Insights from genomic profiling of transcription factors

Peggy J. Farnham

Abstract | A crucial question in the field of gene regulation is whether the location at which a transcription factor binds influences its effectiveness or the mechanism by which it regulates transcription. Comprehensive transcription factor binding maps are needed to address these issues, and genome-wide mapping is now possible thanks to the technological advances of ChIP–chip and ChIP–seq. This Review discusses how recent genomic profiling of transcription factors gives insight into how binding specificity is achieved and what features of chromatin influence the ability of transcription factors to interact with the genome. It also suggests future experiments that may further our understanding of the causes and consequences of transcription factor–genome interactions.

Understanding how genomic information is translated into gene regulation has been the subject of intense scientific investigation over the past several decades. Until recently, most studies focused on detailed characterization of a particular gene or gene family. These studies resulted in the development of general principles of gene regulation, but genome-scale studies are now prompting re-examination of some of these principles.

The established view of transcriptional regulation is that *cis*-regulatory elements, such as promoters and enhancers, and proteins that bind to these elements control different levels of transcription of different genes[1,2]. Promoters are composed of common sequence elements, such as a TATA box and an initiator sequence, and binding sites for other transcription factors, which work together to recruit the general transcriptional machinery to the transcriptional start site (TSS). Enhancers also contain binding sites for transcription factors but are located some distance from the site of transcription initiation. Transcriptional activity that results from general factors binding to the core promoter is usually low, but it can be increased by the binding of site-specific factors to proximal promoter regions, which can help to recruit or stabilize the interaction of the general factors at the core promoter. Promoter activity can be further stimulated by the binding of factors to distal enhancer regions and the subsequent recruitment of a histone-modifying enzyme that creates a more favourable chromatin environment for transcription or of a kinase that induces a bound initiation complex to begin elongation (FIG. 1). Transcription can also be modulated by repressive factors that bind to repressing sequences and/or silencers far from the TSS, which can interfere with activator binding (and thus prevent recruitment of the general transcriptional machinery) or recruit histone-modifying complexes that create repressive chromatin structures.

Recent genome-scale studies have enabled more precise definition of thousands of promoters for known genes and have identified many previously unrecognized transcription units, which has revealed that some previous assumptions about transcriptional regulation are not correct. For example, based on the detailed characterization of a small subset of promoters, a typical RNA polymerase II (RNAPII) promoter was thought to contain a TATA box located 30 bp upstream of the TSS. However, we now know that TATA-driven promoters are the exception and not the rule[3,4]. Other recent genomic studies suggest that ~50% of human genes have alternative promoters[5], indicating that regulatory sequences for a particular gene can be spread over a considerable distance. Clearly, access to large data sets documenting RNA expression and transcription factor binding on a genome-wide scale now provides an exciting opportunity for investigators to re-evaluate previous models of transcriptional regulation. Of particular interest is the role of site-specific DNA-binding factors, which is the focus of this Review.

In humans, it has been estimated that there are 200–300 transcription factors that bind to core promoter elements and that can be considered as components of the general transcriptional machinery; such transcription factors include subunits of RNA polymerases and

TATA box
A consensus sequence in promoters that is enriched in thymine and adenine residues and is important for the recruitment of the general transcriptional machinery at some promoters.

Initiator
An element with a consensus of YYANWYY (in which A is the transcription start site, N is any nucleotide, W is adenosine or thymine, and Y is a pyrimidine) that helps to recruit the general transcriptional machinery to promoters.

Initiation complex
The assembly of RNA polymerase and associated general factors that binds to the core promoter region.

Department of Pharmacology and the Genome Center, University of California Davis, Davis, California 95616, USA.
e-mail: *pjfarnham@ucdavis.edu*
doi:10.1038/nrg2636
Published online
11 August 2009

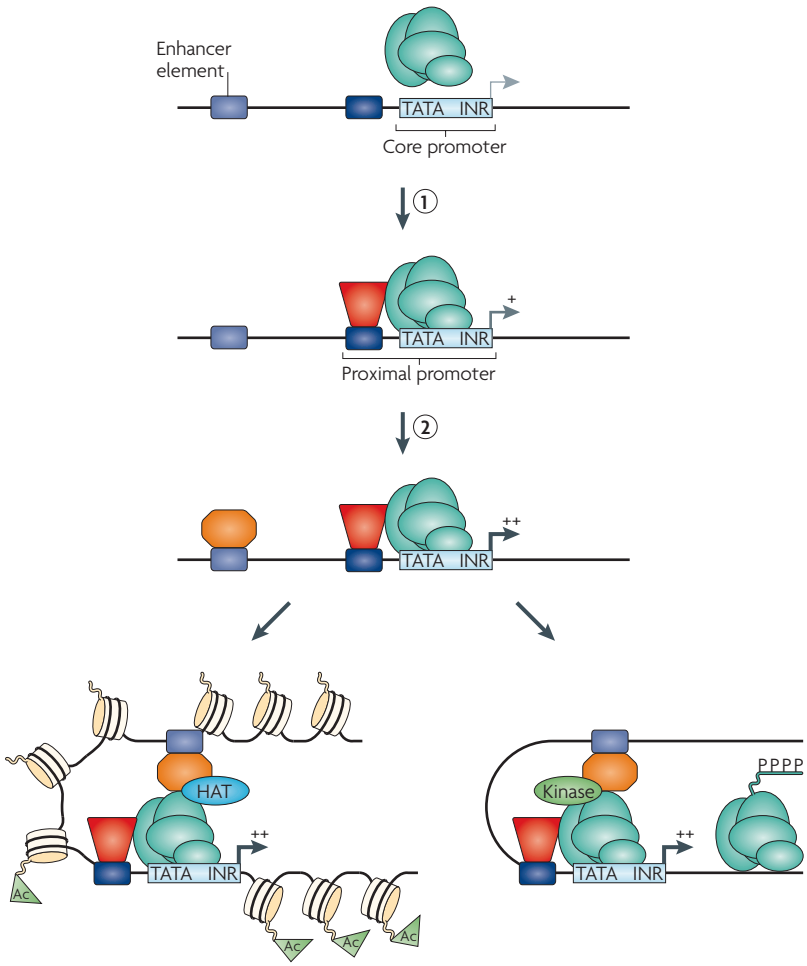


Figure 1 | **Transcriptional regulation by promoters and enhancers.** General transcription factors (green ovals) bind to core promoter regions through recognition of common elements such as TATA boxes and initiators (INR). However, these elements on their own provide very low levels of transcriptional activity owing to unstable interactions of the general factors with the promoter region. Promoter activity can be increased (represented by +) by site-specific DNA-binding factors (red trapezoid) interacting with *cis* elements (dark blue box) in the proximal promoter region and stabilizing the recruitment of the transcriptional machinery through direct interaction of the site-specific factor and the general factors (step 1). Promoter activity can be further stimulated to higher levels by site-specific factors (orange octagon) binding to enhancers (step 2). The enhancer factors can stimulate transcription by (bottom left) recruiting a histone-modifying enzyme (for example, a histone acetyltransferase (HAT)) to create a more favourable chromatin environment for transcription (for example, by histone acetylation (Ac)) or by (bottom right) recruiting a kinase that can phosphorylate (P) the carboxy-terminal domain of RNA polymerase II and stimulate elongation.

Silencer
A DNA sequence capable of binding transcription factors that are termed repressors, which can negatively influence transcription by preventing recruitment of the general transcriptional machinery or by recruiting histone-modifying complexes that create repressive chromatin structures.

of complexes, such as transcription factor II D (TFIID), that are required for transcription of most protein-coding genes. In addition, there are ~1,400 transcription factors that have sequence-specific DNA-binding properties and thus regulate only a subset of genes by binding to site-specific *cis* elements[6–8]. Interestingly, the site-specific factors tend to be expressed either in all (or most) tissues or in one or two tissues, suggesting either a very broad or very specific function[7]. Alterations in gene expression caused by the inappropriate level, structure or function of a transcriptional regulator have been associated with a diverse set of human diseases, including cancers and developmental disorders[9]; for example, 164 transcription factors have been shown to be directly responsible for 277 diseases[7]. This is undoubtedly a large underestimation of the importance of transcription factors in human disease, because most human transcription factors are uncharacterized[7]. Owing to the paucity of our knowledge concerning the function of transcription factors and the likelihood that increased knowledge of transcription factors will lead to increased insight into the causes of human diseases, it is of utmost importance that we expand our understanding of how site-specific transcription factors contribute to gene regulation. Crucial questions that need to be addressed are: where do transcription factors bind in the genome, how is specificity of binding achieved, what features of chromatin influence the ability of transcription factors to stably interact with the genome, and how is binding of a transcription factor related to its subsequent function in respect to regulation of a nearby gene?

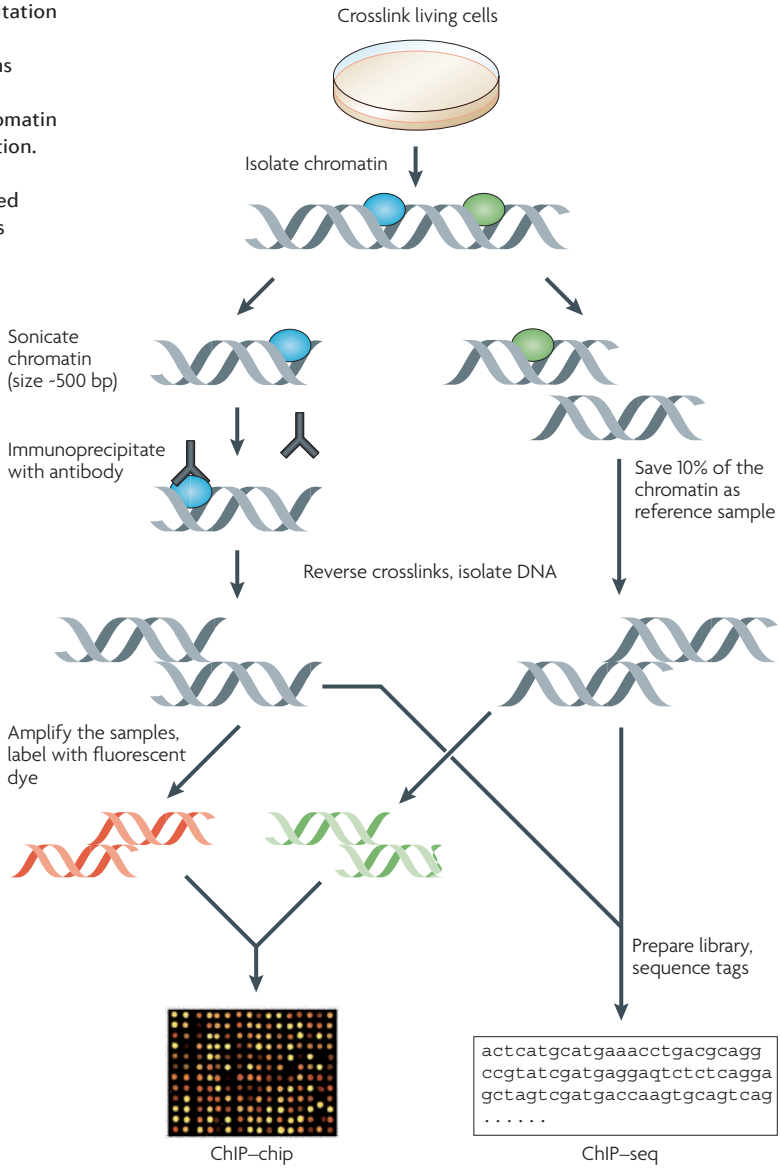
Fortunately, recent advances in chromatin immunoprecipitation followed by microarray (ChIP–chip) or by sequencing (ChIP–seq) (BOX 1), and similar techniques such as DamID, have allowed investigators to create a global map of specific protein–DNA interactions in a given cell type in a single experiment[10–19]. Binding sites identified from these ChIP studies[20–28] are categorized relative to genomic features such as the nearest gene, the frequency of binding relative to gene structure (for example, binding to a promoter, enhancer, exon or intron) and the type of chromatin domain. The cost of ChIP–seq depends partly on the depth of sequencing, but an estimate is that 10–12 million uniquely mapped reads should be sufficient for most human transcription factors, and this can be obtained in 1 or 2 lanes of sequencing for a cost of US$1,000–2,000. Because multiple DNA microarrays are needed to cover the entire human genome, comprehensive studies by ChIP–chip are more expensive. However, for certain applications (such as detailed analyses of a protein complex binding to a small segment of a genome), a focused ChIP–chip experiment currently remains more cost-effective than a genome-wide ChIP–seq analysis.

This Review summarizes recent discoveries provided by genome-wide profiling of site-specific transcription factors and how they have led to new insights regarding patterns of transcription factor binding. I also discuss how binding specificity of transcription factors is achieved and what features of chromatin influence the ability of transcription factors to interact stably with the genome. The focus is on the human genome, although relevant insights from other organisms are also incorporated (in particular when studies using model organisms are more advanced than similar studies of the human genome), as it is likely that the implications of transcription factor recruitment for gene regulation will be similar across all eukaryotes. Importantly, in addition to providing new information, genome-wide studies have challenged our understanding of gene regulation, raising questions such as: why do certain transcription factors bind to so many places in the genome, and why does so much of the regulation seem to be through steps that occur after recruitment

Box 1 | **Chromatin immunoprecipitation methods**

Briefly, chromatin immunoprecipitation (ChIP) (see the figure) involves crosslinking DNA-binding proteins to DNA by treating cells with formaldehyde and preparing chromatin by sonication or enzymatic digestion. An immunoprecipitation of the crosslinked chromatin is performed using an antibody that recognizes a specific transcription factor or histone isoform, which results in the identification of all the binding sites in the genome for the factor of interest. After purification of the precipitated fragments, the sample can be analysed by PCR to study particular genes. However, genome-wide analysis can be performed by microarray (ChIP–chip) or sequencing (ChIP–seq).

For ChIP–chip, the immunoprecipitated sample and input DNA, as a control, are labelled with fluorescent dyes and hybridized to microarrays. Binding sites are identified by the intensity of the signal of the immunoprecipitated sample in relation to the signal of the input DNA sample at each probe on the microarray using various ChIP–chip peak-calling programs[21,22]. For a single ChIP–chip experiment, most investigators use between 10^6 and 10^7 cells; however, recent methodological improvements using amplification methods have enabled successful ChIP–chip experiments with as few as 10^4 cells[76–79].



For ChIP–seq, the immunoprecipitated sample is used to create a library that is analysed using high-throughput next-generation sequencers. Binding sites are identified using various ChIP–seq peak-calling programs[16,26,27,81,82], all of which identify target sites based on the number of sequenced tags from the ChIP library corresponding to each position in the genome. For a ChIP–seq experiment designed to map binding of a site-specific factor, most investigators use 10^7 to 10^8 cells, although 10^4 to 10^5 cells is sufficient for the ChIP–seq analysis of certain histone modifications[82].

It is important to note that because ChIP assays require such large numbers of cells, the observed peaks in either ChIP–chip or ChIP–seq represent an average level of binding of a factor at a particular site in the cell population. Thus, a small peak could represent very strong binding in only a subset of the cells (for example, cells at one stage of the cell cycle) or modest binding in the entire cell population. ChIP–seq experiments, which allow binding to be analysed at all unique overlapping oligomers of a certain length (usually 27–50 nucleotides) in the genome, can provide very high resolution mapping of transcription factor-binding sites. For example, three-quarters of all the ChIP–seq peak positions for the DNA-binding proteins CCCTC-binding factor (CTCF), neuron-restrictive silencer factor (NRSF) and signal transducer and activator of transcription 1 (STAT1) are within 18, 27 and 51 bp, respectively, of the nearest motif for that factor[81]. In general, genome-scale ChIP–chip experiments are less precise in mapping the exact location of a binding site because the oligomers on the array are not overlapping; if overlapping oligomers were used, a prohibitively large number of arrays would be required, so the oligomers are instead spaced approximately 35–100 nucelotides apart.

Transcription factor II D
A protein complex composed of several subunits, called TATA binding protein (TBP)-associated factors (TAFs), and the TBP. It is one of several complexes that make up the RNA polymerase II initiation machinery.

DamID
An alternative method to chromatin immunoprecipitation that uses a DNA-binding protein fused to a DNA methyltransferase. Adenine methylation of a region identifies it as being located near a binding site.

of the site-specific factor to the DNA? Therefore, this Review concludes with suggestions for future experiments that are needed to further our understanding of the causes and consequences of specific transcription factor–genome interactions.

Localization of binding sites

Two decades ago, investigators were using *in vitro* assays or reporter constructs to define the *cis* elements that are necessary for basal transcriptional activity and the regions that control cell-type-specific, hormonal or environmental transcriptional responses. In most cases, small promoter segments (from 500 bp to ~10 kb upstream of a TSS) were used as the starting point for mutational analyses. One common observation was that severe truncation of a fragment could cause large changes in promoter activity but that incremental deletion of the 5′ end of the fragment resulted in only minor changes in activity, suggesting that multiple transcription factor binding sites were scattered throughout the analysed region (for example, see REF. 29). By contrast, other studies found that hormonal regulation or cell-type-specific transcription from a promoter could not be reproduced using reporter assays (for example, see REF. 30). Such results raised two important questions that are now being addressed by genome-wide binding analyses: do different transcription factors bind in clusters near each other, and are most of the binding sites for a given transcription factor located in proximal promoter regions?

Binding to proximal promoters. Transcription factors have been categorized into those that bind proximal promoters and those that bind enhancers[1,2]. However, in most analyses, a single binding site, or in some cases a small set of sites, was studied for a particular factor. Such focused analyses do not allow general conclusions to be drawn as to whether a factor usually binds near or distal to a promoter region. Thus, accurate categorization of factors is not possible without genome-wide analysis of binding sites. Knowing the location, relative to the TSS, at which a factor binds is of interest as it can provide insight into the mechanisms by which the factor regulates transcription (FIG. 1). For example, factors that bind close to TSSs have been proposed to regulate transcription by stabilizing general transcription factors at the core promoter elements; factors that bind to distal regions, either upstream or downstream of a gene, may regulate transcription by mediating, through a looping mechanism, the protein–protein contacts between distal complexes and the general transcriptional machinery bound at TSSs. Thus, comprehensive analysis of the binding locations of a factor not only allows the development of a genomic map but also provides insight into the mechanisms by which the factor regulates transcription.

Initial large-scale ChIP–chip analyses of transcription factor binding focused on the identification of binding sites near CpG islands or within 1–5 kb of the TSS of known genes[15,31–34]. Although these studies identified hundreds, and in some cases thousands, of promoters that were bound by a particular transcription factor, they were limited to target sites in proximal promoter regions, so it was not known whether the identified sites were representative of the majority of the genomic binding sites for a given factor. Analyses of 1% of the human genome, which began as part of the ENCyclopedia Of DNA Elements (ENCODE) pilot project and are being continued by the ENCODE Consortium and others[3,22–24,35,36], have shown that transcription factors that bind almost exclusively at proximal promoters might be the exception, not the rule. Some factors, for example, E2F transcription factor family members, are almost always bound in proximal promoter regions (FIG. 2a). In fact, it is often difficult to distinguish E2F binding patterns from the binding patterns of general transcription factors, such as RNAPII or TATA box binding protein-associated factor 1 (TAF1)[15,22]. However, other factors that have recently been analysed by genome-wide ChIP–chip or ChIP–seq, such as GATA1 and zinc finger protein 263 (ZNF263), bind to diverse regions of the genome (FIG. 2b), including extragenic regions distant from the TSS and intragenic regions (including introns and exons). Other examples of transcription factors that have widespread binding patterns include p53, p63, the oestrogen receptor, forkhead box protein A2 (FOXA2) and transcription factor 4 (TCF4)[10,13,24,36,37].

Although it is difficult to make accurate comparisons of binding patterns generated by different research groups using different experimental platforms, genome-wide profiles for a large number of factors were compared in the ENCODE pilot project[3]. This study found that less than 10% of the factors tested had greater than 50% of their binding sites within 2.5 kb of a transcription start site (see figures in REF. 28). Another study, which analysed 13 site-specific factors in mouse embryonic stem cells using ChIP–seq, also found that many binding sites were located outside proximal promoter regions[38]. Clearly, a typical reporter or *in vitro* assay cannot monitor the contribution to promoter activity of sites distant from the proximal promoter. These new findings of the distribution of factors throughout the genome might explain many of the failed attempts to demonstrate accurate regulation of a target gene using reporter assays or transgenic constructs. Also, the distributive pattern of binding seen for many factors has important implications for subsequent functional analyses. For example, it is not easy to link enhancers to specific promoters if the enhancer is between two genes but at a great distance from both; this is discussed in more detail below.

Binding to enhanceosomes. Early studies of *Drosophila melanogaster* development identified regulatory regions that are bound by combinations of different transcription factors, which led to the concept that transcription factors can cluster near each other to regulate transcription cooperatively[39]. For example, enhancers that regulate *D. melanogaster* segmentation contain a module that typically receives input from multiple transcription factors and that has multiple binding sites for each of the factors; in many cases, the binding sites are clustered in a small interval of 0.5–1 kb. Recently, large-scale profiling of the binding patterns of a set of *D. melanogaster* transcription factors revealed binding hot spots, each 1–5 kb in length

Reporter construct
A plasmid containing a promoter (and sometimes an enhancer) cloned upstream of a reporter gene (often simply called the reporter) that is introduced into cultured cells, animals or plants. Certain genes are chosen as reporters because their products can be easily or quantitatively assayed, or used as selectable markers.

CpG island
A sequence of at least 200 bp with a greater number of CpG sites than expected for its GC content. These regions are often GC rich and usually undermethylated. They correspond to the promoter regions of many mammalian genes.

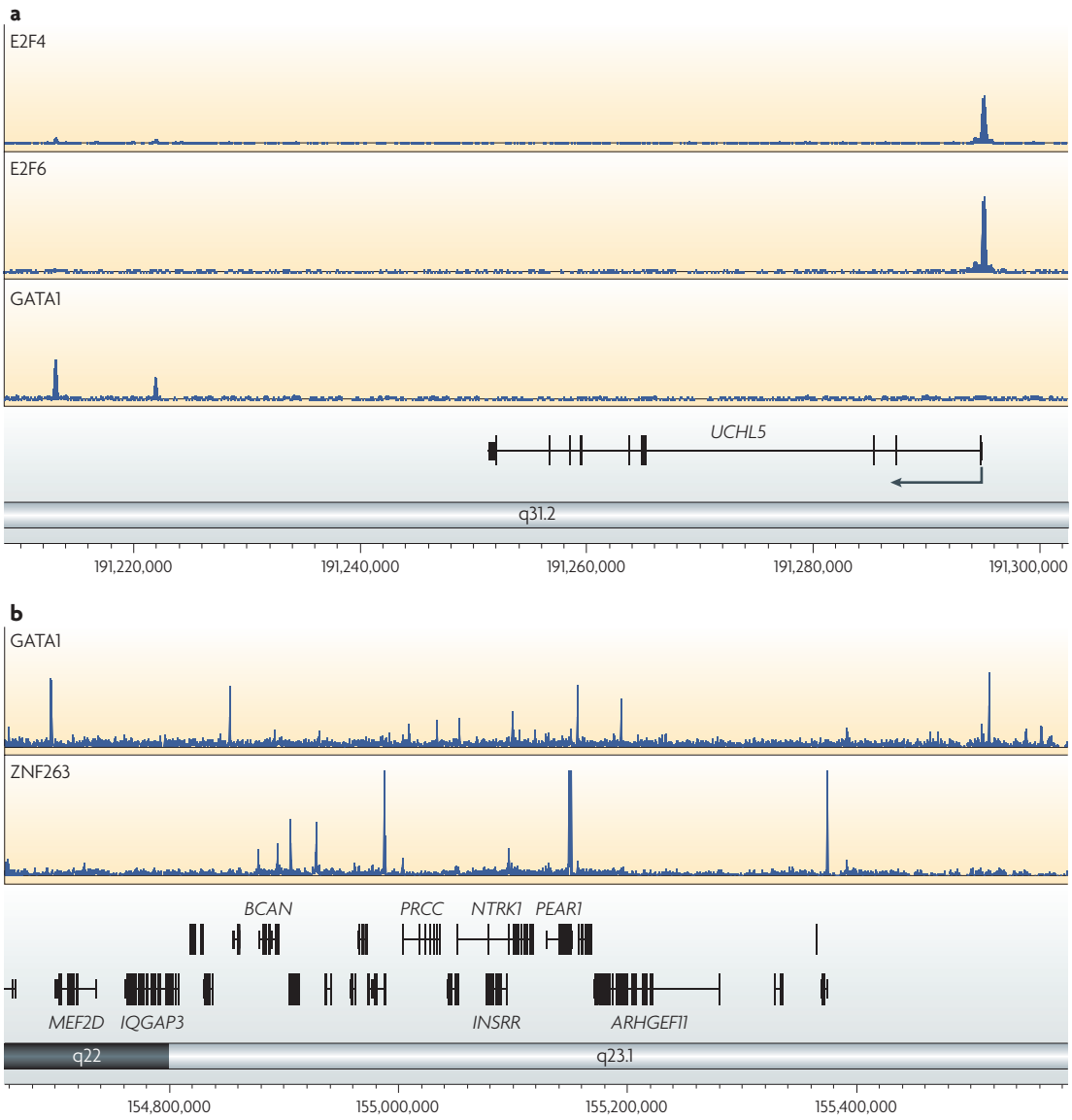


Figure 2 | **Location analysis of transcription factors.** Localization analysis reveals two classes of binding patterns for transcription factors. **a** | Binding sites identified using chromatin immunoprecipitation followed by sequencing (ChIP–seq) for E2F4, E2F6 and GATA1 in a region of chromosome 1 containing the ubiquitin carboxy-terminal hydrolase L5 (*UCHL5*) gene (the direction of transcription is shown by the arrow beginning at the start site). E2F4 and E2F6 bind to the promoter region, whereas GATA1 binds downstream of the gene. **b** | Binding sites identified using ChIP–seq for GATA1 and zinc finger protein 263 (ZNF263) for a region of chromosome 1. The binding sites for these two factors do not cluster at the same genomic locations. *ARHGEF11*, Rho guanine nucleotide exchange factor 11; *BCAN*, brevican; *INSRR*, insulin receptor-related receptor; *IQGAP3*, IQ motif containing GTPase activating protein 3; *MEF2D*, myocyte enhancer factor 2D; *NTRK1*, neurotrophic tyrosine kinase, receptor, type 1; *PEAR1*, platelet endothelial aggregation receptor 1; *PRCC*, papillary renal cell carcinoma (translocation-associated).

Enhanceosome
A protein complex that binds to an enhancer region (which can be located upstream, downstream or in a gene); the transcription factors that compose the enhanceosome are thought to work cooperatively to stimulate transcription.

and spaced ~50 kb apart[40]. The *D. melanogaster* genome is one-tenth the size of the human genome and therefore it is not yet clear whether the same sort of clustering will be commonly found for human transcription factors.

Owing to the large size of the human genome and the large number of transcription factors (~1,400), most investigations into the concept of clustered binding sites creating a regulatory element have used computational tools[41]. As detailed below, bioinformatic analyses are not sufficient to determine which of all possible binding sites are actually occupied by a transcription factor *in vivo*. However, there is some experimental evidence that at least a few binding hot spots do exist in the human genome. An extensively studied mammalian enhancer is the interferon-β enhanceosome[42,43], in which eight transcription factors bind to overlapping elements within a 55 bp region upstream of the interferon-β gene (*IFNB1*). This enhancer was characterized over many years using classical mutational analyses of a single regulatory element. Although very few regions of the human genome

have been characterized in as much detail as the *IFNB1* enhancer, several other enhancer regions have been well studied, including the mouse and chicken β-globin locus control regions and the human growth hormone and major histocompatibility complex II (MHCII) enhancer regions[44].

Chen *et al.* analysed a set of factors that work together to mediate pluripotency and maintain the self-renewal properties of mouse embryonic stem cells[38]. They found that some regions, termed multiple transcription factor-binding loci (MTLs), were bound by several factors. Specifically, clusters of NANOG, OCT4 (also known as POU5F1) and SOX2 sites were identified outside promoter regions, which suggested that these regions might be enhancers, and a subset of MTLs showed strong enhancer activity in follow-up experiments. Identification of these MTLs might have been facilitated by the fact that NANOG, OCT4 and SOX2 were previously known to cooperate in regulating the mouse embryonic stem cell transcriptome.

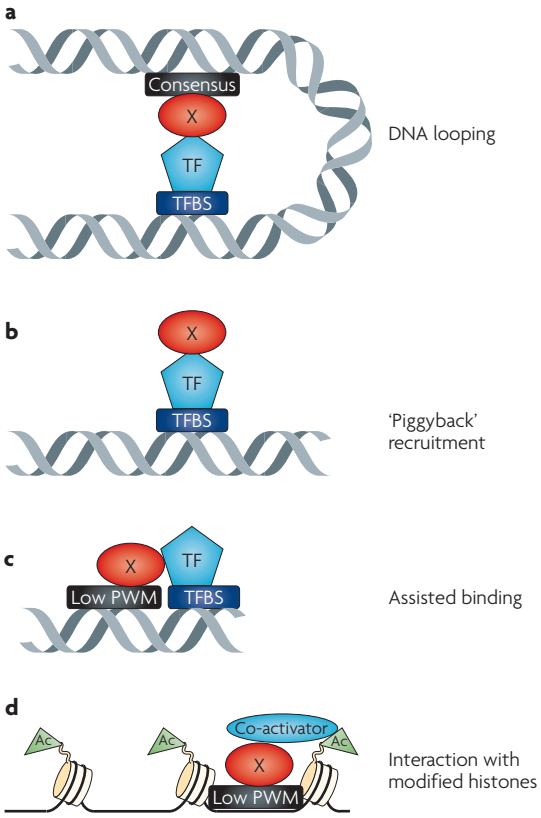


Figure 3 | **Models for recruitment of factors to sites that lack consensus motifs. a** | A transcription factor (X) could bind to its consensus motif and loop, as a result of protein–protein interactions, to another transcription factor (TF) bound to a different binding site (TFBS) that is located at a distant region of the chromosome. In this case, because formaldehyde can create both protein–DNA and protein–protein crosslinks, chromatin immunoprecipitation (ChIP) assays for factor X would enrich for a region containing its own consensus motif and a region bound by the other factor. **b** | Factor X could be recruited to a sequence by protein–protein interactions with another transcription factor in a manner completely independent of its DNA-binding abilities. In this case, ChIP assays would detect binding of factor X at a region that has no match to its consensus or position weight matrix (PWM). **c**,**d** | Factor X could bind to a sequence that has a low match to its PWM and be anchored on the genome by protein–protein interactions with a nearby factor (**c**) or be attached by interaction with a co-activator to a specifically modified — for example, acetylated (Ac) — histone (**d**). In both cases, ChIP assays would detect binding of factor X at a region that contains a low match to its PWM.

Unfortunately, only a handful of human factors (very few of which have been implicated in regulating the same sets of genes) have been analysed using ChIP–seq, and these factors do not seem to show a large degree of overlap in binding at locations outside promoter regions (FIG. 2b). However, it is hard to know whether the lack of observed clustering is due to a lack of hot spots for binding in the human genome or to the possibility that the correct combinations of factors have not yet been studied. Gaining knowledge of the extent of clustered binding in mammalian genomes requires the collection of more ChIP–seq data. Genome-wide analyses of enhancers based on specific histone modification patterns have also recently been initiated[45,46]. However, identifying a potential enhancer region based on histone patterns does not reveal how many site-specific factors bind to the region. If clusters of binding sites are found in mammalian genomes, they may correspond to enhancesomes that are similar to the one at *IFNB1*, in which multiple factors work together to mediate transcriptional activation. Alternatively, they may represent non-functional 'storage bins' for excess transcription factors, provide functional redundancy that decreases the chances that a gene may be turned off owing to mutation, or allow activation of a gene by multiple different signalling cascades.

Do consensus motifs specify binding?

In vitro studies, such as CASTing (cyclic amplification and selection of targets), and sequence comparisons of small sets of promoters known to be bound by a factor have allowed the derivation of consensus binding motifs for some transcription factors[47]. Subsequent bioinformatic analyses that search the human genome using consensus motifs or position weight matrices — a collection of motifs that are similar, but not identical, to the consensus motif — allow the identification of all locations in the genome to which a transcription factor might bind[41,48]. This approach provides the set of all possible locations for a given factor; however, in a mammalian genome there are clearly many more occurrences of a consensus motif for a given factor than there are binding sites[37,49]. Also, the utility of bioinformatic studies relies on the assumption that transcription factors are recruited to the genome *in vivo* by motifs similar to those identified in *in vitro* studies. These caveats have led to uncertainties as to the importance of consensus motifs for *in vivo* binding. ChIP–chip and ChIP–seq studies have allowed investigators to address two important questions concerning motif usage: what percentage of binding sites contain a consensus motif, and what influences whether a specific motif is bound by a particular factor?

Motif enrichment in binding regions. Although some factors seem to be recruited to a majority of their binding sites by a common motif, other factors seem to have a more diverse set of recruitment mechanisms. For example, members of the E2F family do not seem to require a specific motif for binding *in vivo*[49]. By contrast, the binding sites for factors such as p63, signal transducer and activator of transcription 1 (STAT1) and neuron-restrictive silencer factor (NRSF, also known as REST) show high enrichment for a specific motif[16,20,37]. It should be stressed that binding detected at sites that lack a consensus motif is not caused by a general, low-affinity DNA-binding activity. ChIP–chip and ChIP–seq measure DNA–protein interactions as an average of individual binding events in millions of cells, and a peak at a site without a motif can be as high and as sharp as a peak located over a consensus motif, which is inconsistent with random protein–DNA interaction.

Several mechanisms have been proposed to explain how recruitment of a specific transcription factor can occur in the absence of a consensus motif (FIG. 3). These include: binding at a distal site that contains a consensus motif and looping to the site in question through protein–protein interactions (perhaps through a co-activator or corepressor); 'piggyback' binding that is mediated by protein–protein interactions with a second factor and that does not involve the DNA-binding domain of the first factor; or assisted binding to a site that is similar to the consensus site, which is enhanced by protein–protein interaction with another site-specific DNA-binding factor or with a specifically modified histone. Clearly, the greater the contribution of protein–protein interactions to the genomic localization of a factor, the greater the difficulty of using a strictly bioinformatic approach for identifying *in vivo* binding sites.

Sorting binding sites for a factor into subsets that contain or lack a specific motif might eventually provide insight into alternative recruitment or regulatory mechanisms mediated by that factor; the ability of a factor to be recruited to the genome in more than one way might allow it to participate in multiple different signalling pathways. For example, serum response factor (SRF) is ubiquitously expressed, but its activity is modulated at several levels, including protein–protein interaction[50,51]. Perhaps recruitment of SRF by a consensus motif allows the regulation of one set of targets in many cell types, whereas stabilized binding, mediated by protein–protein interaction, to sites lacking the consensus motif allows the constitutively expressed SRF to also have some cell-type-specific functions. It should be noted that even factors that prefer to bind to regions containing a specific motif can also have subsets of binding sites that lack that motif[52,53]. A recent study has shown that the ability of a factor to bind to more than one motif is not necessarily attributable to protein–protein interactions; the same property can be seen for purified proteins in *in vitro* assays. Using protein binding microarrays, Badis *et al.*[54] found that approximately half of a set of 104 mouse DNA-binding proteins recognized multiple different sequence motifs. Such studies suggest that motif analysis of ChIP–seq data should be performed under the assumption that more than one motif can be present in the set of identified binding regions.

Do epigenetic modifications influence motif usage? As discussed above, a major difficulty with using a bioinformatics motif-driven approach for identifying binding sites is that it is clear that only a small percentage of all occurrences of a motif are actually bound by that factor. Therefore, the majority of regions in the genome that contain a consensus motif for a given factor are not occupied. Lack of binding in certain regions of the genome could be a consequence of the chromatin structure (the close packing of nucleosomes in heterochromatin might make binding sites inaccessible) or of DNA methylation (methylation of a crucial residue in the recognition motif might result in reduced binding affinity). However, in a study of unoccupied E2F consensus sites in a human breast cancer cell line, neither repressive histone modifications (that is, histone H3 trimethylated on lysine 9 or lysine 27) nor DNA methylation seemed to account for the lack of E2F binding[49]. An alternative possibility is that specific histone modifications enhance transcription factor recruitment to certain genomic regions. For example, recent ChIP–chip and ChIP–seq studies have shown that histone H3 monomethylated at lysine 4 (H3K4me1) is localized at enhancer regions[45,46]. It is not known whether the histone modification or the binding of a factor comes first, but it is possible that certain factors might have an affinity for a specific histone modification. For example, plant homeodomain finger (PHD finger) domains in several proteins, such as the TAF3 subunit of TFIID, bromodomain PHD finger transcription factor (BPTF) and inhibitor of growth family, member 2 (ING2), can mediate a specific high-affinity interaction with histone H3 trimethylated on lysine 4 (REFS 55–58), which is highly localized to promoter regions[3,46]. PHD domains in site-specific factors or co-activators may help to localize DNA-binding factors to consensus motifs located in proximal promoters; other domains may mediate interactions of transcription factors or co-activators with H3K4me1, which may result in preferential occupancy of motifs located in enhancer regions (FIG. 3d).

Although each of the models presented in FIG. 3 is possible, it is generally not clear why some consensus motifs are occupied and others are not. Once we have binding maps for hundreds of factors, it may become obvious that binding of a factor to one motif commonly prevents another motif from being occupied by a different factor. For example, an ETS1 and an E2F binding site overlap in the *MYC* promoter, and it is only after mutation of the E2F site that ETS1 can bind *in vivo*[59]. Alternatively, as described above, we might find that stable binding is rarely mediated by a single DNA–protein interaction and requires cooperative binding between adjacent site-specific factors, which may be achieved by either direct interaction between the two site-specific factors or indirect interaction through a platform such as a co-activator or corepressor[60].

Are all occupied binding sites important?

The discovery of thousands of binding sites by genome-wide profiling has raised two important questions: can a factor occupy a certain site in many cell types but regulate transcription by binding to that site in only one (or a few) cell types, and is functional redundancy a built-in safeguard for maintaining accurate regulation of the genome?

Heterochromatin
Chromatin that is characterized by very dense packing of DNA, which makes it less accessible to transcription factors. Certain regions of the genome, such as centromeres and telomeres, are always heterochromatinized (constitutive heterochromatin regions), whereas other regions are densely packed and repressed only in certain cells (facultative heterochromatin regions).

DNA methylation
An epigenetic DNA modification that can be added and removed without changing the original DNA sequence and that is characterized by the addition of a methyl group to the number 5 carbon of the cytosine pyrimidine ring.

Plant homeodomain finger
A 50–80 amino acid domain that contains a Cys4-His-Cys3 motif. It is found in more than 100 human proteins, several of which are involved in chromatin-mediated gene regulation.

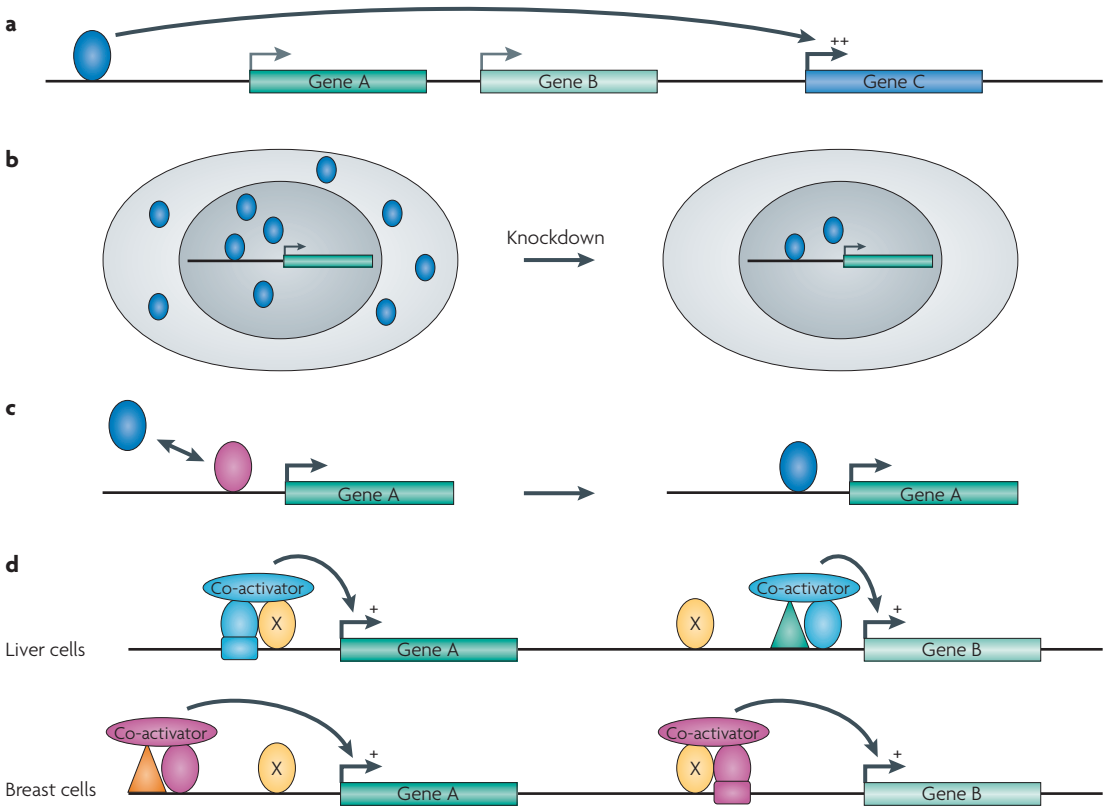


Figure 4 | **Incorrect interpretation of functional assays.** There are a number of reasons, other than a lack of function, why reduction in the level of a transcription factor might not result in a change in expression of the predicted target gene. **a** | The transcription factor (dark blue oval) regulates gene C, which is distal to the binding site; therefore, genes A and B will not show a change in expression following knockdown of the factor, even though they are nearer to the transcription factor than gene C. **b** | Knockdown of a factor (dark blue ovals) with a small interfering RNA (not shown) does not lower the level below that needed for full binding site occupancy; therefore, expression of target genes is not affected. **c** | Knockdown of a factor (pink oval) results in full occupancy by another family member (dark blue oval) at a site that, under normal conditions, is bound interchangeably by both family members; expression of the target gene (A) is not affected because the family members are redundant in function. **d** | Regulation is dependent on the ubiquitous site-specific factor (factor X) in combination with cell-type-specific factors. In this example, factor X is bound to the promoter regions of gene A and gene B in both liver and breast cells, and genes A and B are expressed in both tissues. However, in liver cells, factor X is not involved in regulation of gene B because there is no binding site for the liver-specific factor (light blue oval) near the factor X binding site in the gene B promoter. Conversely, in breast cells, factor X regulates gene B through interaction with the breast-specific factor (pink oval) but does not regulate gene A because there is no binding site for the breast-specific factor near the factor X site in the gene A promoter. Thus, different subsets of target genes may show changes in expression in different cell types when levels of the ubiquitous site-specific factor are reduced. The triangles represent other site-specific factors that cooperate with the liver- or breast-specific factors to activate transcription.

Understanding gene expression data. Several recent studies have attempted to assess the functional importance of each of the thousands of binding sites for a given factor by altering the level of that factor in the cell. A frequent finding is that changing the level of a factor alters the expression level of 1–10% of the potential target genes[12,37,61,62]. One interpretation of these results is that most binding is not functional. There are, however, several caveats to this conclusion.

First, the assignment of a specific binding site to a target gene is not always accurate. Investigators use the most expedient approach, which is to assign the binding site to the nearest known gene, but this can lead to incorrect assumptions in cases of long-range regulation, undiscovered genes or alternative upstream promoters. Changes in the expression level of a gene that does not have a nearby binding site for the factor that is altered might initially be interpreted as indicative of indirect regulation, but might be due to direct regulation by a site many thousands of kilobases away (FIG. 4a).

Second, altering the expression of a human transcription factor is fraught with problems.Downregulation of a transcription factor in human cells is usually accomplished using small interfering RNAs or short hairpin RNAs. However, loss of expression is rarely complete; it is possible that reducing the level of a transcription factor by 90% may not have functional consequences if there is a tenfold excess of the factor under normal conditions.

Small interfering RNAs
Small antisense RNAs (20–25 nucleotides) that can be directly introduced into cells or be generated in cells from longer dsRNAs. They serve as guides for the cleavage of homologous mRNA in the RNA-induced silencing complex.

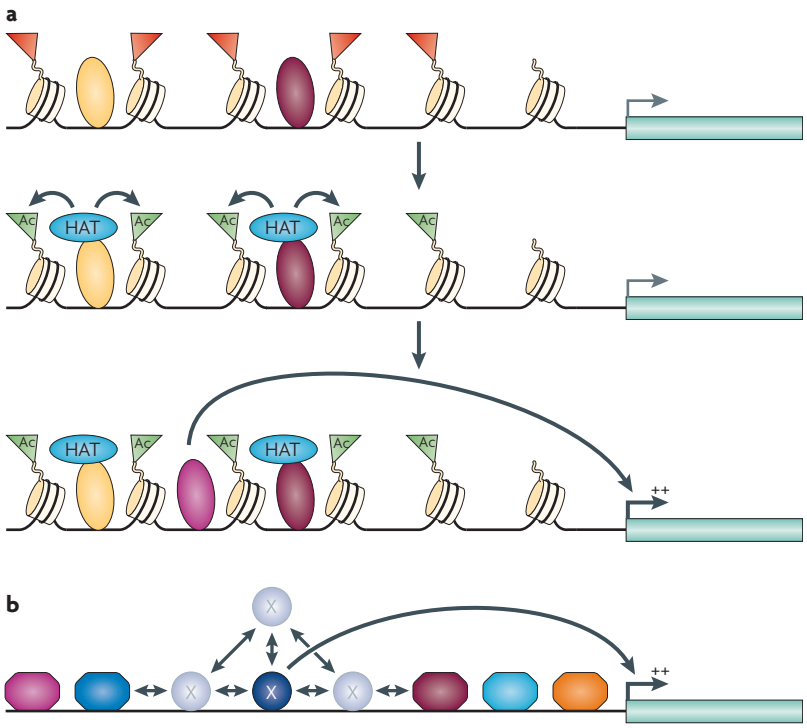


Figure 5 | **Communal action of a set of transcription factors. a** | A possible scenario in which two different factors (large yellow and dark red ovals) can bind near to each other on inactive chromatin (represented by the orange triangles) and each recruit a histone acetyltransferase (HAT), which acetylates histones (Ac) and creates an open chromatin region (green triangles). This allows the binding of another transcription factor (pink oval) that stimulates transcription of a gene (++). In this case, the loss of a single factor that recruits a HAT would not result in a major change in regulation of the gene. **b** | A possible scenario in which multiple factors (octagons) bound on either side of factor X (circle) can create a limited search domain for factor X (which is required for activation of a downstream gene). Factor X binds transiently to its binding site; dissociation from the site is followed by localized rebinding and scanning for the high-affinity binding site. Transcriptional activation can be enhanced if the scanning is spatially limited by adjacent clusters of other bound factors; loss of a single factor in the cluster would not result in a major change in regulation of the gene.

Many studies are performed in cancer cell lines that can have, as shown by western blot, a massive increase in the amount of a particular transcription factor compared with a normal cell. Thus, what seems to be an efficient knockdown in a cancer cell line may leave sufficient levels of the factor for normal regulation (FIG. 4b). Very few studies have actually shown a reduced level of binding of a transcription factor in knockdown cells by ChIP–chip or ChIP–seq. To overcome this problem, mouse knockouts can be used. However, cells from these mice could undergo compensation for loss of a factor during development, which might result in related proteins being selected to regulate the target genes.

Third, closely related family members might bind to the same sites and have the same function. Thus, elimination of one family member could allow a higher level of binding of another family member (FIG. 4c). Finally, only a small proportion of the binding sites for a factor might be functional in a given cell type. For example, if a cell-type-specific partner needs to be recruited for transcriptional activity, then binding of the site-specific factor is necessary but not sufficient for transcription of a target gene (FIG. 4d). Thus, knockdown of a factor in ten cell types may show ten different subsets of affected target genes. To address this possibility, one would have to collect ChIP–chip or ChIP–seq data and gene expression data before and after knockdown of the factor in a diverse set of cell lines. However, most transcription factors have been studied on a genome-wide scale in only one cell type. The ENCODE Consortium has chosen a set of different cell types for thorough characterization of binding of a large number of site-specific factors, and initial studies seem to show that factors can be grouped into those that show very little cell type specificity in binding, such as E2F4 and YY1 (H. O'Geen and P.J.F., unpublished observations), and those that show considerable cell-type-specific binding, such as JunD (D. Raha and M. Snyder, personal communication) and the oestrogen receptor[15,63]. Continuing studies will address whether factors that have small numbers of cell-type-specific binding sites show regulation of a large percentage of their target genes in a given cell type compared with factors that show constitutive binding to a large number of sites and might regulate only a subset of target genes in each cell type.

Functional redundancy in clusters. Many previous analyses of transcriptional regulation assumed that transcription factors act as 'individuals' with each factor having a specific role in regulating a particular gene and a specific mechanism of action. However, a factor might act as an individual at a subset of its sites (perhaps those that show altered regulation of a nearby gene following loss of or enhanced expression of that factor) but have a very different 'community' function at other sites. For example, binding of a set of factors in a cluster might regulate transcription throughout a chromatin domain by helping to keep an open chromatin structure through recruitment of histone acetyltransferases or histone methyltransferases. Loss of a single factor would not affect transcription of the nearby genes; it would take the removal of a large proportion of factors bound in the cluster to alter gene regulation (FIG. 5a). Alternatively, a cluster of bound factors could serve to define a local genomic search space for a second binding factor. Recent studies have shown that many transcription factors have a very fast dissociation rate *in vivo*[64]. A factor might rebind to the same region of DNA but in a non-specific manner and begin scanning for its high-affinity binding site. If the factor moves unimpeded in the wrong direction, there could be a detrimental time lag before it finds another binding site. However, a cluster of bound factors that blocks scanning in the wrong direction might favour release, rebinding and perhaps scanning in the correct direction. That is, binding of a cluster of factors might affect the expression of a nearby gene that is controlled by an entirely different factor. Again, reduced expression of one of the 'bumper proteins' may be fairly inconsequential; loss of several factors from the cluster would be required to cause a significant effect (FIG. 5b). Data to support either of these possibilities are not yet available owing to the lack of genome-wide binding information for most transcription factors.

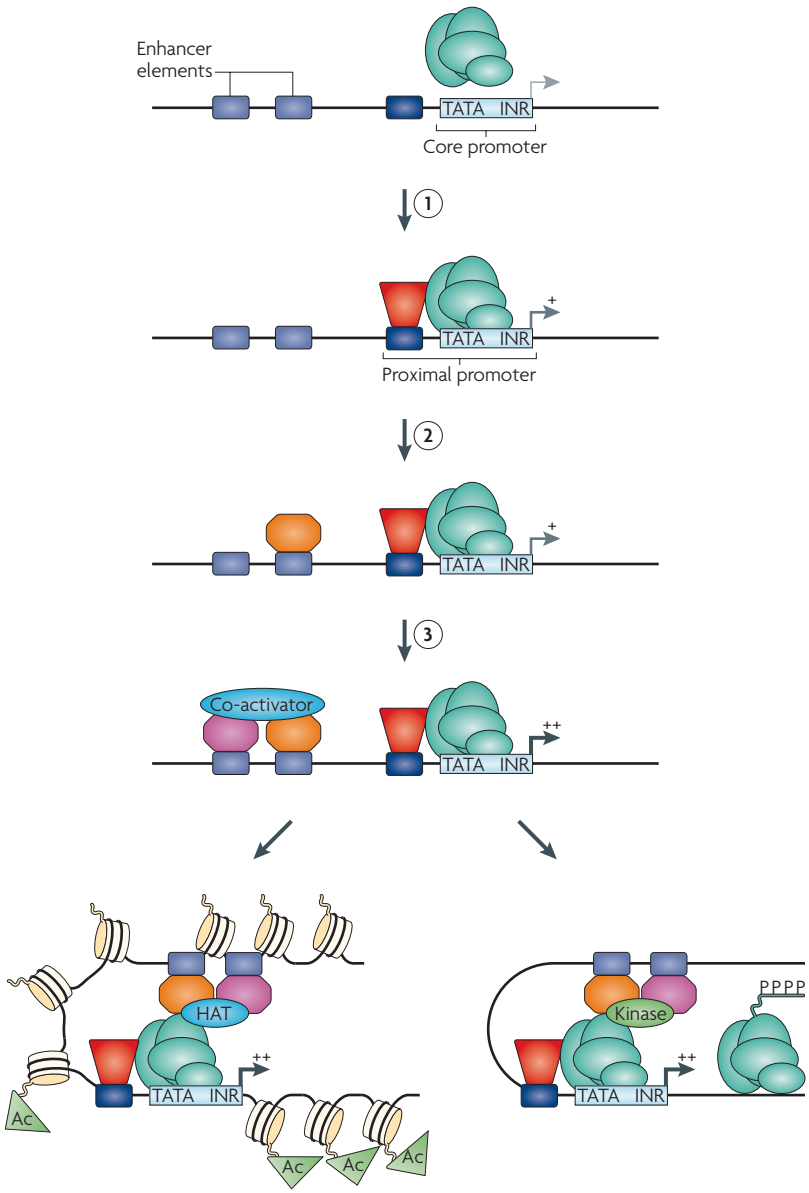


Figure 6 | **Revised model for transcriptional regulation.** Studies using chromatin immunoprecipitation followed by microarray (ChIP–chip) or by sequencing (ChIP–seq) have confirmed that RNA polymerase II (RNAPII) and other general transcription factors (green ovals) bind to thousands of promoter regions at elements such as TATA boxes and initiators (INR) and provide low levels of transcriptional activity (also see FIG. 1). This provides support for step 1, in which promoter activity can be increased by the interaction of site-specific DNA-binding factors (red trapezoid) with *cis* elements (dark blue box) in the proximal promoter region, which stabilizes the recruitment of transcriptional machinery through direct interaction between the site-specific factor and the general factors. Promoter activity can be further increased by the binding of a site-specific factor (orange octagon) to an enhancer region (step 2). However, ChIP–chip and ChIP–seq studies have revealed that step 2 is not sufficient for high levels of promoter activity, and thus a new step has been proposed: the binding of a cell-type-specific partner protein (pink octagon) that allows the recruitment of a co-activator, which provides a constitutively bound factor with a cell-type-specific function (step 3). Currently, the projected later steps remain as shown in FIG. 1: the enhancer factors can stimulate transcription by (bottom left) recruiting a histone-modifying enzyme to create a more favourable chromatin environment for transcription or (bottom right) recruiting a kinase that can phosphorylate (P) the carboxy-terminal domain of RNAPII and stimulate elongation. Ac, acetylated histone; HAT, histone acetyltransferase.

Next steps for genomic landscapes

Significant progress has been made in mapping transcription factor binding sites throughout the genome, and expanding the number of transcription factors for which we have information about global binding patterns is important; however, simply collecting genome-wide data sets will not be sufficient to answer all of the crucial questions. A number of methodological problems now need to be tackled.

Accurate target gene assignment. It is not yet possible to conclusively link a specific binding site with a specific target gene. It remains possible that many binding sites, which may be scattered tens or hundreds of kilobases away from each other (or perhaps even on different chromosomes), all cooperate to regulate a single target gene. If so, linking a binding site to the nearest gene is not appropriate and will lead to both an incorrect assignment of target genes and an underestimation of the number of binding sites that contribute to transcriptional regulation. Methods that define features of chromosomal architecture, such as transcription factories[65,66], could aid in identifying coregulated groups of genes, perhaps by collapsing thousands of seemingly unlinked binding sites into a smaller number of interactomes. For example, chromosome conformation capture (3C), a technique that can identify chromosomal loops mediated by multiple long-range protein–protein interactions[67], may reveal a connection between an enhancer binding protein and the promoter of a distant gene and thereby allow a more accurate interpretation of the regulatory role of that factor in the cell.

Comprehensiveness. Although ChIP–seq can identify all the binding sites for a given factor in a given cell type, researchers may still face the daunting challenge of performing ChIP–seq experiments in many different cell types to determine all possible binding sites for a given factor. The ENCODE Consortium is currently performing studies to estimate how many cell types are needed to identify most binding sites for a set of factors. If a limited, but diverse, set of cell types can be identified that is representative of many different human tissues, then genome-wide analyses may not have to be performed in every possible cell type.

Functional analysis of specific regulatory elements. Most approaches that are designed to study the relationship between a specific *cis* element and a potential target gene involve creating a reporter construct that includes the regulatory element of interest[4,68]. Unfortunately, because reporter analyses remove the *cis* element from its normal genomic context, they cannot reveal effects on long-range regulation. Precise mutation or deletion of a single *cis* element in the genome can be performed in model organisms such as yeast, for which efficient methods for substituting genomic sections have been developed.

Theoretically, mutations could be engineered to alter a specific binding site in animal models or human cell lines. However, mutagenesis of specific small regions of the mouse or human genome is not routinely used to study the significance of individual binding sites owing

Transcription factory
A nuclear subcompartment that is rich in RNA polymerases and transcription factors, and in which there is clustering of active genes.

Interactome
A complete set of macromolecular interactions (physical and genetic). Current use of the word tends to refer to a comprehensive set of protein–protein interactions. However, the protein–DNA interactome (a network formed by transcription factors and their target genes) is also commonly studied.

Artificial zinc finger
Chimaeras of zinc finger domains — small protein domains that coordinate one or more zinc ions and that are commonly found in mammalian transcription factors — and an effector domain (for example, an activator, repressor, methylase or nuclease). Linking together six zinc fingers produces a target site of 18 bp, which is long enough to be unique in all known genomes.

to low frequencies of homologous recombination that limit the efficiency of this technique. New approaches in site-specific targeting of DNAses using artificial zinc fingers[69] might improve the efficiency of genomic replacement, so mutagenesis could become a practical method for dissecting the role of individual *cis* elements. Furthermore, artificial zinc fingers fused to transcriptional activation or repression domains have been used to specifically regulate cellular promoters[70]. It is therefore possible that artificial zinc fingers (without either an activation or repression domain) could be used to simply block access of a factor to a single binding site in the genome, but this has not yet been demonstrated successfully. Other possible methods include the use of pyrrole-imidazole polyamides or peptide nucleic acids to bind to (and perhaps also mutate) specific *cis* elements in the genome[71–73]. Although few studies have used these methods to target a specific site, and even fewer have examined the consequences of such agents on the entire transcriptome, they do hold the promise of providing a method for testing the function of a specific binding site in its natural genomic context.

Conclusions

ChIP–chip and ChIP–seq have greatly advanced our understanding of gene regulation. First, genomic studies have confirmed that RNAPII and general and site-specific factors are bound to thousands of proximal promoters that are active at very low levels[73–75], thus supporting the first step in the model set out at the beginning of this Review. These studies have also revealed that binding of a factor to an enhancer region can be necessary, but not sufficient, for high levels of promoter activity, which leads to the inclusion of a new step in the model (FIG. 6, step 3): the binding of a cell-type-specific partner protein that allows the recruitment of a co-activator, which results in cell-type-specific functioning of a constitutively expressed factor. Although the principle that binding of a transcription factor can be necessary, but not sufficient, for regulation of a specific gene was previously established using 'one-gene-at-a-time' approaches, it was not clear whether a cooperative mode of regulation was the exception or the rule for most genes. Recent genome-wide analyses suggest that this type of regulation is common. For example, of the ~3,700 OCT4, ~4,500 SOX2 and ~10,000 NANOG binding sites identified in mouse embryonic stem cells, only a small number of regions were bound by all three factors and by the co-activator p300 (REF. 38). These studies support the hypothesis that the occupancy of an upstream site by a single factor (OCT4) has no functional effect (as in FIG. 6, step 2), but binding of SOX2 and/or NANOG near to the occupied OCT4 site resulted in recruitment of the p300 co-activator (FIG. 6, step 3) and transcriptional activation.

Other discoveries have also stimulated new ideas concerning long-range and combinatorial regulation. These include the findings that most transcription factors bind to thousands of places in the genome, that binding sites are not localized only in proximal promoter regions and that some binding sites lack sequences similar to the consensus motif. However, current genomic studies have not yet determined whether most transcription factors cluster at hot spots in the human genome or with what frequency binding events have a functional outcome. The answers to these two questions will require the genomic profiling of many more factors. It is likely that a true understanding of the role of a given factor at a particular site in the genome will require the identification of all other factors binding nearby and knowledge of histone modifications in that region. These studies will be best performed by cooperation between large groups and individual investigators. The groups — such as the ENCODE Consortium and the NIH Roadmap Epigenomics Program — can identify binding sites for a large number of transcription factors and develop reference epigenomes in many different cell types. Individual investigators can then perform the follow-up functional analyses of the role of a specific factor in a particular cell type. The next several years of large-scale data collection should provide investigators with a plethora of information that will form the basis for hundreds of follow-up experiments that address important biological questions.

1. Lee, T. I. & Young, R. A. Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.* **34**, 77–137 (2000).
A detailed review of transcriptional regulation, general factors and accessory proteins that control transcription initation and elongation.
2. Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Rev. Genet.* **8**, 424–436 (2007).
3. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
This paper demonstrates how genome-wide studies of transcription factor binding, chromatin structure, DNA replication and sequence conservation can synergize.
4. Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L. & Myers, R. M. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.* **16**, 1–10 (2006).
5. Kimura, K. *et al.* Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* **16**, 55–65 (2006).
6. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
7. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Rev. Genet.* **10**, 252–263 (2009).
A summary of the expression, conservation and activity of the set of human sequence-specific transcription factors.
8. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
9. Jimenez-Sanchez, G., Childs, B. & Valle, D. Human disease genes. *Nature* **409**, 853–855 (2001).
10. Wederell, E. D. *et al.* Global analysis of *in vivo* Foxa2-binding sites in mouse liver using massively parallel sequencing. *Nucleic Acids Res.* **36**, 4549–4564 (2008).
11. Reed, B. D., Charos, A. E., Szekely, A. M., Weissman, S. M. & Snyder, M. Genome-wide occupancy of SREBP1 and its partners NFY and SP1 reveals novel functional roles and combinatorial regulation of distinct classes of genes. *PLOS Genet.* **4**, e1000133 (2008).
12. Scacheri, P. C. *et al.* Genome-wide analysis of menin binding provides insights to MEN1 tumorigenesis. *PLoS Genet.* **2**, e51 (2006).
13. Hatzis, P. *et al.* Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. *Mol. Cell. Biol.* **28**, 2732–2744 (2008).
14. O'Geen, H. *et al.* Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. *PLoS Genet.* **3**, e89 (2007).
15. Xu, X. *et al.* A comprehensive ChIP–chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals iterchangeable roles of E2F family members. *Genome Res.* **17**, 1550–1561 (2007).
16. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**, 1–7 (2007).
An early demonstration that high-throughput sequencing of ChIP samples can be used to identify genome-wide binding sites of site-specific transcription factors.
17. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316**, 1497–1502 (2007).
18. Rada-Iglesias, A. *et al.* Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. *Genome Res.* **18**, 380–392 (2008).
19. Vogel, M. J., Peric-Hupkes, D. & van Steensel, B. Detection of *in vivo* protein–DNA interactions using DamID in mammalian cells. *Nature Protoc.* **2**, 1467–1478 (2007).

20. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP–seq data. *Nature Methods* **5**, 829–834 (2008).
21. Johnson, D. S. *et al.* Systematic evaluation of variability in ChIP–chip experiments using predefined DNA targets. *Genome Res.* **18**, 393–403 (2008).
22. Bieda, M., Xu, X., Singer, M., Green, R. & Farnham, P. J. Unbiased location analysis of E2F1 binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res.* **16**, 595–605 (2006).
A demonstration that some factors bind exclusively to proximal promoters and do not have strict motif requirements for their binding sites.
23. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).
An early demonstration that high-density oligonucleotide arrays can be used to identify genome-wide binding sites for human transcription factors.
24. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
25. Nix, D. A., Courdy, S. J. & Boucher, K. M. Empirical methods for controlling false positives and estimating confidence in ChIP–seq peaks. *BMC Bioinformatics* **9**, 523 (2008).
26. Zhang, Y. *et al.* Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9**, R137 (2008).
27. Fejes, A. P. *et al.* FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24**, 1729–1730 (2008).
28. Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP–seq experiments relative to controls. *Nature Biotech.* **27**, 66–75 (2009).
29. Liu, Y., Michalopoulos, G. K. & Zarnegar, R. Structural and functional characterization of the mouse hepatocyte growth factor gene promoter. *J. Biol. Chem.* **269**, 4152–4160 (1994).
30. Fujishiro, K. *et al.* Analysis of tissue-specific and PPARα-dependent induction of FABP gene expression in the mouse liver by an *in vivo* DNA electroporation method. *Mol. Cell. Biochem.* **239**, 165–172 (2002).
31. Weinmann, A. S., Yan, P. S., Oberley, M. J., Huang, T. H.-M. & Farnham, P. J. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev.* **16**, 235–244 (2002).
32. Li, Z. *et al.* A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl Acad. Sci. USA* **100**, 8164–8169 (2003).
33. Ren, B. *et al.* E2F integrates cell cycle progression with DNA repair, replication, and G_2/M checkpoints. *Genes Dev.* **16**, 245–256 (2002).
34. Odom, D. T. *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378–1381 (2004).
35. Consortium, T. E. P. The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* **306**, 636–640 (2004).
36. Carroll, J. S. *et al.* Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* **122**, 33–43 (2005).
37. Yang, A. *et al.* Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. *Mol. Cell* **24**, 593–602 (2006).
38. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
39. Mann, R. S. & Carroll, S. B. Molecular mechanisms of selector gene function and evolution. *Curr. Opin. Genet. Dev.* **12**, 592–600 (2002).
40. Moorman, C. *et al.* Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **103**, 12027–12032 (2006).
This demonstrates clustering of transcription factors throughout the *D. melanogaster* genome.
41. Elnitski, L., Jin, V. X., Farnham, P. J. & Jones, S. J. M. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res.* **16**, 1455–1464 (2006).
42. Panne, D. The enhanceosome. *Curr. Opin. Struct. Biol.* **18**, 236–242 (2008).
43. Maniatis, T. *et al.* Structure and function of the interferon-β enhanceosome. *Cold Spring Harb. Symp. Quant. Biol.* **63**, 609–620 (1998).
44. Dean, A. On a chromosome far, far away: LCRs and gene expression. *Trends Genet.* **22**, 38–45 (2006).
45. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
This identifies specific histone modifications that are associated with cell-type-specific transcriptional regulation.
46. Heintzman, N. D. *et al.* Distinct predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39**, 311–318 (2007).
47. Wright, W. E. & Funk, W. D. CASTing for multicomponent DNA-binding components. *Trends Biochem. Sci.* **18**, 77–80 (1993).
48. Morgan, X. C., Ni, S., Miranker, D. P. & Iyer, V. R. Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining. *BMC Bioinformatics* **8**, 445 (2007).
49. Rabinovich, A., Jin, V. X., Rabinovich, R., Xu, X. & Farnham, P. J. E2F *in vivo* binding specificity: comparison of consensus versus non-consensus binding sites. *Genome Res.* **18**, 1763–1777 (2008).
50. Gineitis, D. & Treisman, R. Differential usage of signal transduction pathways defines two types of serum response factor target gene. *J. Biol. Chem.* **276**, 24531–24539 (2001).
51. Cooper, S. J., Trinklein, N. D., Nguyen, L. & Myers, R. M. Serum response factor binding sites differ in three human cell types. *Genome Res.* **17**, 136–144 (2009).
52. Jin, V. X., O'Geen, H., Iyengar, S., Green, R. & Farnham, P. J. Identification of an OCT4 and SRY regulatory module using integrated computational and experimental genomics approaches. *Genome Res.* **17**, 807–817 (2007).
53. Li, X.-Y. *et al.* Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol.* **6**, e27 (2008).
54. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
55. Li, H. *et al.* Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* **442**, 91–95 (2006).
56. Pena, P. V. *et al.* Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. *Nature* **442**, 100–103 (2006).
57. Shi, X. *et al.* Proteome-wide analysis in *Saccharomyces cerevisiae* identifies several PHD fingers as novel direct and selective binding modules of histone H3 methylated at either lysine 4 or lysine 36. *J. Biol. Chem.* **282**, 2450–2455 (2007).
58. Vermeulen, M. *et al.* Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131**, 58–69 (2007).
59. Albert, T. *et al.* The chromatin structure of the dual c-*myc* promoter P1/P2 is regulated by separate elements. *J. Biol. Chem.* **276**, 20482–20490 (2001).
60. Jin, V. X., Rabinovich, A., Squazzo, S. L., Green, R. & Farnham, P. J. A computational genomics approach to identify *cis*-regulatory modules from chromatin immunoprecipitation microarray data — a case study using E2F1. *Genome Res.* **16**, 1585–1595 (2006).
61. Krig, S. R. *et al.* Identification of genes directly regulated by the oncogene ZNF217 using chromatin immunoprecipitation (ChIP)–chip assays. *J. Biol. Chem.* **282**, 9703–9712 (2007).
62. Martone, R. *et al.* Distribution of NF-κB-binding sites across human chromosome 22. *Proc. Natl Acad. Sci. USA* **100**, 12247–12252 (2003).
63. Krum, S. A. *et al.* Unique ERα cistromes control cell type-specific gene regulation. *Mol. Endocrinol.* **22**, 2393–2406 (2008).
64. Voss, T. C. & Hager, G. L. Visualizing chromatin dynamics in intact cells. *Biochem. Biophys. Acta* **1783**, 2044–2051 (2008).
65. Osborne, C. S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genet.* **36**, 1065–1071 (2004).
66. Bartlett, J. *et al.* Specialized transcription factories. *Biochem. Soc. Symp.* **73**, 67–75 (2006).
67. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
68. Visel, A. *et al.* ChIP–seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
This study demonstrates that using ChIP–seq to identify binding sites for p300 is a highly accurate method for identifying enhancers that can be shown, using follow-up assays in transgenic mice, to function in a tissue-specific manner.
69. Camenisch, T. D., Brilliant, M. H. & Segal, D. J. Critical parameters for genome editing using zinc finger nucleases. *Mini Rev. Med. Chem.* **8**, 669–676 (2008).
70. Bletran, A., Liu, Y., Parikh, S., Temple, B. & Blancafort, P. Interrogating genomes with combinatorial artificial transcription factor libraries: asking zinc finger questions. *Assay Drug Dev. Technol.* **4**, 317–331 (2006).
71. Faruqi, A. F., Egholm, M. & Glazer, P. M. Peptide nucleic acid-targeted mutagenesis of a chromosomal gene in mouse cells. *Proc. Natl Acad. Sci. USA* **95**, 1398–1403 (1998).
72. Burnett, R. *et al.* DNA sequence-specific polyamides alleviate transcription inhibition associated with long GAA·TCC repeats in Friedreich's ataxia. *Proc. Natl Acad. Sci. USA* **103**, 11497–11502 (2006).
73. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77–88 (2007).
74. Muse, G. W. *et al.* RNA polymerase is poised for activation across the genome. *Nature Genet.* **39**, 1507–1511 (2007).
75. Komashko, V. M. *et al.* Using ChIP–chip technology to reveal common principles of transcriptional repression in normal and cancer cells. *Genome Res.* **18**, 521–532 (2008).
76. Acevedo, L. G. *et al.* Genome-scale ChIP–chip analysis using 10,000 human cells. *Biotechniques* **43**, 791–797 (2007).
77. O'Neill, L. P., VerMilyea, M. D. & Turner, B. M. Epigenetic characterization of the early embryo with a chromatin immunoprecipitation protocol applicable to small cell populations. *Nature Genet.* **38**, 835–841 (2006).
78. Dahl, J. A. & Collas, P. Q2ChIP, a quick and quantitative chromatin immunoprecipitation assay, unravels epigenetic dynamics of developmentally regulated genes in human carcinoma cells. *Stem Cells* **25**, 1037–1046 (2007).
79. Attema, J. L. *et al.* Epigenetic characterization of hematopoietic stem cell differentiation using miniChIP and bisulfite sequencing analysis. *Proc. Natl Acad. Sci. USA* **104**, 12371–12376 (2007).
80. Xu, H., Wei, C.-L., Lin, F. & Sung, W.-K. An HMM approach to genome-wide identification of differential histone modification sites from ChIP–seq data. *Bioinformatics* **24**, 2344–2349 (2008).
81. Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of *in vivo* protein–DNA binding sites from ChIP–seq data. *Nucleic Acids Res.* **36**, 5221–5231 (2008).
82. Hoffman, B. G. & Jones, S. J. Genome-wide identification of DNA–protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing. *J. Endocrinol.* **201**, 1–13 (2009).

Acknowledgements
The author thanks X. Xu, H. O'Geen and S. Frietze for providing data used in figure 2 and the members of the Farnham laboratory for their insights and discussions.

Competing interests statement
The author declares competing financial interests: see Web version for details.

DATABASES
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
IFNB1 | *UCHL5*
UniProtKB: http://www.uniprot.org
FOXA2 | GATA1 | NANOG | NRSF | OCT4 | oestrogen receptor | p53 | p63 | SOX2 | SRF | STAT1 | TAF1 | TCF4 | ZNF263

FURTHER INFORMATION
Peggy J. Farnham's homepage: http://www.genomecenter.ucdavis.edu/farnham
ENCODE Data Coordination Center at UCSC: http://www.genome.ucsc.edu/ENCODE
The ENCODE Project: http://www.genome.gov/10005107
NIH Roadmap Epigenomics Program: http://nihroadmap.nih.gov/epigenomics/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF