Banine F, Bartlett C, Gunawardena R, Muchardt C, Yaniv M, et al. 2005. SWI/SNF chromatin-remodeling factors induce changes in DNA methylation to promote transcriptional activation. *Cancer Res.* 65:3542–47

15. Bank A. 2006. Regulation of human fetal hemoglobin: new players, new complexities. *Blood* 107:435–43

16. Bell AC, Felsenfeld G. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* 405:482–85

17. Bell AC, West AG, Felsenfeld G. 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* 98:387–96

18. Bennett CL, Christie J, Ramsdell F, Brunkow ME, Ferguson PJ, et al. 2001. The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3. *Nat. Genet.* 27:20–21

19. Bird A. 1987. CpG islands as gene markers in the vertebrate nucleus. *Trends Genet.* 3:342–47

20. Blackwood EM, Kadonaga JT. 1998. Going the distance: a current view of enhancer action. *Science* 281:60–63

21. Blanchette M, Tompa M. 2003. FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res.* 31:3840–42

22. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, et al. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299:1391–94

23. Bonifati V, Rizzu P, van Baren MJ, Schaap O, Breedveld GJ, et al. 2003. Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. *Science* 299:256–59

24. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122:947–56

25. Buckland PR, Hoogendoorn B, Coleman SL, Guy CA, Smith SK, O'Donovan MC. 2005. Strong bias in the location of functional promoter polymorphisms. *Hum. Mutat.* 26:214–23

25a. Burgess-Beusse B, Farrell C, Gaszner M, Litt M, Mutskov V, et al. 2002. The insulation of genes from external enhancers and silencing chromatin. *Proc. Natl. Acad. Sci. USA* 99:16433–37

26. **Cameron RA, Chow SH, Berney K, Chiu TY, Yuan QA, et al. 2005. An evolutionary constraint: strongly disfavored class of change in DNA sequence during divergence of *cis*-regulatory modules. *Proc. Natl. Acad. Sci. USA* 102:11769–74**

27. Cantor AB. 2005. GATA transcription factors in hematologic disease. *Int. J. Hematol.* 81:378–84

28. Capelson M, Corces VG. 2004. Boundary elements and nuclear organization. *Biol. Cell* 96:617–29

29. Carey M, Lin YS, Green MR, Ptashne M. 1990. A mechanism for synergistic activation of a mammalian gene by GAL4 derivatives. *Nature* 345:361–64

26. Comparative analyses of regulatory sequences suggest that sequence conservation is caused by the fixation of neutral variants at nonfunctional sites near the TFBS and not the result of selection against mutations in TFBSs themselves.

30. Carey M, Smale ST. 2000. *Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques*. Cold Spring Harbor, NY: Cold Spring Harbor Lab. Press

31. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* 309:1559–63

32. Cavin Perier R, Junier T, Bucher P. 1998. The Eukaryotic Promoter Database EPD. *Nucleic Acids Res* 26:353–57

33. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, et al. 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

34. Chakalova L, Carter D, Debrand E, Goyenechea B, Horton A, et al. 2005. Developmental regulation of the β-globin gene locus. *Prog. Mol. Subcell. Biol.* 38:183–206

35. Chen L, Widom J. 2005. Mechanism of transcriptional silencing in yeast. *Cell* 120:37–48

36. Chen Z, Manley JL. 2003. Core promoter elements and TAFs contribute to the diversity of transcriptional activation in vertebrates. *Mol. Cell. Biol.* 23:7350–62

37. Claessens F, Gewirth DT. 2004. DNA recognition by nuclear receptors. *Essays Biochem.* 40:59–72

38. Conaway JW, Florens L, Sato S, Tomomori-Sato C, Parmely TJ, et al. 2005. The mammalian Mediator complex. *FEBS Lett* 579:904–8

39. Consortium IHGS. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–45

40. Cook PR. 2003. Nongenic transcription, gene regulation and action at a distance. *J. Cell. Sci.* 116:4483–91

41. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* 15:901–13

42. Crawford GE, Holt IE, Mullikin JC, Tai D, Blakesley R, et al. 2004. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc. Natl. Acad. Sci. USA* 101:992–97

43. Crossley M, Brownlee GG. 1990. Disruption of a C/EBP binding site in the factor IX promoter is associated with haemophilia B. *Nature* 345:444–46

44. Davuluri RV, Grosse I, Zhang MQ. 2001. Computational identification of promoters and first exons in the human genome. *Nat. Genet.* 29:412–17

45. de Kok YJ, van der Maarel SM, Bitner-Glindzicz M, Huber I, Monaco AP, et al. 1995. Association between X-linked mixed deafness and mutations in the POU domain gene *POU3F4*. *Science* 267:685–88

46. de Kok YJ, Vossenaar ER, Cremers CW, Dahl N, Laporte J, et al. 1996. Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene *POU3F4*. *Hum. Mol. Genet.* 5:1229–35

47. de la Serna IL, Ohkawa Y, Berkes CA, Bergstrom DA, Dacwag CS, et al. 2005. MyoD targets chromatin remodeling complexes to the myogenin locus prior to forming a stable DNA-bound complex. *Mol. Cell. Biol.* 25:3997–4009

48. Defossez PA, Kelly KF, Filion GJ, Perez-Torrado R, Magdinier F, et al. 2005. The human enhancer blocker CTC-binding factor interacts with the transcription factor Kaiso. *J. Biol. Chem.* 280:43017–23

49. Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* 19:1114–21

50. Dike S, Balija VS, Nascimento LU, Xuan Z, Ou J, et al. 2004. The mouse genome: experimental examination of gene predictions and transcription start sites. *Genome Res.* 14:2424–29

35. Results reported here indicate that, in some cases, silencing does not prevent activator binding to DNA, but does prevent RNA polymerase II, TFIIB, and TFIIE recruitment at the core promoter.

47. Explains a case where chromatin remodeling, and thus tissue-specific regulation, depends on the direct interaction between a specific activator and a ubitquitous activator.

51. Down TA, Hubbard TJ. 2002. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res*. 12:458–61

52. Driscoll MC, Dobkin CS, Alter BP. 1989. Gamma Δ β-thalassemia due to a de novo mutation deleting the 5′ β-globin gene activation-region hypersensitive sites. *Proc. Natl. Acad. Sci. USA* 86:7470–74

53. Ek J, Andersen G, Urhammer SA, Gaede PH, Drivsholm T, et al. 2001. Mutation analysis of peroxisome proliferator-activated receptor-gamma coactivator-1 (PGC-1) and relationships of identified amino acid polymorphisms to Type II diabetes mellitus. *Diabetologia* 44:2220–26

54. Ellingsen S, Laplante MA, Konig M, Kikuta H, Furmanek T, et al. 2005. Large-scale enhancer detection in the zebrafish genome. *Development* 132:3799–811

55. Esteller M. 2005. DNA methylation and cancer therapy: new developments and expectations. *Curr. Opin. Oncol*. 17:55–60

56. Euskirchen G, Royce TE, Bertone P, Martone R, Rinn JL, et al. 2004. CREB binds to multiple loci on human chromosome 22. *Mol. Cell. Biol*. 24:3804–14

57. Felsenfeld G, Burgess-Beusse B, Farrell C, Gaszner M, Ghirlando R, et al. 2004. Chromatin boundaries and chromatin domains. *Cold. Spring Harb. Symp. Quant. Biol*. 69:245–50

58. Fickett JW, Hatzigeorgiou AG. 1997. Eukaryotic promoter recognition. *Genome Res*. 7:861–78

59. Florquin K, Saeys Y, Degroeve S, Rouze P, Van de Peer Y. 2005. Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucleic Acids Res*. 33:4255–64

60. Fogel GB, Weekes DG, Varga G, Dow ER, Craven AM, et al. 2005. A statistical analysis of the TRANSFAC database. *Biosystems* 81:137–54

61. Follows GA, Tagoh H, Lefevre P, Morgan GJ, Bonifer C. 2003. Differential transcription factor occupancy but evolutionarily conserved chromatin features at the human and mouse M-CSF (CSF-1) receptor loci. *Nucleic Acids Res*. 31:5805–16

62. Forget BG. 1998. Molecular basis of hereditary persistence of fetal hemoglobin. *Ann. N. Y. Acad. Sci*. 850:38–44

63. Foster JW, Dominguez-Steglich MA, Guioli S, Kowk G, Weller PA, et al. 1994. Campomelic dysplasia and autosomal sex reversal caused by mutations in an SRY-related gene. *Nature* 372:525–30

64. Fourel G, Magdinier F, Gilson E. 2004. Insulator dynamics and the setting of chromatin domains. *Bioessays* 26:523–32

65. Freson K, Devriendt K, Matthijs G, Van Hoof A, De Vos R, et al. 2001. Platelet characteristics in patients with X-linked macrothrombocytopenia because of a novel GATA1 mutation. *Blood* 98:85–92

66. Gabellini D, Green MR, Tupler R. 2002. Inappropriate gene activation in FSHD: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle. *Cell* 110:339–48

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

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

69. Geserick C, Meyer HA, Haendler B. 2005. The role of DNA response elements as allosteric modulators of steroid receptor function. *Mol. Cell. Endocrinol*. 236:1–7

70. Gibbons RJ, Picketts DJ, Villard L, Higgs DR. 1995. Mutations in a putative global transcriptional regulator cause X-linked mental retardation with α-thalassemia (ATR-X syndrome). *Cell* 80:837–45

71. Gowri PM, Yu JH, Shaufl A, Sperling MA, Menon RK. 2003. Recruitment of a represso-some complex at the growth hormone receptor promoter and its potential role in diabetic nephropathy. *Mol. Cell. Biol.* 23:815–25

72. Green MR. 2005. Eukaryotic transcription activation: right on target. *Mol. Cell* 18:399–402

72a. Gross DS, Garrard WT. 1988. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* 57:159–97

72b. Grosveld F, van Assendelft GB, Greaves DR, Kollias G. 1987. Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell* 51:975–85

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

74. Harris MB, Mostecki J, Rothman PB. 2005. Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function. *J. Biol. Chem.* 280:13114–21

75. Hatton CS, Wilkie AO, Drysdale HC, Wood WG, Vickers MA, et al. 1990. α-thalassemia caused by a large (62 kb) deletion upstream of the human α globin gene cluster. *Blood* 76:221–27

76. Hersh BM, Carroll SB. 2005. Direct regulation of knot gene expression by Ultrabithorax and the evolution of *cis*-regulatory elements in Drosophila. *Development* 132:1567–77

77. Hitzler JK, Zipursky A. 2005. Origins of leukaemia in children with Down syndrome. *Nat. Rev. Cancer* 5:11–20

78. Hobbs K, Negri J, Klinnert M, Rosenwasser LJ, Borish L. 1998. Interleukin-10 and transforming growth factor-β promoter polymorphisms in allergies and asthma. *Am. J. Respir. Crit. Care Med.* 158:1958–62

79. Hoogendoorn B, Coleman SL, Guy CA, Smith K, Bowen T, et al. 2003. Functional analysis of human promoter polymorphisms. *Hum. Mol. Genet.* 12:2249–54

80. Horak CE, Mahajan MC, Luscombe NM, Gerstein M, Weissman SM, Snyder M. 2002. GATA-1 binding sites mapped in the β-globin locus by using mammalian chIp-chip analysis. *Proc. Natl. Acad. Sci. USA* 99:2924–29

81. Hughes JD, Estep PW, Tavazoie S, Church GM. 2000. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 296:1205–14

82. Huttenhofer A, Schattner P, Polacek N. 2005. Non-coding RNAs: hope or hype? *Trends Genet.* 21:289–97

83. Ioshikhes IP, Zhang MQ. 2000. Large-scale human promoter mapping using CpG islands. *Nat. Genet.* 26:61–63

84. Jiang J, Levine M. 1993. Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell* 72:741–52

85. Jones PL, Veenstra GJ, Wade PA, Vermaak D, Kass SU, et al. 1998. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat. Genet.* 19:187–91

86. Jordan T, Hanson I, Zaletayev D, Hodgson S, Prosser J, et al. 1992. The human *PAX6* gene is mutated in two patients with aniridia. *Nat. Genet.* 1:328–32

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

88. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. 2003. MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* 31:3576–79

89. Kim TH, Barrera LO, Qu C, Van Calcar S, Trinklein ND, et al. 2005. Direct isolation and identification of promoters in the human genome. *Genome Res.* 15:830–39

76. Uses both comparative genomics and transgenic functional assays to show that there are multiple independent silencer sites that apparently have redundant effects on the *Drosophila knot* gene.

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

91. Klug A. 2005. Towards therapeutic applications of engineered zinc finger proteins. *FEBS Lett.* 579:892–94

92. Koivisto UM, Palvimo JJ, Janne OA, Kontula K. 1994. A single-base substitution in the proximal Sp1 site of the human low density lipoprotein receptor promoter as a cause of heterozygous familial hypercholesterolemia. *Proc. Natl. Acad. Sci. USA* 91:10526–30

93. Kulkarni MM, Arnosti DN. 2005. *cis*-regulatory logic of short-range transcriptional repression in *Drosophila melanogaster*. *Mol. Cell. Biol.* 25:3411–20

94. Kulozik AE, Bellan-Koch A, Bail S, Kohne E, Kleihauer E. 1991. Thalassemia intermedia: moderate reduction of β-globin gene transcriptional activity by a novel mutation of the proximal CACCC promoter element. *Blood* 77:2054–58

95. Lalioti MD, Scott HS, Antonarakis SE. 1999. Altered spacing of promoter elements due to the dodecamer repeat expansion contributes to reduced expression of the cystatin B gene in EPM1. *Hum. Mol. Genet.* 8:1791–98

96. Lalioti MD, Scott HS, Buresi C, Rossier C, Bottani A, et al. 1997. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature* 386:847–51

97. Lang G, Mamalaki C, Greenberg D, Yannoutsos N, Kioussis D. 1991. Deletion analysis of the human *CD2* gene locus control region in transgenic mice. *Nucleic Acids Res.* 19:5851–56

98. LeBrun DP. 2003. E2A basic helix-loop-helix transcription factors in human leukemia. *Front. Biosci.* 8:206–22

99. Lee B, Vouthounis C, Stojadinovic O, Brem H, Im M, Tomic-Canic M. 2005. From an enhanceosome to a repressosome: Molecular antagonism between glucocorticoids and EGF leads to inhibition of wound healing. *J. Mol. Biol.* 345:1083–97

100. Lee DH, Gershenzon N, Gupta M, Ioshikhes IP, Reinberg D, Lewis BA. 2005. Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. *Mol. Cell. Biol.* 25:9674–86

101. Lee GR, Fields PE, Griffin TJ, Flavell RA. 2003. Regulation of the Th2 cytokine locus by a locus control region. *Immunity* 19:145–53

102. Lee MP, Howcroft K, Kotekar A, Yang HH, Buetow KH, Singer DS. 2005. ATG deserts define a novel core promoter subclass. *Genome Res.* 15:1189–97

103. Lee TI, Young RA. 2000. Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.* 34:77–137

104. Lefstin JA, Yamamoto KR. 1998. Allosteric effects of DNA on transcriptional regulators. *Nature* 392:885–88

105. Lehmann AR. 2001. The xeroderma pigmentosum group D (XPD) gene: one gene, two functions, three diseases. *Genes Dev.* 15:15–23

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

107. Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, et al. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* 12:1725–35

108. Leung TH, Hoffmann A, Baltimore D. 2004. One nucleotide in a kappaB site can determine cofactor specificity for NF-kappaB dimers. *Cell* 118:453–64

109. Lewis BA, Kim TK, Orkin SH. 2000. A downstream element in the human β-globin promoter: evidence of extended sequence-specific transcription factor IID contacts. *Proc. Natl. Acad. Sci. USA* 97:7172–77

102. Describes how a single protein can be generated at a locus in which transcription is initiated from numerous dispersed sites in the absence of any known core promoter.

110. Li L, He S, Sun JM, Davie JR. 2004. Gene regulation by Sp1 and Sp3. *Biochem. Cell. Biol.* 82:460–71

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

112. Li Q, Zhang M, Duan Z, Stamatoyannopoulos G. 1999. Structural analysis and mapping of DNase I hypersensitivity of HS5 of the β-globin locus control region. *Genomics* 61:183–93

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

114. Lin YS, Carey M, Ptashne M, Green MR. 1990. How different eukaryotic transcriptional activators can cooperate promiscuously. *Nature* 345:359–61

115. Lonard DM, O'Malley BW. 2005. Expanding functional diversity of the coactivators. *Trends Biochem. Sci.* 30:126–32

116. Loots GG, Kneissel M, Keller H, Baptist M, Chang J, et al. 2005. Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. *Genome Res.* 15:928–35

117. Ludlow LB, Schick BP, Budarf ML, Driscoll DA, Zackai EH, et al. 1996. Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibβ promoter resulting in the Bernard-Soulier syndrome. *J. Biol. Chem.* 271:22076–80

118. Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403:564–67

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

120. Manco L, Ribeiro ML, Maximo V, Almeida H, Costa A, et al. 2000. A new *PKLR* gene mutation in the R-type promoter region affects the gene transcription causing pyruvate kinase deficiency. *Br. J. Haematol.* 110:993–97

121. Martin N, Patel S, Segre JA. 2004. Long-range comparison of human and mouse Sprr loci to identify conserved noncoding sequences involved in coordinate regulation. *Genome Res.* 14:2430–38

122. Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, et al. 2003. Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc. Natl. Acad. Sci. USA* 100:12247–52

123. Masotti C, Armelin-Correa LM, Splendore A, Lin CJ, Barbosa A, et al. 2005. A functional SNP in the promoter region of TCOF1 is associated with reduced gene expression and YY1 DNA-protein interaction. *Gene* 359:44–52

124. Massari ME, Murre C. 2000. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol. Cell. Biol.* 20:429–40

125. Matsuda M, Sakamoto N, Fukumaki Y. 1992. Δ-thalassemia caused by disruption of the site for an erythroid-specific transcription factor, GATA-1, in the Δ-globin gene promoter. *Blood* 80:1347–51

126. McKnight SL, Kingsbury R. 1982. Transcriptional control signals of a eukaryotic protein-coding gene. *Science* 217:316–24

127. Mehaffey MG, Newton AL, Gandhi MJ, Crossley M, Drachman JG. 2001. X-linked thrombocytopenia caused by a novel mutation of GATA-1. *Blood* 98:2681–88

128. Merika M, Williams AJ, Chen G, Collins T, Thanos D. 1998. Recruitment of CBP/p300 by the IFN β enhanceosome is required for synergistic activation of transcription. *Mol. Cell.* 1:277–87

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

118. Demonstrates that regulatory elements in the *even-skipped* gene stripe 2 enhancer from two related *Drosophila* species are not equivalent; chimeric enhancer constructs are shown to not reproduce the native expression pattern.

130. Moi P, Faa V, Marini MG, Asunis I, Ibba G, et al. 2004. A novel silent β-thalassemia mutation in the distal CACCC box affects the binding and responsiveness to EKLF. *Br. J. Haematol.* 126:881–84

131. Mooney S. 2005. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief. Bioinform.* 6:44–56

132. Morris JR, Petrov DA, Lee AM, Wu CT. 2004. Enhancer choice in *cis* and in *trans* in *Drosophila melanogaster*: role of the promoter. *Genetics* 167:1739–47

133. Muller F, Tora L. 2004. The multicoloured world of promoter recognition complexes. *EMBO J.* 23:2–8

134. Muller F, Williams DW, Kobolak J, Gauvry L, Goldspink G, et al. 1997. Activator effect of coinjected enhancers on the muscle-specific expression of promoters in zebrafish embryos. *Mol. Reprod. Dev.* 47:404–12

135. Murata T, Kurokawa R, Krones A, Tatsumi K, Ishii M, et al. 2001. Defect of histone acetyltransferase activity of the nuclear transcriptional coactivator CBP in Rubinstein-Taybi syndrome. *Hum. Mol. Genet.* 10:1071–76

136. Murayama A, Kim MS, Yanagisawa J, Takeyama K, Kato S. 2004. Transrepression by a liganded nuclear receptor via a bHLH activator through co-regulator switching. *EMBO J.* 23:1598–608

137. Naar AM, Lemon BD, Tjian R. 2001. Transcriptional coactivator complexes. *Annu. Rev. Biochem.* 70:475–501

138. Nichols KE, Crispino JD, Poncz M, White JG, Orkin SH, et al. 2000. Familial dyserythropoietic anaemia and thrombocytopenia due to an inherited mutation in GATA1. *Nat. Genet.* 24:266–70

139. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* 302:413

140. Ogbourne S, Antalis TM. 1998. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem. J.* 331(Pt. 1):1–14

141. Orphanides G, Lagrange T, Reinberg D. 1996. The general transcription factors of RNA polymerase II. *Genes Dev.* 10:2657–83

142. Pabo CO, Sauer RT. 1992. Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* 61:1053–95

143. Palhan VB, Chen S, Peng GH, Tjernberg A, Gamper AM, et al. 2005. Polyglutamine-expanded ataxin-7 inhibits STAGA histone acetyltransferase activity to produce retinal degeneration. *Proc. Natl. Acad. Sci. USA* 102:8472–77

144. Papatsenko D, Levine M. 2005. Quantitative analysis of binding motifs mediating diverse spatial readouts of the Dorsal gradient in the Drosophila embryo. *Proc. Natl. Acad. Sci. USA* 102:4966–71

145. Perissi V, Aggarwal A, Glass CK, Rose DW, Rosenfeld MG. 2004. A corepressor/coactivator exchange complex required for transcriptional activation by nuclear receptors and other regulated transcription factors. *Cell* 116:511–26

146. Popescu NC, Zimonjic DB. 2002. Chromosome-mediated alterations of the MYC gene in human cancer. *J. Cell. Mol. Med.* 6:151–59

147. Prawitt D, Enklaar T, Gartner-Rupprecht B, Spangenberg C, Oswald M, et al. 2005. Microdeletion of target sites for insulator protein CTCF in a chromosome 11p15 imprinting center in Beckwith-Wiedemann syndrome and Wilms' tumor. *Proc. Natl. Acad. Sci. USA* 102:4085–90

148. Privalsky ML. 2004. The role of corepressors in transcriptional regulation by nuclear hormone receptors. *Annu. Rev. Physiol.* 66:315–60

139. Excellent demonstration of the combined use of comparative genomics and functional transgenic testing to identify enhancers in a vertebrate species.

149. Ptashne M, Gann A. 1997. Transcriptional activation by recruitment. *Nature* 386:569–77

150. Quandt K, Frech K, Karas H, Wingender E, Werner T. 1995. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res*. 23:4878–84

151. Rebar EJ. 2004. Development of pro-angiogenic engineered transcription factors for the treatment of cardiovascular disease. *Exp. Opin. Investig. Drugs* 13:829–39

152. Recillas-Targa F, Pikaart MJ, Burgess-Beusse B, Bell AC, Litt MD, et al. 2002. Position-effect protection and enhancer blocking by the chicken β-globin insulator are separable activities. *Proc. Natl. Acad. Sci. USA* 99:6883–88

153. Recillas-Targa F, Valadez-Graham V, Farrell CM. 2004. Prospects and implications of using chromatin insulators in gene therapy and transgenesis. *Bioessays* 26:796–807

154. Remenyi A, Scholer HR, Wilmanns M. 2004. Combinatorial control of gene expression. *Nat. Struct. Mol. Biol*. 11:812–15

155. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* 290:2306–9

156. Rockman MV, Wray GA. 2002. Abundant raw material for *cis*-regulatory evolution in humans. *Mol. Biol. Evol*. 19:1991–2004

157. Saglio G, Cilloni D. 2004. Abl: the prototype of oncogenic fusion proteins. *Cell. Mol. Life Sci*. 61:2897–911

158. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*. 32:D91–94

159. Schaffner G, Schirm S, Muller-Baden B, Weber F, Schaffner W. 1988. Redundancy of information in enhancers as a principle of mammalian transcription control. *J. Mol. Biol*. 201:81–90

160. Scherf M, Klingenhoff A, Werner T. 2000. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol*. 297:599–606

161. **Schmitt S, Prestel M, Paro R. 2005. Intergenic transcription through a polycomb group response element counteracts silencing. *Genes Dev*. 19:697–708**

162. Schott JJ, Benson DW, Basson CT, Pease W, Silberbach GM, et al. 1998. Congenital heart disease caused by mutations in the transcription factor NKX2–5. *Science* 281:108–11

163. Scully KM, Jacobson EM, Jepsen K, Lunyak V, Viadiu H, et al. 2000. Allosteric effects of Pit-1 DNA sites on long-term repression in cell type specification. *Science* 290:1127–31

164. Sertil O, Kapoor R, Cohen BD, Abramova N, Lowry CV. 2003. Synergistic repression of anaerobic genes by Mot3 and Rox1 in *Saccharomyces cerevisiae*. *Nucleic Acids Res*. 31:5831–37

165. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 15:1034–50

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

167. Solis C, Aizencang GI, Astrin KH, Bishop DF, Desnick RJ. 2001. Uroporphyrinogen III synthase erythroid promoter mutations in adjacent GATA1 and CP2 elements cause congenital erythropoietic porphyria. *J. Clin. Invest*. 107:753–62

168. Spiegelman BM, Heinrich R. 2004. Biological control through regulated transcriptional coactivators. *Cell* 119:157–67

169. Spilianakis CG, Lalioti MD, Town T, Lee GR, Flavell RA. 2005. Interchromosomal associations between alternatively expressed loci. *Nature* 435:637–45

161. Identification of noncoding transcription suggests the possibility that there are many more transcriptional regulatory regions than genes in the human genome.

170. Srinivasan L, Atchison ML. 2004. YY1 DNA binding and PcG recruitment requires CtBP. *Genes Dev.* 18:2596–601

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

172. Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16:16–23

173. Sun X, Frierson HF, Chen C, Li C, Ran Q, et al. 2005. Frequent somatic mutations of the transcription factor ATBF1 in human prostate cancer. *Nat. Genet.* 37:407–12

174. Suzuki Y, Yamashita R, Nakai K, Sugano S. 2002. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.* 30:328–31

175. Szutorisz H, Dillon N, Tora L. 2005. The role of enhancers as centres for general transcription factor recruitment. *Trends Biochem. Sci.* 30:593–99

176. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT. 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* 203:439–55

177. Tanimoto K, Liu Q, Bungert J, Engel JD. 1999. Effects of altered gene order or orientation of the locus control region on human β-globin gene expression in mice. *Nature* 398:344–48

178. Thanos D, Maniatis T. 1995. Virus induction of human IFN β gene expression requires the assembly of an enhanceosome. *Cell* 83:1091–100

179. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. 2002. Looping and interaction between hypersensitive sites in the active β-globin locus. *Mol. Cell.* 10:1453–65

180. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, et al. 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 310:644–48

181. Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* 23:137–44

182. Vavouri T, Elgar G. 2005. Prediction of *cis*-regulatory elements using binding site matrices—the successes, the failures and the reasons for both. *Curr. Opin. Genet. Dev.* 15:395–402

183. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51

184. Vilar JM, Saiz L. 2005. DNA looping in gene regulation: from the assembly of macromolecular complexes to the control of transcriptional noise. *Curr. Opin. Genet. Dev.* 15:136–44

185. Vollrath D, Jaramillo-Babb VL, Clough MV, McIntosh I, Scott KM, et al. 1998. Loss-of-function mutations in the LIM-homeodomain gene, LMX1B, in nail-patella syndrome. *Hum. Mol. Genet.* 7:1091–98

186. Wagner T, Wirth J, Meyer J, Zabel B, Held M, et al. 1994. Autosomal sex reversal and campomelic dysplasia are caused by mutations in and around the SRY-related gene SOX9. *Cell* 79:1111–20

187. Wang HL, Wu T, Chang WT, Li AH, Chen MS, et al. 2000. Point mutation associated with X-linked dominant Charcot-Marie-Tooth disease impairs the P2 promoter activity of human connexin-32 gene. *Brain. Res. Mol. Brain Res.* 78:146–53

188. Wang X, Greenberg JF, Chamberlin HM. 2004. Evolution of regulatory elements producing a conserved gene expression pattern in *Caenorhabditis*. *Evol. Dev.* 6:237–45

189. Weinmann AS, Bartley SM, Zhang T, Zhang MQ, Farnham PJ. 2001. Use of chromatin immunoprecipitation to clone novel E2F target promoters. *Mol. Cell. Biol.* 21:6820–32

190. West AG, Fraser P. 2005. Remote control of gene transcription. *Hum. Mol. Genet.* 14(Spec. No. 1):R101–11

191. West AG, Gaszner M, Felsenfeld G. 2002. Insulators: many functions, many mechanisms. *Genes Dev.* 16:271–88

192. Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* 24:238–41

193. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* 3:e7

194. Yamashita R, Suzuki Y, Sugano S, Nakai K. 2005. Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene* 350:129–36

195. Yang WS, Nevin DN, Peng R, Brunzell JD, Deeb SS. 1995. A mutation in the promoter of the lipoprotein lipase (LPL) gene in a patient with familial combined hyperlipidemia and low LPL activity. *Proc. Natl. Acad. Sci. USA* 92:4462–66

RELATED RESOURCES

Ashurst JL, Collins JE. 2003. Gene annotation: prediction and testing. *Annu. Rev. Genomics Hum. Genet.* 4:69–88

Gerasimova TI, Corces VG. 2001. Chromatin insulators and boundaries: effects on transcription and nuclear organization. *Annu. Rev. Genet.* 35:193–208

Lee TI, Young RA. 2000. Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.* 34:77–137

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

Smale ST, Kadonaga JT. 2003. The RNA Polymerase II core promoter. *Annu. Rev. Biochem.* 72:449–79

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

AlignACE: **http://atlas.med.harvard.edu/**

DbTSS: **http://dbtss.hgc.jp/**

EPD: **http://www.epd.isb-sib.ch/**

FirstEF: **http://rulai.cshl.org/tools/FirstEF/**

FootPrinter: **http://wingless.cs.washington.edu/htbin-post/unrestricted/FootPrinter Web/FootPrinterInput2.pl**

JASPAR: **http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl**

MEME: **http://meme.sdsc.edu/meme/website/**

PathoDB: **http://www.gene-regulation.com/pub/databases.html#pathodb**

PhastCons: **http://www.cse.ucsc.edu/~acs/phastCons-HOWTO.html#paper**

The ENCODE Project: **http://www.genome.gov/ENCODE**

TRANSFAC: **http://www.gene-regulation.com/pub/databases.html#transfac**

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* chapters
may be found at http://genom.annualreviews.org/