Published in final edited form as:

Subcell Biochem. 2011 ; 52: 175–191. doi:10.1007/978-90-481-9069-0_8.

Identification of Transcription Factor-DNA Interactions *in vivo*

Duncan T Odom

University of Cambridge Cancer Research UK

Abstract

Recent technological developments have revolutionized our understanding of transcriptional regulation by providing an unprecedented ability to interrogate *in vivo* transcription factor binding. The combination of high-throughput sequencing with chromatin precipitation of transcription factors and specifically labeled histones has allowed direct protein-DNA contacts to be visualized across genomes as large and complex as mammals at base-pair resolution. This chapter reviews the developments that led to these insights, with particular focus on examples of early protein-DNA localization experiments using genomic microarrays in mammals and yeast. Four state of the art research directions are highlighted as examples of previously unimaginable frontiers that are now under active investigation.

Keywords

Protein-DNA contact; chromatin immunoprecipitation; transcription; transcriptional regulation; gene expression

7.1 The biochemistry and regulation of the genome

Living organisms use nucleic acids to store genetic information compactly and reliably. Although ribonucleic acids (RNA) were likely used in the earliest life forms to store genetic information [1], virtually all present-day organisms from prokaryotes to mammals use deoxyribonucleic acid (DNA) to store the instructions for creating proteins and RNAs. DNA is a highly negatively charged polymer that self-assembles into a stacked double helix of unsaturated nitrogenous bases (Figure 1A). This helix has a major groove with substantial potential for hydrogen bonding and van der Waals interactions to other molecules, with a somewhat less information-rich minor groove; structural implications for site-recognition were in part recognized in the initial reports of the DNA double-helical structure [2-4] (Figure 1B).

The proper deployment of the instructions found in DNA is controlled by a specialized class of proteins called transcription factors (TFs), numbering from scores in bacteria to thousands in vertebrates (reviewed in [5]) and higher plants [6, 7] (see **Chapters 2 and 3** for description and enumeration of TFs in bacteria and eukarya). These proteins have evolved specialized amino acid surfaces to interact most often with the DNA major groove (an example is shown in Figure 1A). Proteins accomplish this by using highly positively charged residues, directed hydrogen bonding, and architectural van der Waal interactions to select and bind to specific biochemical structures formed by the exposed surfaces of stacked base-pairs within the DNA major groove [8]. These major groove interactions can be

augmented by specific contacts in the minor groove (for instance, polyamides [9]). The DNA sequence preferences of individual TFs can be determined *in vitro* (see **Chapter 8**), but *in vivo*, the precise locations where TFs bind can be influenced by a variety of other factors also present in the cell, including nucleosomes and other TFs. This is presumably due to competition and cooperation, which can be either direct or indirect (see **Chapter 9**). Therefore, examining the genomic locations bound by TFs is critical to understanding the functions of the individual TFs as well as the operational principles of genomes.

7.2 Early approaches to identifying TF binding locations: nuclease sensitivity

One of the earliest methodologies used to identify where TFs bind DNA exploits the protection that a protein-DNA contact affords DNA when exposed to an exonuclease, in particular DNase I [10]. A bound TF physically protects stretches of DNA of varying base length from digestion, depending on the context and other possible neighboring TFs bound to the same promoter region.

This procedure has been used extensively to explore the structure of the hepatitis promoter, beta-globin [11], interleukins [12], many liver-specific gene classes [13, 14] (and others), and many prokaryotic genes. The utilization of this method is reviewed also in **Chapter 8**.

Originally, this technique referenced results against a combination of Maxim-Gilbert sequencing reactions to identify the sites protected *in vitro* against digestion. The same general approach has been adapted recently in *in vivo* work with isolated mammalian nuclei to identify regions protected from DNase digestion, methods that have been coupled with the methods described below for TF binding assays to obtain global views of open versus closed chromatin [15]. When combined with bioinformatic analysis of the protected sequences, these methods can suggest which TFs bind to a particular regulatory region (for instance, [16]).

7.3 Antibodies specifically targeting DNA-binding proteins allowed identification of *in vivo* binding events

The use of formaldehyde to crosslink proteins to nucleic acids was reported in the 1960s for ribonucleotides [17], and continues to be used in numerous RNA-protein identification protocols [18-20]. The ability of formaldehyde to reversibly crosslink proteins with DNA gradually evolved from work with SV40 minichromosomes and nucleosomes (for example [21, 22]).

By using antibodies against specific TFs of interest, the DNA regulatory regions bound *in vivo* can be isolated as nucleic acids, and then further interrogated, a process known as chromatin immunoprecipitation (ChIP). Historically, the enrichment of particular TFs at specific sites has been established using pairs of oligonucleotide primers at pre-selected promoter region(s) (Figure 2). Direct comparison can be made of the number of copies of a potentially bound region versus random and unbound regions in the genome by

simultaneous amplification of these regions, followed by gel electrophoresis and quantitation of the nucleic acid bands.

7.4 Microarrays first allowed the genome-wide determination of TF binding in the yeast *Saccharomyces cerevisiae*

In the early 2000s, ChIP experiments were combined with the then-nascent technology of microarrays. The most popular method for gene expression microarray synthesis in the first years of the technology was to PCR-amplify mRNA sequences, print them onto glass slides, and fix chemically. Gene expression arrays had been successfully reported to interrogate yeast [23, 24] and mammalian gene expression in tissues [25, 26] and in response to stimuli [27] (and many, many other publications on gene expression). Since the early days of gene expression microarray analysis, scores of species have had microarrays designed to interrogate gene expression.

In principle, ChIP experiments such as those described above simultaneously isolate and enrich all promoter regions that are bound by a protein, even if only a small subset are interrogated for ChIP enrichment using specific primers. This fact led to a number of groups realizing that one possible method for obtaining genome-wide information on TF-DNA binding would be to create promoter-sequence microarrays, as opposed to coding-sequence gene expression microarrays (Figure 2). Synthesis of these microarrays was combined with methods to fluorescently label ChIP DNA one color and input (or a mock ChIP experiment performed without the specific antibody) DNA a second color, followed by co-hybridization against the promoters present on the promoter microarrays. The creation of promoter microarrays that tile the noncoding regions of the yeast genome was greatly facilitated by the extraordinarily dense yeast genome. In contrast to higher eukaryotes, yeast has few repeated sequences. When combined with ChIP experiments, the use of a microarray to identify TF binding became a technique that quickly gained the name *ChIP-chip*.

The first reports of genome-wide TF binding *in vivo* were both reported in yeast, essentially simultaneously [28, 29]. Three major genetics research groups were active in this then-nascent field, and used as a proof-of-principle TFs that had been well-studied by yeast transcriptional biologists for years. Richard Young's laboratory at the Whitehead Institute used Ste12 and Gal4, both tagged with a myc epitope and induced with either mating hormone (Ste12) or galactose (Gal4), to perform genome-wide location analysis using an anti-myc antibody [29]. In addition, to showing that this technology yielded results consistent with site-specific analysis, these authors were able to identify a number of novel components of the carbon metabolic pathways involved in galactose utilization, as well as new mating genes regulated by Ste12.

The Snyder and Brown laboratories, at Yale and Stanford respectively, collaboratively performed ChIP-chip experiments against two TFs active at the G1/S transition in yeast cell cycle [28]. On a technical basis, their experiments demonstrated that *in vivo* tagging of yeast TFs (here, SBF and MBF) afforded the same results as antibodies raised specifically against the TF proteins themselves. In addition, their study demonstrated for the first time the power of using TF binding and systematic gene expression perturbations as independent

methodologies to deconstruct on a genome-wide basis the direct versus indirect effects of TF binding. As noted below, and in other chapters in this book, the relationships between TF binding and expression output can be complex. Notably, similar approaches in mammals have revealed that transcription factor binding rarely if ever follows the intuitively predicted expectation of direct regulation (Text Box 1)—nevertheless, few researchers, even in the field of transcription, appear to have absorbed this lesson

Shortly thereafter, the Brown laboratory also reported the genome-wide binding of the Rap1 protein [30]. Among their notable observations was the large-scale binding of this protein to the telomeric regions of yeast, as well as to the ribosomal protein genes. This link between a site-specific TF that regulates a large number of coding regions in the genome with regions containing noncoding RNA was among the first of many major surprises in investigations utilizing genome wide TF binding.

7.5 Combining multiple TF binding experiments on the same platform allows network elaboration

The first systems-biology application of the ChIP-chip technology was to deconstruct the combinatorial transcriptional networks active in the model organism *Saccharomyces cerevisiae*. It has been long known that TFs at specific promoters bind to and regulate gene expression combinatorially. Two publications reporting large numbers of TF binding experiments exploiting a carefully matched yeast genetic system were reported within a year and half of the first report of ChIP-chip as a technique [31, 32].

One publication pioneered a comprehensive systems-wide characterization of the transcriptional regulatory networks active in yeast by controlling the growth condition (YPD media), genetic background (yeast strain), microarrays, and experimental protocols and reagents [32]. Over a hundred independent yeast strains containing epitope-tagged TFs were created and grown, followed by ChIP-chip analysis. Two crucial observations made in the simplest of eukaryotes indicated that (1) on a genome-wide basis, TFs bind to promoters in a highly combinatorial manner, and (2) that regulators not only bound transcriptional regulators in the same functional class, but that the interconnections between TFs and the promoters of other TFs often crosses functional classes (Figure 3). For instance, cell cycle regulators were found to bind to the promoters of TFs that had well-characterized metabolic functions.

A simultaneous publication pioneered a different strategy of characterizing regulatory networks combinatorially by identifying other regulators downstream of a known TF's binding (MBF), followed by characterization of their genome-wide binding and further analysis [31]. The G1/S transition was carefully dissected by identifying the TFs downstream of MBF, tagging these with a biochemical handle, and performing matched ChIP-chip experiments at the same point of the cell cycle.

7.6 The vast size and repeat content of mammalian genomes provoked four different approaches to microarray analysis of TF binding

Mammalian genomes like human and mouse can be over a hundred times the size of yeast genomes, as well as being composed of over 50% repeated sequences found at high density in introns, promoters, enhancers, gene deserts--basically everywhere outside coding regions. These facts greatly complicated the original assemblies of the human and mouse genomes [33, 34].

When considering strategies to interrogate TF binding, the enormous size and complex, repeat-laden sequence content were also very complicating. Four microarray strategies were simultaneously pioneered to interrogate on a genome-wide scale human TF binding and the histone mark patterns obtainable from ChIP experiments: (1) CpG island microarrays, (2) PCR-based promoter microarrays, (3) PCR-based chromosome tiling microarrays, (4) short oligonucleotide chromosomal tiling microarrays.

(7.6.1) *CpG island microarrays*

Many mammalian regulatory regions have an enrichment in CpG dinucleotides, and as such are named CpG islands. One strategy to capture on a genome-scale TF binding in mammals was to PCR amplify specifically the regions of the mammalian genome that contain CpG islands for printing onto a microarray platform. Such a microarray was first reported in combination with E2F4 ChIP experiments to reveal scores of newly discovered binding events across the genome [35].

(7.6.2) *Promoter region PCR microarrays*

A substantial fraction of transcriptional regulation has long been known to occur within the first few kilobases of transcription start sites in mammals. This fact was exploited to create PCR products against non-repeated sequence found within the first two kilobases of potentially regulatory sequence upstream of transcription start sites across the human genome, which were then used to create microarrays. The first report using this form of microarray characterized the cell-cycle promoter regions bound by E2F1 in *in vitro* human fibroblast cells [36]. A somewhat later report used a slightly larger platform that represented the promoters from most human genes, yet reported for the first time ChIP experiments performed against disease-related TFs in primary human tissues, here liver and pancreatic islets [37].

(7.6.3) *Regional and chromosomal tiling PCR microarrays*

The first microarrays designed to represent entire regions [38] and chromosomes [39] were created by using PCR to amplify overlapping, or nearly overlapping DNA regions, followed by printing onto microarrays. The first chromosome-wide interrogation of TF binding was performed against NFkB sites across human chromosome 22, and was the first genome-scale experimental evidence that TFs bind extensively far from gene loci. This observation led eventually to similar pioneering discoveries in tissue-specific transcriptional regulation, including with Estrogen Receptor [40] and Androgen Receptor [41].

(7.6.4) Short oligonucleotide chromosomal tiling microarrays

Finally, it was quickly realized that commercially available Affymetrix microarrays designed to tile human chromosomes for other purposes could be adapted to ChIP studies. The first report exploiting this described the comparison of histone marks in human and mouse lung fibroblasts across microarrays that tiled human chromosomes 21 and 22, finding strong similarities in the histones found in a few selected human and mouse orthologous regions [42]. This study was among the first to focus on understanding histone marks; since then, similar studies are far more common than genome-wide TF binding studies, due to commercial availability of high affinity antisera.

7.6.5 Limitations of microarrays

Though the ability to print DNA sequences at high density onto slides was quickly developed, thus overcoming a major roadblock towards having a genome on a single microarray, other hurdles remained. Most important of these is the requirement to omit repeated regions from microarray designs, because of their extremely high cross-hybridization with multiple regions in the mammalian genome. In effect, typically half of a mammalian genome cannot be interrogated with microarrays regardless of the microarray technology being used, and this is an unavoidable design limitation. Finally, microarrays are by their nature designed to be species-specific, which is a significant limitation if a researcher wishes to investigate non-standard model organisms. For instance, how large could a market be for microarrays to interrogate butterfly gene expression [43]?

7.7 High-throughput sequencing has revolutionized interrogation of *in vivo* binding in complex higher genomes

A variety of methods have been developed to sequence nucleic acids (reviewed in [44]), and the most widely used method until recently was Sanger sequencing [45]. This methodology sequences stretches of hundreds of bases, which were painstakingly computationally sewn together using chromosomal maps to assemble the human [33] and mouse [34] genomes. Most placental mammals sequenced more recently have relied heavily on these early, carefully annotated mammalian genomes to facilitate the assembly of genomes (for instance [46-48], as synteny is widespread).

The recent development and rapid widespread adaptation of technology that uses millions of parallel sequencing reactions [49] is currently displacing these methods in techniques as diverse as genome resequencing, mRNA sequencing, and characterization of TF binding globally by directly sequencing nucleic acids bound by TFs (ChIP-seq), instead of microarray hybridization (ChIP-chip) (HTP sequencing is reviewed in [50], and different platforms that can be used as of 2008 is reviewed in [51]).

To interrogate TF binding globally, oligonucleotides between 20-100 bases are read from the end of the ChIP-isolated DNA, aligned to the genome, and compiled into a global picture of TF binding. Typical reactions use >10 million reads per experiment, and, importantly, resolve up to 90% of the mammalian genome by covering sufficient nonrepeat region to allow unambiguous mapping of the reads to occur. The speed of data acquisition, and

tremendous coverage of the mammalian genome, has revolutionized the frontiers of modern genomic biology.

7.8 Major new frontiers in mammalian transcription are possible with genome-wide interrogation of *in vivo* TF binding

The ability to identify the global binding of mammalian TFs has created new opportunities to revisit questions that have been raised over the last hundred years in biology. How does one genome create multiple tissues? What mechanistically directs this process? How do intra-species variations and evolution of different species interface? I present here three nascent research areas that exploit how HTP sequencing enhances our ability to answer ambitious research questions.

7.8.1 Variation of TF binding within species

Genetic differences drive phenotypic differences within individuals of a species. How these genetic differences impact the gene expression and transcriptional regulatory differences between species has been a long-standing question, and a number of microarray based experiments have characterized gene expression differences between different individual humans [52] and mice [53]. How the regulation of these differences is driven has more recently been addressed using human cell lines [54, 55].

The variation among humans in matched cell lines for the same polymerase machinery and TFs has been found to be remarkably high, on the order of 10-25% variation between individual human genotypes [54]. Closer inspection of inter-allelic regulation indicates that up to one in ten active chromatin sites can differ between the two chromosome copies within the same individual [55]. Taken together, these results suggest that the natural assumption of large-scale similarity between individuals within a species has important limitations. If up to a quarter of genomic regulation can vary between one individual and another, then assigning a consensus regulatory pattern becomes problematic; this observation has clear relevance to the ENCODE project's objectives, described below.

7.8.2 Evolution of TF binding among mammals

Using microarrays, my laboratory showed that TF binding has diverged significantly between human and mouse in matched tissues [56]. We showed using ChIP-chip on specially designed proximal promoter microarrays that a set of liver-specific TFs (FOXA2, HNF1A, HNF4A, HNF6) whose function, amino acid sequence, and targeted binding motif are highly conserved throughout mammals, change both their potentially targeted genes and also their binding locations globally. However, our early study was limited to the proximal promoters around 4,000 transcription start sites in human and mouse, purely due to the technical limitations imposed on our experiments by the extant microarray densities.

By performing similar genome-wide binding studies using ChIP-seq for two TFs (CEBPA and HNF4A) in five vertebrates, we unambiguously identified how each species has tens of thousands of binding events that are unique to each evolutionary lineage (Figure 4) [57]. The role of repeat elements could now be addressed, and it was shown that substantial numbers

of binding events in liver were found on repeat elements in each lineage. Similar results have been found also for the OCT4 and NANOG transcriptional regulators in mouse and human in embryonic stem cells [58] [59], indicating that in mammals, rapid evolution of transcriptional regulation is the general rule.

7.8.3 Comprehensive analysis of multiple cell lines from one species—the ENCODE project

Inspired by the success of the human, mouse, and other mammalian genome projects, the National Human Genome Research Institute have undertaken an effort to fully annotate all functional elements in the human genome [60] (see also <http://www.genome.gov>). The ENCyclopedia Of DNA Elements (ENCODE) pilot project began by exhaustive characterization of 1% of the human genome using different microarray platforms and many of the techniques reviewed in this and other Chapters in this book. More recently, with the advent of HTP sequencing, genome-wide analysis has displaced the more limited views obtainable by microarrays. Scores of laboratories have been funded to characterize a handful of human cell lines via TF binding profiles, gene expression, small RNA expression, chromatin structure, histone marks, proteomics, and CpG methylation, among other techniques.

7.8.4 Widespread adaptation of ChIP-seq experiments in the transcriptional community

Many laboratories have now adapted sequencing to interrogate the genome-wide binding of transcription factors in cultured cell lines and *in vivo*. Virtually all are (re-)discovering that (i) tens of thousands of transcription factor binding events exist in the mammalian genome, (ii) that TFs bind combinatorially in tissue-specific patterns, and (iii) that TF binding rarely controls gene expression in a direct or intuitive manner. By community agreement, most datasets are freely available from the European Nucleotide Archive at the EMBL EBI (<http://www.ebi.ac.uk/ena/>) and from the Gene Expression Omnibus in America at the NCBI (<http://www.ncbi.nlm.nih.gov/geo/>). The sequencing data in ChIP-seq experiments is literally growing exponentially, and will continue to pose unprecedented challenges in storage of such massive amounts of data, much less its analysis and biological interpretation.

7.9 Future directions

What will the future hold for understanding how transcriptional regulation operates in the vast mammalian genome? Clearly, efforts like the ENCODE consortium will clearly produce enormous insight into how one species can deploy a single genome to create hundreds to thousands of distinct cell types. Combining TF binding and chromatin status for all possible proteins and post-transcriptional modifications may allow the development of models with true predictive value. Though simple, this goal has eluded computational and experimental science for decades. Intriguingly, one species of placental mammal (mouse) can readily and accurately interpret the genetic instructions embedded in the sequence of another mammal (human) [61] (Figure 5), suggesting that an understanding of the underlying grammar is indeed an obtainable goal.

Nevertheless, the most likely result, as is always the case for high impact research areas, will be more questions. Many of these questions are predictable and clear extrapolations of the trends in genomics and proteomics. For instance, how does the three dimensional genome interact with the proteomic components of the rest of the cell? Other predictable directions may include controlled engineering of development: tailoring phenotypic traits using a soon-to-be-known regulatory toolbox in mammals. Many vertebrates can regenerate limbs using ancient developmental systems driven by TFs [62], which may be applicable to higher mammals that may have lost these functions.

Other questions will certainly be driven by unpredictable discoveries, much like the realization in the last years of the twentieth century that small RNAs can and do play integral regulatory roles in mammalian transcription. The ability to identify TF binding across the entire mammalian genome will be an important tool; one whose value will become fully apparent only after it is as routine and reliable as gene expression microarrays.

Literature cited

1. Gilbert W. Evolution of antibodies. The road not taken. *Nature*. 1986; 320(6062):485–6. [PubMed: 3960131]
2. Wing R, et al. Crystal structure analysis of a complete turn of B-DNA. *Nature*. 1980; 287(5784): 755–8. [PubMed: 7432492]
3. Watson JD, Crick FH. Genetical implications of the structure of deoxyribonucleic acid. *Nature*. 1953; 171(4361):964–7. [PubMed: 13063483]
4. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*. 1953; 171(4356):737–8. [PubMed: 13054692]
5. Vaquerizas JM, et al. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*. 2009; 10(4):252–63. [PubMed: 19274049]
6. Riechmann JL, et al. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*. 2000; 290(5499):2105–10. [PubMed: 11118137]
7. Riano-Pachon DM, et al. PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics*. 2007; 8:42. [PubMed: 17286856]
8. Jordan SR, Pabo CO. Structure of the lambda complex at 2.5 Å resolution: details of the repressor-operator interactions. *Science*. 1988; 242(4880):893–9. [PubMed: 3187530]
9. Chenoweth DM, Dervan PB. Allosteric modulation of DNA by small molecules. *Proc Natl Acad Sci U S A*. 2009; 106(32):13175–9. [PubMed: 19666554]
10. Galas DJ, Schmitz A. DNase footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res*. 1978; 5(9):3157–70. [PubMed: 212715]
11. Hardison RC, et al. The structure and transcription of four linked rabbit beta-like globin genes. *Cell*. 1979; 18(4):1285–97. [PubMed: 519769]
12. Stranick KS, et al. Identification of transcription factor binding sites important in the regulation of the human interleukin-5 gene. *J Biol Chem*. 1997; 272(26):16453–65. [PubMed: 9195954]
13. Kardassis D, et al. Characterization of the promoter elements required for hepatic and intestinal transcription of the human apoB gene: definition of the DNA-binding site of a tissue-specific transcriptional factor. *Mol Cell Biol*. 1990; 10(6):2653–9. [PubMed: 2342459]
14. Kardassis D, Zannis VI, Cladaras C. Organization of the regulatory elements and nuclear activities participating in the transcription of the human apolipoprotein B gene. *J Biol Chem*. 1992; 267(11): 7956. [PubMed: 1560025]
15. Shibata Y, Crawford GE. Mapping regulatory elements by DNaseI hypersensitivity chip (DNase-Chip). *Methods Mol Biol*. 2009; 556:177–90. [PubMed: 19488879]

16. Quitschke WW, et al. Differential effect of zinc finger deletions on the binding of CTCF to the promoter of the amyloid precursor protein gene. *Nucleic Acids Res.* 2000; 28(17):3370–8. [PubMed: 10954607]
17. Perry RP, Kelley DE. Evidence for specific association of protein with newly formed ribosomal subunits. *Biochem Biophys Res Commun.* 1966; 24(3):459–65. [PubMed: 5967111]
18. Economidis IV, Rousseau GG. Association of the glucocorticoid hormone receptor with ribonucleic acid. *FEBS Lett.* 1985; 181(1):47–52. [PubMed: 2578982]
19. Licatalosi DD, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature.* 2008; 456(7221):464–9. [PubMed: 18978773]
20. Wang Z, et al. CLIP: construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo. *Methods.* 2009; 48(3):287–93. [PubMed: 19272451]
21. Solomon MJ, Varshavsky A. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proc Natl Acad Sci U S A.* 1985; 82(19):6470–4. [PubMed: 2995966]
22. Varshavsky AJ, Sundin O, Bohn M. A stretch of “late” SV40 viral DNA about 400 bp long which includes the origin of replication is specifically exposed in SV40 minichromosomes. *Cell.* 1979; 16(2):453–66. [PubMed: 222461]
23. Gasch AP, et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell.* 2000; 11(12):4241–57. [PubMed: 11102521]
24. Causton HC, et al. Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell.* 2001; 12(2):323–37. [PubMed: 11179418]
25. Ross DT, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet.* 2000; 24(3):227–35. [PubMed: 10700174]
26. Alizadeh AA, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature.* 2000; 403(6769):503–11. [PubMed: 10676951]
27. Boldrick JC, et al. Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proc Natl Acad Sci U S A.* 2002; 99(2):972–7. [PubMed: 11805339]
28. Iyer VR, et al. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature.* 2001; 409(6819):533–8. [PubMed: 11206552]
29. Ren B, et al. Genome-wide location and function of DNA binding proteins. *Science.* 2000; 290(5500):2306–9. [PubMed: 11125145]
30. Lieb JD, et al. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet.* 2001; 28(4):327–34. [PubMed: 11455386]
31. Horak CE, et al. Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev.* 2002; 16(23):3017–33. [PubMed: 12464632]
32. Lee TI, et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science.* 2002; 298(5594):799–804. [PubMed: 12399584]
33. Finishing the euchromatic sequence of the human genome. *Nature.* 2004; 431(7011):931–45. [PubMed: 15496913]
34. Waterston RH, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature.* 2002; 420(6915):520–62. [PubMed: 12466850]
35. Weinmann AS, et al. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev.* 2002; 16(2):235–44. [PubMed: 11799066]
36. Ren B, et al. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev.* 2002; 16(2):245–56. [PubMed: 11799067]
37. Odom DT, et al. Control of pancreas and liver gene expression by HNF transcription factors. *Science.* 2004; 303(5662):1378–81. [PubMed: 14988562]
38. Horak CE, et al. GATA-1 binding sites mapped in the beta-globin locus by using mammalian ChIP-chip analysis. *Proc Natl Acad Sci U S A.* 2002; 99(5):2924–9. [PubMed: 11867748]
39. Martone R, et al. Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc Natl Acad Sci U S A.* 2003; 100(21):12247–52. [PubMed: 14527995]
40. Carroll JS, et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell.* 2005; 122(1):33–43. [PubMed: 16009131]

41. Massie CE, et al. New androgen receptor genomic targets show an interaction with the ETS1 transcription factor. *EMBO Rep.* 2007; 8(9):871–8. [PubMed: 17721441]
42. Bernstein BE, et al. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell.* 2005; 120(2):169–81. [PubMed: 15680324]
43. Vera JC, et al. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol.* 2008; 17(7):1636–47. [PubMed: 18266620]
44. Schuster SC. Next-generation sequencing transforms today's biology. *Nat Methods.* 2008; 5(1):16–8. [PubMed: 18165802]
45. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.* 1977; 74(12):5463–7. [PubMed: 271968]
46. Gibbs RA, et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature.* 2004; 428(6982):493–521. [PubMed: 15057822]
47. Gibbs RA, et al. Evolutionary and biomedical insights from the rhesus macaque genome. *Science.* 2007; 316(5822):222–34. [PubMed: 17431167]
48. Mikkelsen TS, et al. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature.* 2007; 447(7141):167–77. [PubMed: 17495919]
49. Shendure J, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science.* 2005; 309(5741):1728–32. [PubMed: 16081699]
50. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol.* 2008; 26(10):1135–45. [PubMed: 18846087]
51. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet.* 2008; 9:387–402. [PubMed: 18576944]
52. Schadt EE, et al. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* 2008; 6(5):e107. [PubMed: 18462017]
53. Yang X, et al. Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res.* 2006; 16(8):995–1004. [PubMed: 16825664]
54. Kasowski M, et al. Variation in transcription factor binding among humans. *Science.* 328(5975):232–5. [PubMed: 20299548]
55. McDaniel R, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science.* 328(5975):235–9. [PubMed: 20299549]
56. Odom DT, et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet.* 2007; 39(6):730–732. [PubMed: 17529977]
57. Schmidt D, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science.* 2010; 328(5981):1036–40. [PubMed: 20378774]
58. Bourque G. Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr Opin Genet Dev.* 2009; 19(6):607–12. [PubMed: 19914058]
59. Kunarso G, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet.* 2007; 39(6):631–4. [PubMed: 20526341]
60. Birney E, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature.* 2007; 447(7146):799–816. [PubMed: 17571346]
61. Wilson MD, et al. Species-specific transcription in mice carrying human chromosome 21. *Science.* 2008; 322(5900):434–8. [PubMed: 18787134]
62. Brockes JP, Kumar A. Appendage regeneration in adult vertebrates and implications for regenerative medicine. *Science.* 2005; 310(5756):1919–23. [PubMed: 16373567]
63. Schmidt D, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science.* 328(5981):1036–40. [PubMed: 20378774]
64. MacArthur S, et al. Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol.* 2009; 10(7):R80. [PubMed: 19627575]

Inset Box 1. The elephant in the room: TFs rarely regulate the expression of bound genes simplistically or intuitively. A major finding of the simultaneous comparison of gene expression and TF binding experiments in the same tissues and cells has been that our early and overly-simplistic model of TF binding leading directly to transcriptional regulation (panel A) is rarely true in complex eukaryotes like mammals. It is common for a TF to bind tens or even hundreds of thousands of locations in the mammalian genome, yet only a few hundred transcripts noticeably change their level in response to removal via siRNA or genetic deletion of this TF— and few of these transcriptionally altered genes are bound directly by the TF (Panel B). One (controversial) model pioneered by part of the fly transcriptional community that may help account for this disparity is that the regional concentration of **all** protein-DNA interactions that may control transcription [64]. This model could also help explain the surprisingly rapid divergence of mammalian TFs seen in recent experiments [61].

Figure 1.

The structure of deoxyribonucleic acid (DNA). (A) A schematic model of the stacked B-form 12 base sequence (blue and red, indicating the two halves of a palindromic sequence) is shown with the DNA binding domain of a leucine zipper protein (green) inserted into the major groove symmetrically. (B) The adenine-thymine (A-T) and cytosine-guanine (C-G) base pairings proposed in the original 1953 Nature report from Watson and Crick are shown. The hydrogen-bonding potential in the major groove (upper surface of the base pairs) is suggested by the nitrogen-hydrogen bond in the A-T pair, as well as the (implied) lone pairs

directed from the nitrogens and oxygens in both A-T and C-G (reprinted by permission from [4]).

Figure 2.

Identification of protein-DNA contacts using chromatin immunoprecipitation. (A) Microarrays that contain the genetic sequence of promoter regions can be used to interrogate the complete set of nucleic acids enriched by antibody binding to TF-DNA complexes (reprinted by permission from [28]). (B) Primer sets can also be used to interrogate or confirm limited subsets of these binding events (reprinted by permission from [29]).

Figure 3.

Transcriptional regulators often bind to the genes coding for other transcriptional regulators in eukaryotes. Yeast transcriptional regulators are shown in a circle and organized by their functional categories. Potential regulatory linkages are shown as arrows between one TF and the promoter of another that is bound; lines between categories are very common (reprinted by permission from [32]).

Figure 4.

Tissue-specific TF binding evolves rapidly among mammals. (A) The in vivo binding of CEBPA is shown in five species near the PCK1 locus. (B) The pairwise overlap between two of five vertebrates is shown. The pie chart indicates the location in the genome of the total set of binding events (red: intergenic; green: promoter [TSS+/- 3 KB]; yellow: intronic). The left most piechart is the whole genome, the right most diagonal shows where the set of all binding events fall. (C&D) A sizable fraction of binding events are found in regions that are >50% repeat for CEBPA (C) and HNF4A (D). Reprinted by permission from [63].

Figure 5.

A mouse nucleus can accurately read human genetic instructions to place the CEBPA TF in adult liver. (A) The chromosome-wide binding events for CEBPA in human liver were sorted first by whether they were shared with mouse, and then within these two classes by intensity of binding. The Tc1-Hs-chr21 lane is the signal obtained from the human chromosome in a mouse environment, which is largely indistinguishable from the native human case. (B) The CEBPA binding events near the ETS2 locus reflect the largely identical signal in human or mouse for this chromosome. Reprinted by permission from [63]; see also [61].