STUDY DESIGNS

ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions

Terrence S. Furey

Abstract | Chromatin immunoprecipitation experiments followed by sequencing (ChIP–seq) detect protein–DNA binding events and chemical modifications of histone proteins. Challenges in the standard ChIP–seq protocol have motivated recent enhancements in this approach, such as reducing the number of cells that are required and increasing the resolution. Complementary experimental approaches — for example, DNaseI hypersensitive site mapping and analysis of chromatin interactions that are mediated by particular proteins — provide additional information about DNA-binding proteins and their function. These data are now being used to identify variability in the functions of DNA-binding proteins across genomes and individuals. In this Review, I describe the latest advances in methods to detect and functionally characterize DNA-bound proteins.

DNA-binding proteins play crucial roles in many major cellular processes, such as transcription, splicing, replication and DNA repair. These proteins include transcription factors that bind preferentially to certain DNA sequences, as well as histone proteins that form the core of nucleosomes, which are the basic units of chromatin. The genomic locations of neither bound factors nor modified histones can be accurately predicted in a particular cell type using DNA sequence features alone, and functional assays are necessary to identify these cellular characteristics. Chromatin immunoprecipitation coupled with microarrays (ChIP–chip) or short-tag sequencing (ChIP–seq) has become the standard technique for identifying the locations and biochemical modifications of bound proteins genome-wide[1–3]. Recent advances in ChIP methodology have overcome some of the limitations of the 'standard' ChIP experiment, and the development of complementary assays and analyses have expanded the number, types and resolution of protein–DNA interactions that have been discovered.

In this Review, I discuss the current state of ChIP-based experiments, including modifications of the standard ChIP protocol, and I review basic features of ChIP–seq analysis pipelines. I then describe alternatives to ChIP, including open chromatin assays such as DNase–seq[4–7], formaldehyde-assisted identification of regulatory elements (FAIRE–seq)[8–10], and genome-wide DNaseI footprinting[11–14]. Finally, I discuss approaches for characterizing protein–DNA interactions that are improving our understanding of function. These include three-dimensional chromatin assays — such as chromatin conformation capture (3C) and its derivatives[15–17], and chromatin interaction analysis with paired-end tag sequencing (ChIA-PET)[18,19]— that provide evidence for functional targets of DNA-bound proteins, and analyses of sequence-based data from ChIP[20,21] and other experiments[22–24] that reveal allele-specific effects on protein–DNA binding.

ChIP–seq experiments

Current ChIP–seq experiments. ChIP is the most direct way to identify the binding sites of a single DNA-binding protein or the locations of modified histones. The basic steps of the ChIP–seq assay have been reviewed elsewhere[25–27] and are depicted in FIG. 1a for transcription factors and in FIG. 1b for histone modifications. The Encyclopedia of DNA Elements (ENCODE) Consortium[28] has carried out hundreds of ChIP–seq experiments and has used this experience to develop a set of working standards and guidelines[29] (BOX 1). It must be noted that given the diversity of cell types, conditions, factors and modifications being assayed, it is near impossible to define common guidelines that will be appropriate for all situations. From a technical perspective, the success of a ChIP experiment depends on the development and validation of a highly specific antibody

Departments of Genetics and Biology, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, 120 Mason Farm Road, CB#7264, Chapel Hill, North Carolina 27599, USA.
e-mail: tsfurey@email.unc.edu
doi:10.1038/nrg3306
Published online
23 October 2012

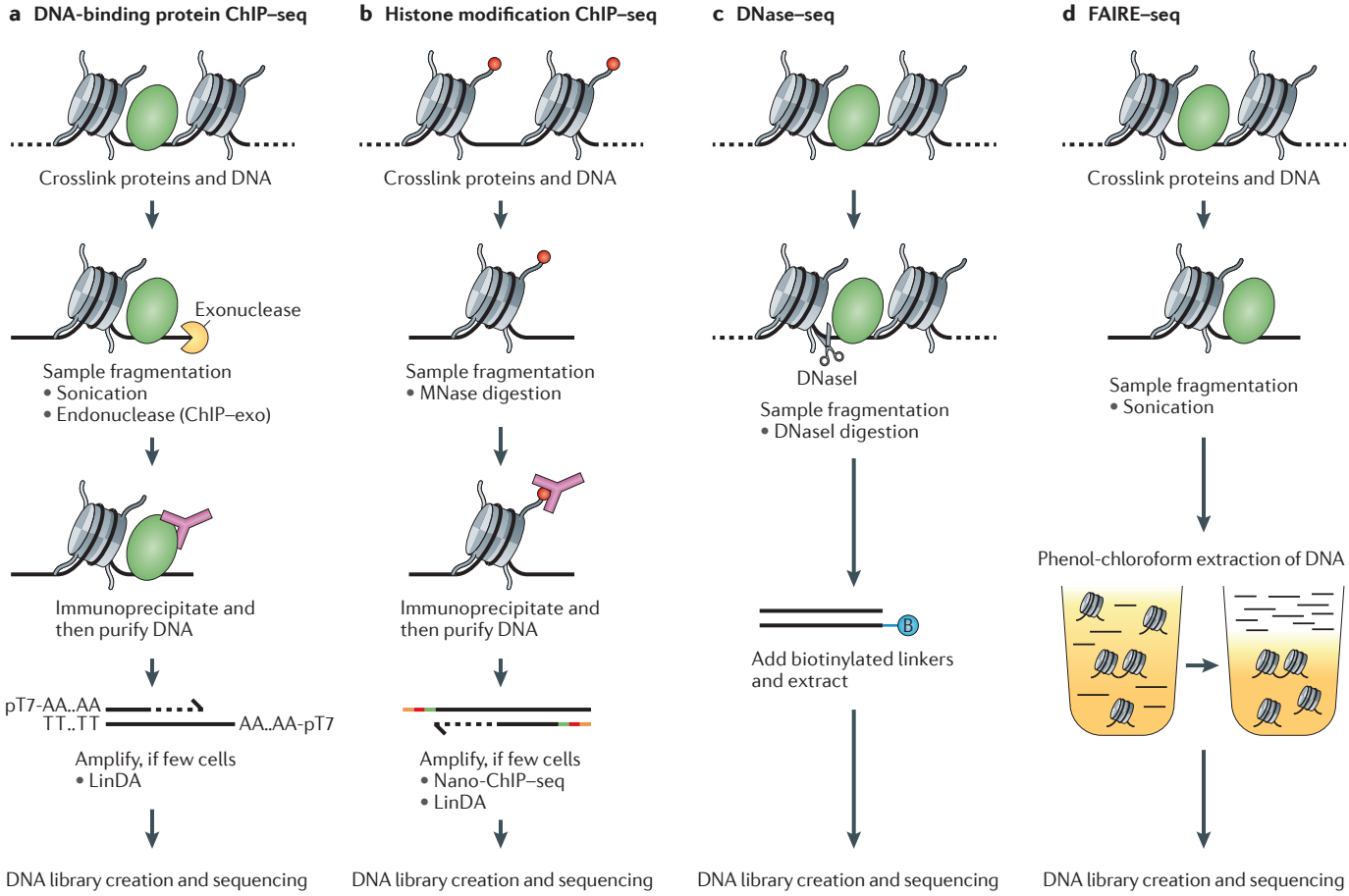


Figure 1 | **Comparison of experimental protocols.** Experiments to detect different aspects of DNA-binding proteins share many of the same steps; simplified schematics of the main steps are shown. **a** | Chromatin immunoprecipitation followed by sequencing (ChIP–seq) for DNA-binding proteins such as transcription factors. Recent variations on the standard protocol include using endonuclease digestion instead of sonication (ChIP–exo) to increase the resolution of binding-site detection and to eliminate contaminating DNA, and DNA amplification after ChIP for samples with limited cells. **b** | ChIP–seq for histone modifications uses micrococcal nuclease (MNase) digestion to fragment DNA and can also now be run on low-quantity samples when combined with the additional post-ChIP amplification. **c** | DNase–seq relies on digestion by the DNaseI nuclease to identify regions of nucleosome-depleted open chromatin where there are binding sites for all types of factors, but it cannot identify what specific factors are bound. **d** | Formaldehyde-assisted identification of regulatory elements (FAIRE–seq) similarly identifies nucleosome-depleted regions by extracting fragmented DNA that is not crosslinked to nucleosomes. LinDA, single-tube linear DNA amplification; T7, T7 phage RNA polymerase.

to the bound protein or modification. Antibody quality varies, even between independently prepared lots of the same antibody, as demonstrated in a recent assessment of over 200 human, fly and worm antibodies as part of the ENCODE and the model organism ENCODE (modENCODE) projects[30]. In this study, 25% failed specificity tests and 20% failed immunoprecipitation experiments. In addition, multiple histone modifications can alter the efficacy of certain antibodies[31]. Other technical challenges include the requirement for large numbers of cells and prior knowledge of the existence of a DNA-binding protein or histone modification. Possible solutions to these issues are considered below and in later sections.

Limited cells. Typically, large numbers of cells (~10 million) are required for a ChIP experiment, thus limiting the types of cells that can be assayed and the number of ChIP experiments that can be carried out on a valuable sample. It can be especially challenging in small model organisms, for which multiple whole animals may be necessary to achieve these quantities. Two protocols have been developed recently to address this problem through post-ChIP DNA amplification (FIG. 1a,b).

Nano-ChIP–seq[32] has been successfully carried out on as few as 10,000 cells for histone modifications. The authors recommend using variable sonication times and antibody concentrations that are scaled in proportion to the number of starting cells. The small amount of DNA that is extracted following the ChIP experiment is PCR amplified using custom primers that form a hairpin structure at their 5′ end to prevent self-annealing when being added. The primers also contain a BciVI restriction site that allows the direct addition of Illumina sequencing adaptors to the resulting amplified DNA,

Sonication
The fragmenting of DNA sequence by exposing it to high-frequency sound waves.

Box 1 | **Recommended ChIP–seq standards**

Based on the collective experience of laboratories involved in the Encyclopedia of DNA Elements (ENCODE) and model organism ENCODE (modENCODE) projects, which have carried out hundreds of chromatin immunoprecipitation followed by sequencing (ChIP–seq) experiments, a set of standards and guidelines for carrying out ChIP–seq has been written[29]. Experiments are classified as point source (highly localized signals, such as for transcription factors), broad source (signals that span large domains; for example, as for some histone modifications such as H3K36me3) or mixed source (signals that have elements of both, such as for RNA polymerase II). If the type of signal is unknown, multiple peak callers focusing on point-source or broad-source peaks may be applied to determine the best fit to the data. These standards are summarized below.

Antibody validation

The primary characterization of transcription factor antibodies is carried out using immunoblot or immunofluorescence analysis. Secondary characterization is carried out using one of: factor knockdown by mutation or RNAi; independent ChIP experiments using alternative epitopes or protein members of a complex; immunoprecipitation using epitope-tagged constructs; mass spectrometry; or binding-site motif analyses. The primary characterization of histone modification antibodies is carried out using immunoblot analysis. The secondary characterization is carried out using one of: peptide binding tests; mass spectrometry; immunoreactivity analysis in cell lines containing knockdowns of relevant histone modification enzymes or mutant histones; or genome annotation enrichment.

Sequencing depth

Generating a ChIP–seq profile requires different amounts of sequencing data depending on the size of the genome and the peak type. For human genomes, 20 million uniquely mapped read sequences are suggested for point-source peaks, or 40 million for broad-source peaks. For fly or worm genomes, these values are 8 million and 10 million reads, respectively. An increased sequencing depth allows the detection of more sites that have lower levels of enrichment over the genomic background. It is noted that setting a minimal signal strength threshold, usually based on a *P* value or false-discovery rate calculation, to identify peaks does not guarantee the discovery of all functional sites. It is also noted that DNA sequencing library complexity (that is, the amount of unique DNA molecules) must be sufficient, such that the sequencing depth does not exceed the library complexity. It is suggested that at least 80% of 10 million or more reads be mapped to distinct genomic locations. Low-complexity libraries generally indicate a failed experiment in which not enough DNA was recovered; this causes the same PCR-amplified products to be sequenced repeatedly and many small peaks to be detected with a high false-positive rate.

Experimental replication

A minimum of two replicates should be carried out per experiment. Each replicate of a human genome experiment should have 10 million uniquely mapped reads per replicate for point-source peaks or 20 million for broad-source peaks. For fly or worm genomes, these values should be 4 million and 5 million reads, respectively. Each replicate should be a biological rather than a technical replicate; that is, it represents an independent cell culture, embryo pool or tissue sample. For two replicates, either 80% of the top 40% of identified targets in one replicate must be among the targets in the second replicate; alternatively, 75% of target lists must be in common between both replicates.

Data quality assessment

No single test is universally suitable for all experiments, nor is always necessary. Recommended assessments include: investigating signals at known sites using a genome browser; calculating the fraction of reads in peaks (FRiP), which is recommended to be >1%; and calculating cross-correlations. Cross-correlations are defined as the correlation of the density of sequences aligned to the Watson strand with the density of sequences aligned to the Crick strand after shifting the Watson strand alignments by the average distance between opposite strands reads.

Data and metadata reporting

ChIP results should be submitted to the Gene Expression Omnibus (GEO)[109]. The provided experimental and analysis information should include ChIP procedures, antibody validation, DNA sequencing information, identified regions of enrichment and their method of identification, and any other analysis.

which makes DNA library preparation and sequencing straightforward. The number of cells required is dependent on multiple factors, including antibody efficiency and the abundance of the target protein. Therefore, although 10,000 cells were sufficient to assay the chromatin mark histone H3 trimethylated on lysine 4 (H3K4me3), ChIPs for less-abundant histone modifications or transcription factors will probably require more cells and may require further optimization of certain steps such as sonication time.

The second protocol uses single-tube linear DNA amplification (LinDA) and has been successfully applied for the oestrogen receptor-α (ERα) transcription factor using 5,000 cells and for the H3K4me3 histone modification using 10,000 cells[33]. The key to this technique is an optimized T7 phage RNA polymerase linear amplification protocol[34]. A major concern in any amplification protocol is that technical biases could result in uneven amplification of the starting material. LinDA was shown to be robust for the even amplification of starting material; importantly, it seemed to avoid bias in relation to GC content, which is generally problematic for PCR-based approaches.

Increased precision. Standard ChIP–seq experiments that use sonication to fragment chromatin result in libraries containing DNA molecules that are ~200 bases long, even though each protein typically binds only 6–20 bases. In addition, the resulting libraries are often contaminated with DNA that was not bound by the target factor. This contamination is responsible for some common systematic biases and has necessitated the use of input control experiments.

ChIP–exo[35] uses lambda (λ) phage exonuclease to digest the 5′ end of protein-bound and formaldehyde-crosslinked DNA fragments to a fixed distance from the bound protein (FIG. 1a); fixation is a barrier to 5′–3′ digestion. As DNA fragments are produced from both strands during ChIP, the 5′ ends of sequence tags align primarily at two genomic locations corresponding to the barriers on each strand. The protein is bound to the region between these locations. In addition, the exonuclease largely eliminates contaminating DNA. Experiments in yeast for the Reb1 transcription factor[35] showed that ChIP–exo could identify binding sites with single base-pair precision (which is a 90-fold greater precision than when using the standard protocol), and with a 40-fold increase in the signal-to-noise ratio, thus indicating a lower background (contaminating) signal.

Multiple binding events. DNA-bound proteins and histone modifications work together and with other genomic modifications to carry out cellular functions. When multiple experiments indicate different proteins or modifications at the same genomic location, it is not clear whether these are simultaneously present or are present on different chromosomes in the same cell or in different cells. Sequential ChIP (also known as re-ChIP or co-immunoprecipitation)[36] uses antibodies to different proteins in successive experiments to determine the genomic locations where both targets are present.

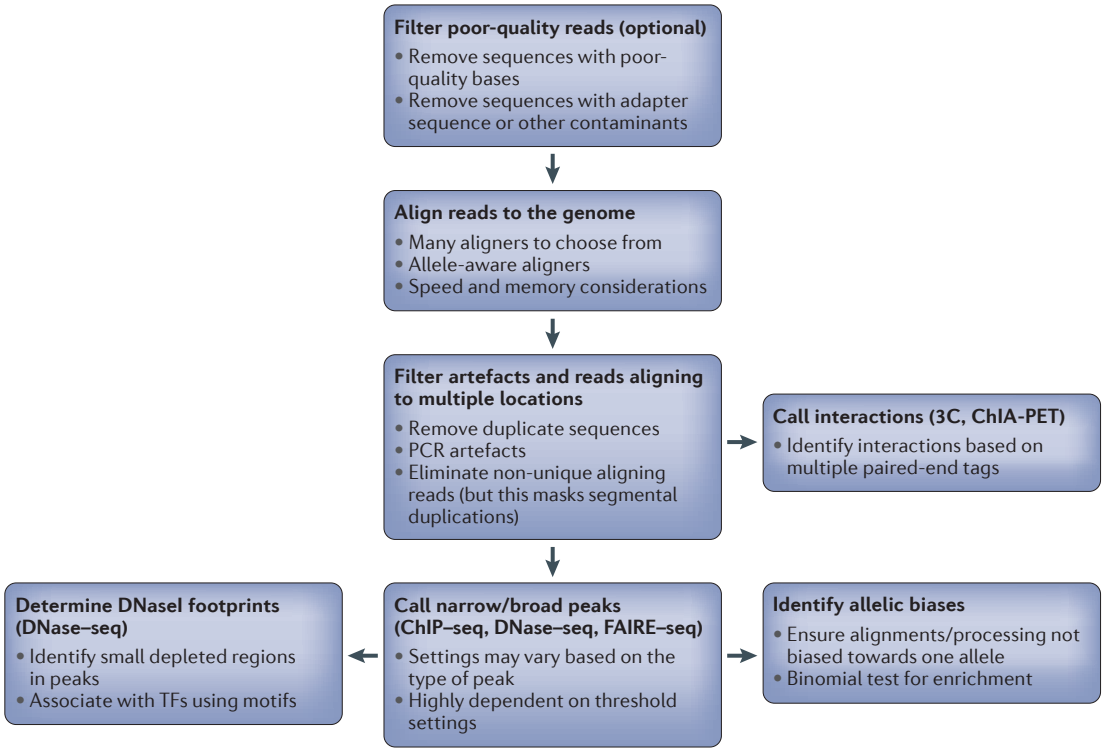


Figure 2 | **General analysis pipeline for sequence-tag experiments.** Different experiments that use short sequence reads to identify regions with a particular molecular characteristic share many of the same analysis steps. Poor-quality reads can be filtered initially, but often the inability to align these reads to the genome sufficiently removes bad sequences. Alignment using one of many possible software programs (TABLE 1) is followed by filtering artefacts that arose during the PCR amplification step when sequencing, or that appear owing to the under-representation of certain sequences in the reference genome, such as peri-centromeric satellite sequences. Often, reads aligning to more than a chosen number of genomic locations are removed. For experiments to identify independent locations, 'peak'-calling tools (TABLE 1) identify genomic regions of signal enrichment, which indicate a bound protein, histone modification or open chromatin. By contrast, chromatin interaction experiments use aligned paired-end reads to find evidence of interacting distal genomic regions. DNaseI footprints (FIG. 3) indicate local protection from DNaseI digestion within a larger DNaseI-hypersensitive site (DHS) region due to a bound protein. The distribution of alleles in sequences spanning heterozygous variants can be analysed to determine if a bias towards sequences with one of the two alleles exists (FIG. 5). This may reflect a functional difference caused by the underlying genotype.3C, chromatin conformation capture; ChIA-PET, chromatin interaction analysis with paired-end tag sequencing; ChIP–seq, chromatin immunoprecipitation followed by sequencing; DNase–seq, DNaseI-hypersensitive sites sequencing; FAIRE–seq, formaldehyde-assisted identification of regulatory elements; TFs, transcription factors.

However, experiments have only been carried out at individual loci and not in conjunction with high-throughput sequencing. Recently, assays have been developed that use bisulphite sequencing to identify methylated DNA in immunoprecipitated chromatin fragments[37,38]. These genome-wide experiments showed that DNA methylation and the histone modification H3K27me3 can occur simultaneously. More generally, new techniques have been developed to reveal the identities of individual proteins interacting in larger complexes in human and model organisms[39–47], thus providing evidence for combinations of factors that bind together.

ChIP–seq analysis pipelines

There has also been a large effort to improve analytical tools that are necessary to interpret the sequence data output from ChIP–seq experiments. Computational processing pipelines are generally implemented to progress from raw sequence reads to usable annotations. Steps common to many pipelines are depicted in FIG. 2. Each step has led to the development of specialized software tools, which are briefly discussed below.

Sequence aligners must be fast and accurate, and several strategies have been developed to achieve these goals (TABLE 1; see REF. 48 for a recent review). Given a final set of aligned sequences, genomic regions are identified that contain enriched signals (that is, 'peaks') where more sequences are aligned than would be expected by chance, thus indicating locations of binding sites or histone modifications. Several software programs have been developed to identify these peaks (TABLE 1; see REFS 49–52 for recent comparisons of the methods). When available, data from input control experiments are used by most peak callers to represent the background levels of signal.

Exonuclease
An enzyme that cleaves a single nucleotide from the end of a DNA molecule.

Crosslinked
The strong binding of DNA to interacting proteins through covalent bonds.

Table 1 | **A subset of software tools available for three key steps in the analysis of sequence data**

Software tool	Web address	Notes
Short-read aligners		
BWA	http://bio-bwa.sourceforge.net	Fast and efficient; based on the Burrows–Wheeler transform
Bowtie	http://bowtie-bio.sourceforge.net	Similar to BWA, part of suite of tools that includes TopHat and CuffLinks for RNA-seq processing
GSNAP	http://research-pub.gene.com/gmap	Considers a set of variant allele inputs to better align to heterozygous sites
Wikipedia list of aligners	http://en.wikipedia.org/wiki/List_of_sequence_alignment_software#Short-Read_Sequence_Alignment	A comprehensive list of available short-read aligners, with descriptions and links to download the software
Peak callers		
MACS	http://liulab.dfci.harvard.edu/MACS	Fits data to a dynamic Poisson distribution; works with and without control data
PeakSeq	http://info.gersteinlab.org/PeakSeq	Takes into account differences in mappability of genomic regions; enrichment based on FDR calculation
ZINBA	http://code.google.com/p/zinba	Can incorporate multiple genomic factors, such as mappability and GC content; can work with point-source and broad-source peak data
Differential peak calling		
edgeR	http://www.bioconductor.org/packages/2.9/bioc/html/edgeR.html	Uses negative binomial distribution to model differences in tag counts; uses replicates to better estimate significant differences
DESeq	http://www-huber.embl.de/users/anders/DESeq	Also uses negative binomial distribution modelling, but differs in the calculation of the mean and variance of the distribution
baySeq	http://www.bioconductor.org/packages/release/bioc/html/baySeq.html	Uses empirical Bayes approach to identify significant differences; assumes negative binomial distribution of data
SAMSeq	http://www.stanford.edu/~junli07/research.html#SAM	Based on the popular SAM software; a non-parametric method that uses resampling to normalize for differences in sequencing depth

BWA, Burrows–Wheeler Aligner; DESeq, analysis of high-throughput sequencing to detect differential expression; FDR, false-discovery rate; GSNAP, Genomic Short-read Nucleotide Alignment Program; MACS, Model-based Analysis for ChIP–seq; RNA-seq, high-throughput RNA sequencing; SAM, significance analysis of microarrays; ZINBA, Zero-Inflated Negative Binomial Algorithm.

Many programs also control for differences in mappability to regions of the genome. As described in BOX 1, peaks can be point source (highly localized signals, such as for transcription factors), broad source (signals that span large domains; for example, for some histone modifications such as H3K36me3) or mixed source (signals that have elements of both, such as for RNA polymerase II (Pol II) binding). Each of these requires different detection strategies; some software is focused primarily on one type of peak, whereas others offer different settings that tune the software on the basis of the peak shape.

It is often desirable to compare data from multiple experiments; for example assaying the same transcription factor in two different cell types or conditions, to investigate common and cell-type-specific activity. Simply comparing peaks from each experiment is often used to identify regions that are differentially bound or modified. However, this approach may not identify regions that are called as peaks in both experiments but which have very different strengths of signal, and it may incorrectly identify regions that were just above the peak threshold in one experiment but just below in the other.

Several software packages, which were originally developed for high-throughput RNA sequencing (RNA-seq) data, are now available that can be adapted to identify statistically significant differences directly on the basis of ChIP–seq read count data (TABLE 1; see REFS 53, 54 for a comparison).

Using experimental evidence of factor binding sites, there is an opportunity to improve the characterization of preferred DNA binding motifs for each factor. Several groups have developed software that uses information from ChIP–seq experiments during motif discovery[55–60]. More-accurate modelling of binding preferences allows for better prediction of significant signals and the precise DNA contact site for factor binding events identified by ChIP–seq.

Sequencing considerations. We are still discovering biases and systematic errors in sequence data that result from combinations of genomic characteristics, experimental protocols, specific sequencing technologies, batch effects and analytical methods[61]. Biases have been studied in multiple types of experimental data, mainly

Mappability
The uniqueness of a stretch of DNA sequence compared with a whole-genome sequence. Short sequence reads can be confidently mapped to unique sequence, but less confidently mapped to sequence that occurs multiple times in a genome.

DNA binding motifs
A degenerate pattern of DNA sequences to which transcription factors prefer to bind. They are often represented as a probabilistic matrix.

generated using Illumina's Genome Analyzer sequencer, to better understand how to detect biases and correctly normalize data to uncover true signals[62–67]. These studies indicate the need to normalize for chromatin structure (which may affect fragmentation), for uneven nucleotide distributions across read base positions and for GC content. All of these can affect the number of sequence reads generated from particular genomic regions. Using an input control to correct for these assumes that biases are similar between the paired ChIP and input experiments, which is not necessarily the case[65], and thus biases may need independent correction in input controls. Mappability can also affect accurate signal detection; this can partially be remedied by using paired-end sequencing. In one study, sequencing paired-ends was shown to nearly double the effective genomic coverage in repeat regions, but with increased sequencing costs[62]. The effect of sequencing depth on accuracy and sensitivity was also assessed, and it was found that some binding sites were missed even at high depths (16.2 million reads across the *Drosophila melanogaster* genome, which is equivalent to approximately 327 million reads across the human genome)[62]. Individual base sequencing errors in reads are also not uniform. For example, it is well known that base-pair quality degrades towards the 3′ end of Illumina-sequenced reads[63]. In addition, certain substitutions are more prevalent (that is, A→C and G→T), inverted repeats and G-rich sequences (especially GGC) often precede errors, and quality scores at the high and low end often overestimate and underestimate, respectively, the true error rate[61,63,66,68].

Further analytical challenges. Despite the progress, several challenges remain. As read length increases, the current short-read aligners will probably require further modification[48], and alignments to repetitive sequences will remain a challenge[69–71]. Continued effort is needed to develop or improve methods to identify real events, given the inherent biases and errors described above, and to enable a better interpretation. For example, although we would like to think of the assayed binding or modification events as binary — that is, a protein is or is not bound to a given location — the data are more continuous in nature. Signal strength at a particular location is influenced by the strength of the interaction, which can be modulated by variations in genotype, and by the percentage of the population of cells assayed that have the binding or modification event. Signals may reflect not only direct binding events, but also indirect binding in which one factor interacts with another factor that is bound to DNA. Distinguishing between direct and indirect events is important but cannot be achieved directly from ChIP data.

Open chromatin

Most transcription factors cannot stably interact with their DNA targets if the DNA is nucleosomal. For stable binding to occur, nucleosomes must be displaced or translocated to create a nucleosome-depleted, open chromatin region. Detecting open chromatin complements ChIP–seq data and can identify binding sites for nearly all factors simultaneously. Two distinct assays, DNase–seq and FAIRE–seq, have been developed to detect open chromatin directly (see REF. 72 for a review of genome accessibility experiments).

DNase–seq and FAIRE–seq. The DNaseI endonuclease non-specifically digests DNA, but in the normal context of chromatin structure it will preferentially digest unbound, open chromatin. As most DNA is wrapped in a nucleosome, DNaseI-hypersensitive (DHS) sites largely correspond to nucleosome-depleted regions, and these are primarily the regions that have gene-regulatory functions, such as promoters, enhancers, silencers, insulators and locus control regions[73–75]. DNase–seq experiments (FIG. 1c) combine traditional DHS assays with high-throughput sequencing to simultaneously identify all types of regulatory regions genome-wide[4,7,76]. The 5′ end of a sequence tag generated by DNase–seq indicates the site of a DNaseI digestion event, and regions of enrichment in digestion events are identified as DHS sites, each of which can contain binding sites of multiple factors. Comparisons with ChIP–seq data indicate that DNase–seq captures the vast majority of binding sites for most factors[4,6,7].

The FAIRE–seq assay[8,9] starts with formaldehyde crosslinking, similarly to ChIP, but then instead of using an antibody to target specific factors, DNA is sonicated and the extract is subjected to phenol-chloroform extraction. The nucleosome-depleted fraction of DNA is preferentially segregated to the aqueous phase. FAIRE-enriched DNA has been shown to correspond to regulatory regions[8].

Enriched regions from these two assays are highly overlapping but are not identical[6]. In a comparison[6], both showed good correspondence to ChIP–seq data for multiple factors, and most factor binding sites were found by both methods. However, each method identified a subset of putative regulatory elements that are not seen in the other. Binding sites of certain factors (such as FOXA1, FOXA3 and GATA1) were better identified by FAIRE–seq, whereas others (such as ZNF263 and CTCF) were more often seen in DNase–seq data. Sites that were only found in DNase–seq assays were enriched at promoter regions and in regions that have the promoter-associated histone modifications H3K4me3 and H3K9 acetylation (H3K9ac), whereas sites that were specific to FAIRE–seq were more often in introns and exons, intergenic regions and H3K4me1 regions[6].

The FAIRE–seq assay is fairly easy to carry out, although some optimization of crosslinking times may be needed for different cell types or tissues owing to variation in fixation efficiency[10]. DNase–seq can be more difficult at the bench as optimization is required for cell lysis procedures and DNaseI concentration[5]. The signal-to-noise ratio — that is, the fraction of sequences in enriched regions versus non-enriched regions — is higher for DNase–seq than for FAIRE–seq, which contributes to the identification of more-precise DNA binding sites (known as DNaseI footprints), as described below. Advantages of DNase–seq and FAIRE–seq

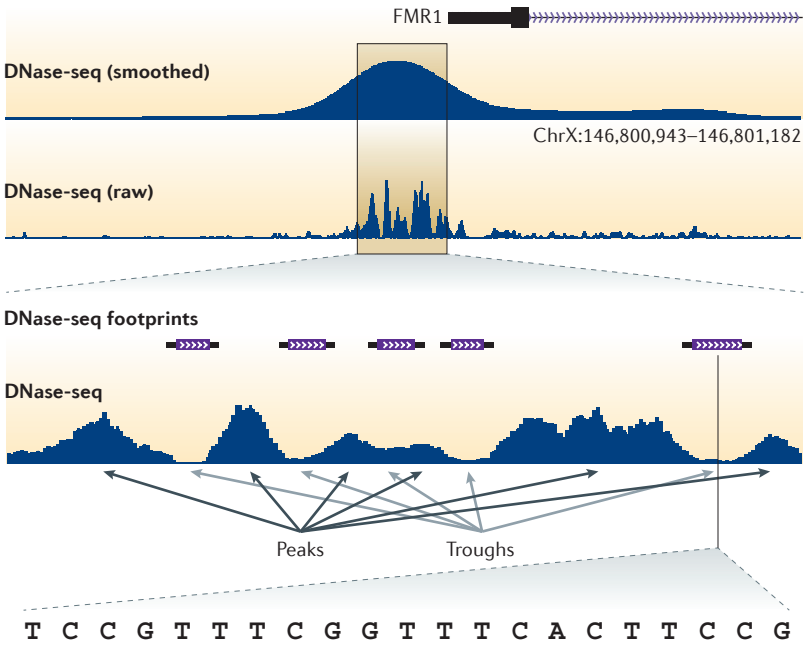
Promoters
DNA sequences immediately upstream of transcription start sites at which RNA polymerases and transcription factors bind to initiate gene transcription.

Enhancers
DNA sequences at which transcription factors bind that increase the transcription rate of one or more target genes that can be at varying distances from the enhancer.

Silencers
DNA sequences at which transcription factors bind that decrease the transcription rate of one or more target genes that can be at varying distances from the silencer.

Insulators
DNA sequences that interfere with enhancer and/or silencer activity.

Locus control regions
Regulatory elements that generally control transcription of multiple genes in a single locus.



Motifs from JASPAR database

Model name	Score	Relative score	Start	End	Strand	Predicted site sequence
IRF1	12.986	0.904279917181229	3	14	–1	GAAACCGAAACG
IRF2	17.216	0.907706906384892	4	21	–1	CGGAAGTGAAACCGAAAC
SPIB	4.820	0.806987596140569	5	11	–1	ACCGAAA
BRCA1	4.228	0.802287513481405	8	14	–1	GAAACCG

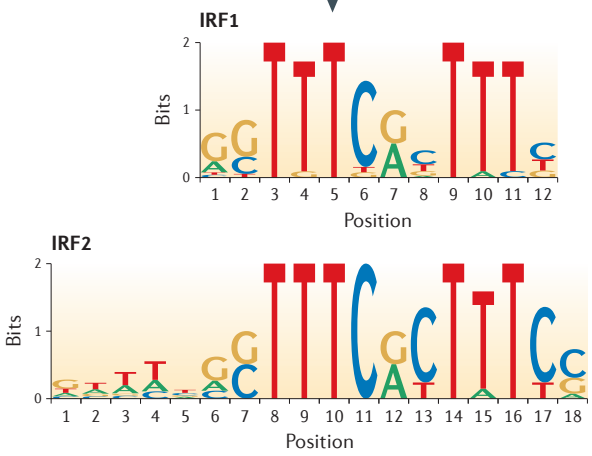


Figure 3 | **DNaseI footprints correspond to bound proteins.** The distribution of DNaseI digestion sites with DNaseI hypersensitive regions is not uniform; peaks and troughs occur in the signal, where troughs are due to the protection of DNA sequences by bound proteins. Transcription factor binding motif databases such as JASPAR[81] can be searched using the sequence from each footprint to predict what factor is bound. Shown here are data from the proximal promoter region of the human fragile-X mental retardation 1 (*FMR1*) gene, with motif-matching results for one footprint indicating that potentially bound factors are interferon regulatory factor 1 (IRF1) or IRF2. DNaseI footprints had been identified previously at this locus[110] in lymphoblastoid cells. More recent data from DNase–seq was used to recapitulate these results in a single experiment[12]. The upper panel is modified, with permission, from REF 12 © (2010) Cold Spring Harbor Laboratory Press.

compared with ChIP–seq are that they can identify genomic locations bound by proteins that are uncharacterized or for which antibodies do not exist. However, standard open chromatin analysis does not allow the determination of which protein or proteins are present in these regions.

Nucleosome positioning experiments such as MNase–seq[77,78] use micrococcal nuclease (MNase) digestion to determine where nucleosomes are present and, by extension, nucleosome-free regions. For large genomes, such as the human genome, MNase–seq may not be as economically practical because >90% of the genome is nucleosomal. Considerably greater sequencing coverage is required for MNase–seq to obtain the same level of resolution of nucleosome-free regions as the open chromatin assays described above.

DNaseI footprinting. Within a DHS site there are smaller, more-focal areas of DNaseI protection, called DNaseI footprints (FIG. 3), which result from the binding of individual proteins or complexes. Single-site DNaseI footprinting has been used to identify binding sites at individual loci for over 30 years[79], and DNase–seq now enables the discovery of footprints genome-wide[7,11–13].

Two different basic strategies have been used for predicting protein binding sites using DNaseI footprints in DNase–seq data. The first tries initially to delineate individual footprints solely based on the distribution of the sequence reads; a depletion of the 5′ ends of reads within the footprint would be expected compared with the immediately adjacent, non-footprint bases. This strategy has been used in the yeast and human genomes to identify 8–30 bp footprint regions of significantly reduced DNaseI digestion compared to a random background distribution[11,14] and in the human genome using a hidden Markov model (HMM) to model the characteristic changes in sequence read density in footprints[12]. To predict what factor might be bound at each identified footprint, transcription factor binding motif databases such as TRANSFAC[80], JASPAR[81] and UniPROBE[82] can be scanned using the sequence in the footprint. Footprints can also be used to identify DNA binding motifs for novel transcription factors. A recent analysis of 41 diverse cell-types showed that approximately 90% of all motifs in TRANSFAC, JASPAR, and UniPROBE could be identified using footprinted sequences, and an additional 289 distinct motifs could be defined[14]. Comparing ChIP–seq data with motifs in footprints also provides the ability to estimate which sites are being directly versus indirectly bound by a factor[14]. As these are predictions, it is recommended that specific binding events are tested experimentally.

An alternative strategy, which is implemented in the CENTIPEDE software tool[13], essentially carries out the above steps in the reverse order. First, the genome is scanned to identify all potential binding sites for a given DNA-binding protein based on its motif. CENTIPEDE then uses an unsupervised Bayesian mixture model to predict which of these sites are bound by protein and which are not bound in a particular cell type. This probabilistic model uses evidence based

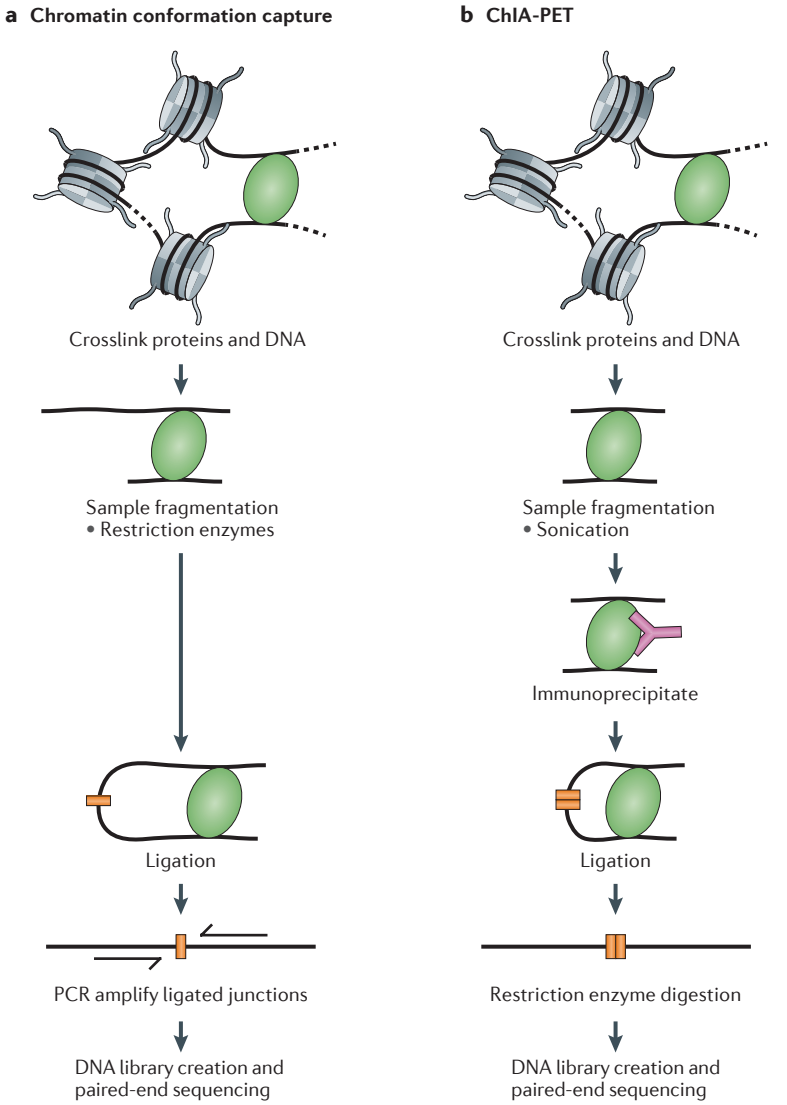


Figure 4 | **Detecting chromatin interactions.** In three-dimensional space, distal genomic regions on the same or different chromosomes interact, and this can be mediated by one or more DNA-binding proteins. **a** | Chromatin conformation capture experiments use a ligation step to join distant fragments that are interacting in three-dimensional chromatin space, thus providing information on possible targets for DNA-bound proteins. **b** | Chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) similarly detects chromatin interactions using a ligation step to pair non-adjacent interacting regions. However, ChIA-PET uses a chromatin immunoprecipitation (ChIP) step to more specifically identify interactions with a particular bound protein, such as RNA polymerase II. It should be noted that the DNA that is actually sequenced as part of the paired-end sequencing does not necessarily correspond to the precise region of interaction but is dictated by the presence of restriction enzyme targets.

primarily on DNaseI digestion, but can also incorporate evidence from the evolutionary conservation of bases and the presence of histone modifications, if those data are available. A second analysis in this study[13] using all 10-mers that were enriched in DHS sites predicted 49 novel motifs not found in existing motif databases, demonstrating that CENTIPEDE can also find binding sites of undefined factors.

A comparison of the accuracies of the two methods has not been carried out. The first method may be more appropriate for a more global annotation of potential binding sites regardless of the existence of a motif, whereas CENTIPEDE provides a more straightforward method to identify footprints for particular factors with known binding site preferences. Both methods are constrained by sequencing depth — which can limit their ability to identify footprints in the DHS sites that have reduced signals in DNase–seq data — and by the lack of knowledge of binding site preferences for factors. Increased sequencing depths will enable further refinement of footprint models. As DNaseI footprint annotations are generated for more cell types, motif-finding algorithms may help to predict new factor-binding motifs that in turn will help with the annotation of footprints.

Mapping chromatin interactions

Identifying protein–DNA binding sites is important, but that by itself does not lead to an understanding of the regulatory programs and other biological processes in cells. ChIP–seq, DNase–seq, and FAIRE–seq do not map each bound protein to the target gene or genes that it is helping to regulate, nor to the genomic region or regions with which it is interacting to form a higher order chromatin structure. Towards this end, approaches have been developed that are based on the 3C method[15]. This method has been extended to improve the scope and/or precision, resulting in methods known as chromatin conformation capture carbon copy (5C)[16] and Hi-C[17]. Furthermore, 3C has been adapted to identify interactions that are associated with specific proteins, resulting in the ChIA-PET sequencing method[18,19].

The principal steps of chromatin conformation capture experiments (FIG. 4a) are to: crosslink genomic regions that are in close proximity (analogous to the crosslinking that is used in ChIP–seq to find DNA–protein interactions); digest the DNA using restriction enzymes to create pairs of crosslinked DNA fragments that originated from distinct genomic locations; and identify these pairs of fragments (for example, using paired-end sequencing after the ligation and amplification of the fragments). 3C experiments require PCR primers that are designed for regions of interest and thus are low-throughput. However, designing primers for promoter regions of genes and for regulatory regions that have been identified through ChIP–seq or DNase–seq experiments can identify potential interactions between specific bound proteins and their target genes. 5C experiments simultaneously use thousands of primers in one experiment to detect millions of interactions[16]. 5C is still limited in the size of the genomic region that can be assayed, both by the number of primers that are incorporated and by sequencing depth to confidently detect interactions. 5C was used to analyse a 400 kb region that included the human β-globin locus and was able to confirm known interactions between regulatory elements and genes in the locus, as well as to identify new looping interactions[16]. Hi-C does not depend on primers but instead incorporates biotinylated residues after restriction enzyme digestion that allow

these fragments to be pulled down using streptavidin beads and the detection of interactions genome-wide. Extremely deep sequencing is required to confidently identify all interactions. Although this represents a substantial increase in throughput, the resolution is limited to a megabase scale owing to the frequency of restriction sites in the genome[83]. This limits the ability to confidently associate individual factor binding sites with target genes. A recent study showed that Hi-C was able to identify correctly interaction domains in the mouse and human *HOXA* locus that are separated by a known CTCF insulator element[83]. Thus, chromosome conformation information can provide boundaries for potential factor–gene interactions.

ChIA-PET (FIG. 4b) also starts with formaldehyde-based crosslinking, but this is followed by fragmentation by sonication and an immunoprecipitation step using a specific antibody, as is done in a ChIP experiment. DNA ligase is added to create chimeric DNA fragments, followed by restriction enzyme digestion and paired-end tag sequencing. ChIP–seq experiments for the factor of interest are also carried out to support the interaction data and to annotate where the factor is bound.

ChIA-PET provides genome-wide high-resolution data for interactions that involve a given DNA-binding protein. An initial study of the ERα protein revealed that ERα binding sites are involved in long-range looping interactions to gene promoters, and these interactions affect transcription rates[84]. Knockdown of ERα by short interfering RNA (siRNA) led to at least some of the interactions disappearing and transcriptional regulation being affected. As with Hi-C, the resolution of ChIA-PET is limited by the frequency and distribution of restriction enzyme digestion sites. Because ChIA-PET relies on an antibody targeted to the factor of interest, as for ChIP an increase in available antibodies will increase the scope of interactions that can be discovered by this method.

Data from ChIA-PET and 5C experiments are available in the University of California, Santa Cruz (UCSC) Genome Browser, which provides a visual representation of the sequenced paired-end tags. Together, the chromatin conformation capture and ChIA-PET technologies offer the ability to generate evidence of what genes are being targeted by DNA-bound proteins and regions with specific histone modifications.

Variation in protein binding

ChIP–seq, DNase–seq, and chromatin interaction experiments generate complex data sets that reflect the dynamic nature of the biological processes being measured. The results of these experiments provide a snapshot of varying chromatin states and protein binding events across millions of cells that are subject to genetic and environmental influences. Signals from these data reveal a spectrum of intensities, but the molecular underpinnings of this variation — among loci in the genome of an individual and among multiple individuals — remains unclear. Using data from these experiments, we can begin to understand both types of variation more completely.

Variation across loci. DNA-binding proteins can generally interact with a range of DNA sequences, giving rise to a sequence 'motif' to describe the binding preference of a protein. A motif, which is often more specifically defined as a position weight matrix, describes the nucleotide preferences (usually defined as probabilities) at each position in a binding site. These probabilities are usually based on the frequency at which each nucleotide is present in known binding sites that have been identified across the genome. It is generally thought that the presence of the higher probability nucleotides at a locus indicates an increase in binding affinity and/or specificity. Binding affinity refers to the strength of an interaction and is generally specified in terms of a dissociation constant, whereas binding specificity refers to the preference for binding to specific sequences. Higher affinity or specificity sites may be expected to generate higher signals in protein-binding assays owing to increased occupancy and/or stability of the interaction.

Several high-throughput methods are now available to determine binding specificities of proteins in an unbiased manner (see REF. 85 for a more detailed review). Protein-binding microarrays have been developed that contain all possible 10 bp sequences[86] and have been used, for example, to determine the binding specificities for 104 diverse factors in the mouse[87]. The binding preferences of factors are largely unique, and approximately half of the factors show preferences for two motifs. More recently, a similar study was carried out in *D. melanogaster* using the novel method protein–DNA binding followed by sequencing (PB–seq). In this approach, the protein of interest — in this case, heat shock factor (HSF) — was fused to the 3×FLAG epitope and allowed to bind to fragmented DNA. The HSF-bound DNA was recovered and sequenced[88]. This study compared the binding preferences of HSF defined by PB–seq *in vitro* to binding sites defined by ChIP–seq *in vivo*. Interestingly, *in vitro* and *in vivo* binding intensities were not highly correlated when all possible binding sites in the genome were considered. A chromatin environment data model was then generated using available DNaseI hypersensitivity data, MNase data and ChIP–chip data for 21 histone modifications, and this model was used with the *in vitro* results to predict binding intensities. This resulted in a high correlation with *in vivo* data, underscoring the influence of chromatin on protein–DNA binding. In fact, a prior model based solely on DNaseI data produced the highest correlation, suggesting that DNA accessibility generally corresponds to the actual binding of factors *in vivo*.

Chromatin is dynamic and has substantial, stable differences between phenotypically different cell types and also smaller, more variable differences across a population of similar cells. ChIP–seq and other protein-binding experiments provide a snapshot of the occupancy of binding sites, but do not describe the dynamics or function of factor binding. Competition ChIP assays[89,90] have enabled the investigation of binding site turnover in yeast. These studies integrated into a single strain two copies of a factor-encoding gene; the two copies had different epitope tags, and one copy was constitutively expressed whereas the other was inducible.

Hidden Markov model
(HMM). A statistical model consisting of states that represent an aspect of a sequence (such as in a footprint), which transitions between states; it is used to label bases in a sequence with the modelled property. HMMs are also used in many gene prediction programs.

Bayesian mixture model
A probabilistic model that is used to represent the presence of multiple subpopulations (such as DNaseI footprints) within the whole population (such as the whole genome sequence). Bayesian mixture models allow for the incorporation of prior knowledge about subpopulation frequencies.

Biotinylated
A protein or nucleic acid to which a small biotin molecule has been attached. Biotin binds to streptavidin, thus allowing for the isolation of biotinylated molecules.

Dissociation constant
A constant that reflects the amount of energy that is required to separate two interacting molecules, often referred to as K_d.

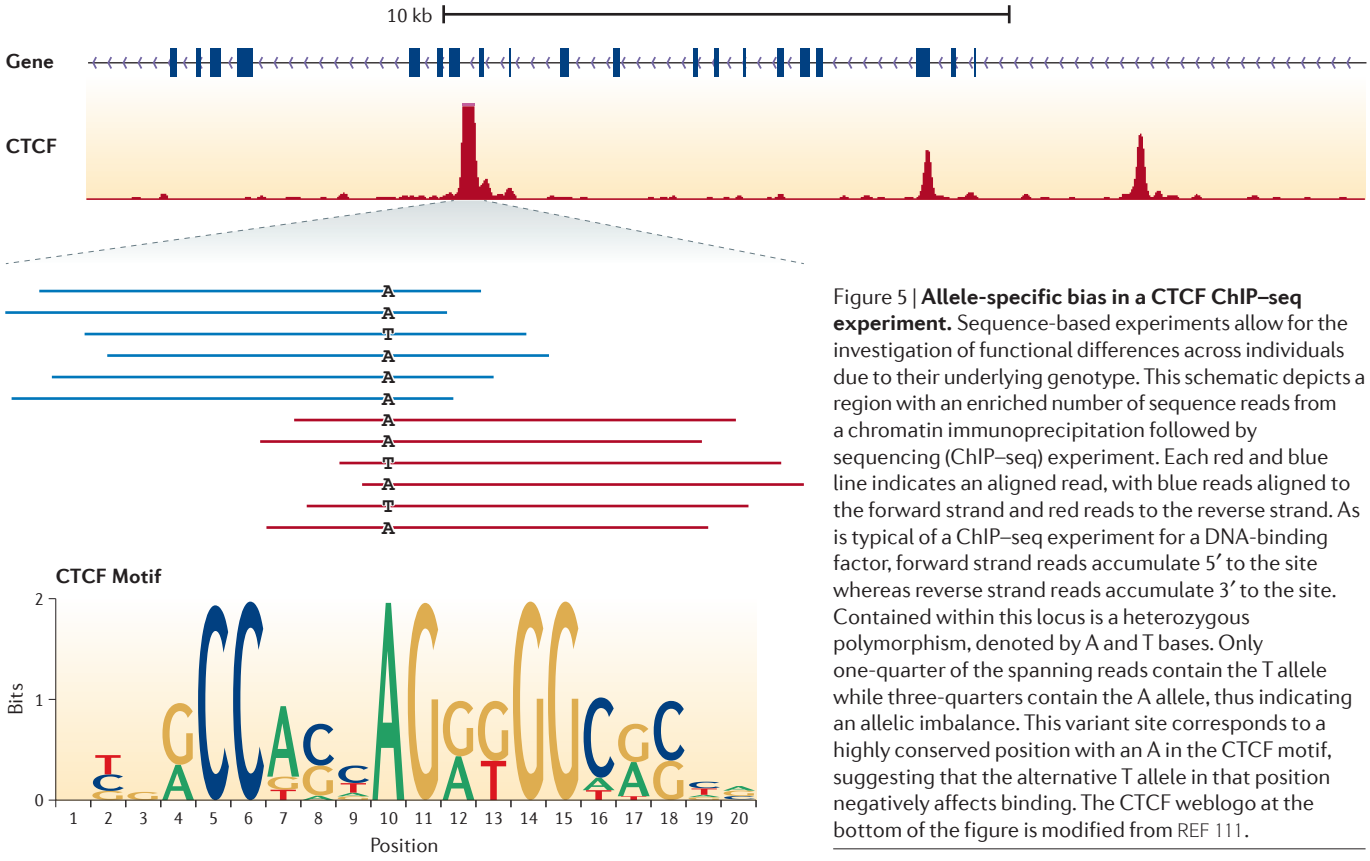


Figure 5 | **Allele-specific bias in a CTCF ChIP–seq experiment.** Sequence-based experiments allow for the investigation of functional differences across individuals due to their underlying genotype. This schematic depicts a region with an enriched number of sequence reads from a chromatin immunoprecipitation followed by sequencing (ChIP–seq) experiment. Each red and blue line indicates an aligned read, with blue reads aligned to the forward strand and red reads to the reverse strand. As is typical of a ChIP–seq experiment for a DNA-binding factor, forward strand reads accumulate 5′ to the site whereas reverse strand reads accumulate 3′ to the site. Contained within this locus is a heterozygous polymorphism, denoted by A and T bases. Only one-quarter of the spanning reads contain the T allele while three-quarters contain the A allele, thus indicating an allelic imbalance. This variant site corresponds to a highly conserved position with an A in the CTCF motif, suggesting that the alternative T allele in that position negatively affects binding. The CTCF weblogo at the bottom of the figure is modified from REF 111.

ChIP for each epitope was carried out on samples collected at multiple time points after the induction of the inducible gene to show the dynamics of factor binding (specifically to show at which sites there is stable binding and at which there is turnover). A study[91] of the Rap1 transcription factor showed that sites stably bound by the same factor (resident sites) were associated with efficient transcriptional activation, whereas high-turnover sites (treadmilling sites) were associated with lower transcriptional output, even under similar rates of occupancy.

These studies demonstrate that binding sites across a genome are not functionally equivalent and reveal influences on this variation. Complementary information about factor binding, chromatin state and binding dynamics provides a more complete picture of how protein–DNA interactions at particular loci contribute to cellular processes.

Variation across individuals. The adaptation of ChIP and other experiments to sequencing technologies also provides the opportunity to investigate potential functional effects of the underlying DNA sequence on the presence or absence of a particular event, such as the binding of a protein. Polymorphic bases within regulatory regions can affect the stability of a bound protein or the ability of a region to acquire or propagate chromatin marks. These, in turn, can affect the ability of that locus to regulate the transcription of its target gene.

To identify polymorphic sites that are associated with functional variation, we can investigate sequences in individual ChIP–seq peaks that align across a heterozygous base in a particular sample; a significant difference in the distribution of sequences containing one allele versus the other indicates a potential allelic effect on protein binding (FIG. 5). For example, given ChIP–seq data for transcription factor F, we can investigate each heterozygous site that falls within a called peak (binding site) in that data. For a site with alleles A and B, if the presence of A or B has no effect, we would expect an even distribution of sequences containing A and B at that binding site. If sequences at that site predominantly contain allele A, we could hypothesize that A provides a more favourable binding sequence for that protein, or conversely that B interferes with binding.

Allelic analysis of sequence data requires modifications to the standard analysis pipelines described above (FIG. 2). Aligning short-read sequences to a single reference sequence creates a bias at heterozygous loci where reads containing the allele present in the reference genome are aligned at a higher rate owing to the inherent 'mismatch' penalty incurred by the non-reference allele sequences. Ideally, sequences would be aligned to fully defined haplotype genomes, as described in the AlleleSeq computation pipeline[21]. These are rarely available, but more often the genotype of each individual has been obtained. This can be used to create two reference genomes, each one containing one allele for

DNaseI-sensitivity quantitative trait loci (dsQTL). A locus whose sensitivity to DNaseI digestion varies based on the presence of different alleles in that locus. An allelic difference may influence the binding of proteins at this locus, causing the variation in digestion.

each heterozygous location, thus enabling the merging of separate alignments of sequences to each of these genome sequences. Alternatively, allele-aware aligners such as the Genomic Short-read Nucleotide Alignment Program (GSNAP)[92] can be used that dynamically consider multiple alleles during alignments. In addition, the alignability of a sequence containing each variant must be considered. The presence of allele A may make a particular sequence unique with respect to the rest of the genome, whereas allele B of that sequence might be found one or more times elsewhere in the genome. This can be determined by aligning all possible sequences overlapping the site of interest back to the genome and analysing the uniqueness of these alignments. Overall, a more careful consideration of non-reference-sequence bases is necessary to accurately detect signals at these locations.

Allelic biases have been detected in data from several sequencing-based experiments, including ChIP–seq[20,93–96] and DNase–seq[22,24]. In one study, analysis of ChIP–seq data from 10 human lymphoblastoid cell lines showed that the occupancy of 7.5% of nuclear factor-κB (NF-κB) binding sites and 25% of Pol II binding sites differed significantly between individuals, and that 35% and 26% of these corresponded with genetic variations, respectively[20]. Another study, also using human lymphoblastoid cells, found that 7% of DHS sites and 11% of CTCF binding sites showed allele-specific effects[22]. Both studies were carried out on family trios that showed evidence of the heritability of these allelic functional traits. A more recent study of DNase–seq and expression data from lymphoblastoid cell lines from 70 individuals uncovered just under 9,000 DNaseI-sensitivity quantitative trait loci (dsQTLs) for which genetic variants with allelic biases in DHS sites are associated with changes in expression levels of nearby genes[24]. Many dsQTLs could also be mapped to previously identified DNaseI footprints[12,13], suggesting that the binding of specific factors is altered. Analysis of the footprints with predicted binding factors showed that there was enrichment for allelic biases in CTCF binding sites, cAMP-response-elements (CREs) and interferon-stimulated response elements (ISREs), but depletion for allelic biases in myocyte-specific enhancer factor 2A (MEF2A) sites.

Perspective

The importance of DNA-binding proteins has motivated the continued development of experimental and analytical methods to better identify and characterize these interactions. ChIP–seq remains the standard for identifying binding site locations for individual proteins and histone modifications. However, practical limitations of antibody development, the limit of a single factor or modification per experiment, the lack of functional annotation and the static snapshots of a dynamic cell that are provided necessitate the use of complementary methods or extensions of ChIP–seq to provide a more complete picture of biological processes in the cell, especially transcriptional regulation.

Open chromatin assays, such as DNase–seq and FAIRE–seq, provide a more comprehensive status of all active regulatory elements in a single experiment. Comparisons of changes in open chromatin profiles across cell types[6,7,97,98], differentiation states[99,100], disease states[101–104] and species[105] are revealing key changes in factor binding that underlie functional differences across cells. Reduced sequencing costs are enabling a deeper coverage in these experiments, thus uncovering more-precise positioning of bound proteins in the form of footprints.

Identifying the genomic locations of protein–DNA interactions is just the start. Bound proteins interact with other proteins in complexes, create higher order chromatin structures, are involved in specific cellular processes (such as the regulation of a particular gene) and vary across time, cell types and genetic background. Answering these questions requires complementary assays, many of which are presented here. As data from complementary assays accumulate, the challenge will be to integrate these to provide a more complete understanding of transcriptional networks and cellular processes[106,107]. Comparisons across cell types will provide new insights into the properties of individual factors (and their combinations) that drive cell-type-specific functions. These will require the further development of new analytical and computational modelling techniques, as well as focused validation experiments to support model hypotheses.

Results from these studies continue to further our understanding of normal cell biology, but also provide crucial information that will benefit efforts to determine the causes and consequences of abnormal cellular states that are associated with disease. Genome-wide association studies in humans have identified thousands of loci that are strongly associated with a complex disease or a related trait[108], most of which are located in non-coding genomic regions and lack functional annotation. Characterizing the effects of different SNP alleles on DNA–protein interactions provide potential functional consequences of the alleles. These can then be used to suggest testable hypotheses for observed associations of individual SNPs with complex diseases, potentially leading to the development of better diagnoses and treatment options.

1. Bhinge, A. A., Kim, J., Euskirchen, G. M., Snyder, M. & Iyer, V. R. Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE). *Genome Res.* **17**, 910–916 (2007).
2. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods* **5**, 829–834 (2008).
3. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnol.* **26**, 1351–1359 (2008).
4. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell.* **132**, 311–322 (2008).
5. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* **2010**, pdb.prot5384 (2010).
6. Song, L. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* **21**, 1757–1767 (2011).
7. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
8. Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17**, 877–885 (2007).
9. Giresi, P. G. & Lieb, J. D. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of

Regulatory Elements). *Methods* **48**, 233–239 (2009).
10. Simon, J. M., Giresi, P. G., Davis, I. J. & Lieb, J. D. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nature Protoc.* **7**, 256–267 (2012).
11. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods* **6**, 283–289 (2009).
12. Boyle, A. P. *et al.* High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.* **21**, 456–464 (2010).
13. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447–455 (2010).
14. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
This paper describes the identification and analysis of 8.4 million DNaseI footprints across 41 human cell types corresponding to putative factor binding events and predicting ~300 novel motifs for factor binding.
15. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
This paper described the first general approach to characterize interactions between any two genomic loci and provided the first glimpse of the three-dimensional structure of chromatin in the nucleus.
16. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).
17. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
18. Li, G. *et al.* ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.* **11**, R22 (2010).
19. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
20. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).
This paper demonstrated that functional variation in transcription factor binding due to differences in genotype could be uncovered using data from ChIP–seq experiments.
21. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7**, 522 (2011).
22. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**, 235–239 (2010).
This paper similarly demonstrated that differences in chromatin structure due to genotype variation could be seen using data from DNase–seq data.
23. Gertz, J. *et al.* Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLoS Genet.* **7**, e1002228 (2011).
24. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
25. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Rev. Genet.* **10**, 669–680 (2009).
26. Farnham, P. J. Insights from genomic profiling of transcription factors. *Nature Rev. Genet.* **10**, 605–616 (2009).
27. Ku, C. S., Naidoo, N., Wu, M. & Soong, R. Studying the epigenome using next generation sequencing. *J. Med. Genet.* **48**, 721–730 (2011).
28. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
29. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
This paper provides practical guidelines for conducting and analysing ChIP–seq experiments.
30. Egelhofer, T. A. *et al.* An assessment of histone-modification antibody quality. *Nature Struct. Mol. Biol.* **18**, 91–93 (2011).
31. Fuchs, S. M., Krajewski, K., Baker, R. W., Miller, V. L. & Strahl, B. D. Influence of combinatorial histone modifications on antibody and effector protein recognition. *Curr. Biol.* **21**, 53–58 (2011).
32. Adli, M. & Bernstein, B. E. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nature Protoc.* **6**, 1656–1668 (2011).
33. Shankaranarayanan, P. *et al.* Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. *Nature Methods* **8**, 565–567 (2011).
34. Liu, C. L., Schreiber, S. L. & Bernstein, B. E. Development and validation of a T7 based linear amplification for genomic DNA. *BMC Genomics* **4**, 19 (2003).
35. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–1419 (2011).
This paper describes a modification to the traditional ChIP–seq protocol that allows for greater resolution in identifying the binding sites of factors. The key advance is the use of an exonuclease to generate more consistent signals of binding locations.
36. Markham, K., Bai, Y. & Schmitt-Ulms, G. Co-immunoprecipitations revisited: an update on experimental concepts and their implementation for sensitive interactome investigations of endogenous proteins. *Anal. Bioanal. Chem.* **389**, 461–473 (2007).
37. Brinkman, A. B. *et al.* Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res.* **22**, 1128–1138 (2012).
38. Statham, A. L. *et al.* Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. *Genome Res.* **22**, 1120–1127 (2012).
39. Havugimana, P. C. *et al.* A census of human soluble protein complexes. *Cell* **150**, 1068–1081 (2012).
40. Butland, G. *et al.* Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433**, 531–537 (2005).
41. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
42. Gavin, A. C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
43. Guruharsha, K. G. *et al.* A protein complex network of *Drosophila melanogaster*. *Cell* **147**, 690–703 (2011).
44. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
45. Hu, P. *et al.* Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. *PLoS Biol.* **7**, e96 (2009).
46. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).
47. Kuhner, S. *et al.* Proteome organization in a genome-reduced bacterium. *Science* **326**, 1235–1240 (2009).
48. Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform.* **11**, 473–483 (2010).
49. Kim, H. *et al.* A short survey of computational analysis methods in analysing ChIP-seq data. *Hum. Genom.* **5**, 117–123 (2011).
50. Wilbanks, E. G. & Facciotti, M. T. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* **5**, e11471 (2010).
51. Malone, B. M., Tan, F., Bridges, S. M. & Peng, Z. Comparison of four ChIP-Seq analytical algorithms using rice endosperm H3K27 trimethylation profiling data. *PLoS ONE* **6**, e25260 (2011).
52. Laajala, T. D. *et al.* A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics* **10**, 618 (2009).
53. Gao, D. *et al.* A survey of statistical software for analysing RNA-seq data. *Hum. Genom.* **5**, 56–60 (2010).
54. Kvam, V. M., Liu, P. & Si, Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am. J. Bot.* **99**, 248–256 (2012).
55. Guo, Y., Mahony, S. & Gifford, D. K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.* **8**, e1002638 (2012).
56. Boeva, V. *et al.* De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res.* **38**, e126 (2010).
57. Wu, S., Wang, J., Zhao, W., Pounds, S. & Cheng, C. ChIP-PaM: an algorithm to identify protein-DNA interaction using ChIP-Seq data. *Theor. Biol. Med. Model.* **7**, 18 (2010).
58. Hu, M., Yu, J., Taylor, J. M., Chinnaiyan, A. M. & Qin, Z. S. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.* **38**, 2154–2167 (2010).
59. Kulakovskiy, I. V., Boeva, V. A., Favorov, A. V. & Makeev, V. J. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics.* **26**, 2622–2623 (2012).
60. Georgiev, S. *et al.* Evidence-ranked motif identification. *Genome Biol.* **11**, R19 (2010).
61. Taub, M. A., Corrada Bravo, H. & Irizarry, R. A. Overcoming bias and systematic errors in next generation sequencing data. *Genome Med.* **2**, 87 (2010).
62. Chen, Y. *et al.* Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods* **9**, 609–614 (2012).
63. Khrameeva, E. E. & Gelfand, M. S. Biases in read coverage demonstrated by interlaboratory and interplatform comparison of 117 mRNA and genome sequencing experiments. *BMC Bioinformatics.* **13**, S4 (2012).
64. Schwartz, S., Oren, R. & Ast, G. Detection and removal of biases in the analysis of next-generation sequencing reads. *PLoS ONE* **6**, e16685 (2011).
65. Cheung, M. S., Down, T. A., Latorre, I. & Ahringer, J. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res.* **39**, e103 (2011).
66. Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* **12**, R112 (2011).
67. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72 (2012).
68. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39**, e90 (2011).
69. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Rev. Genet.* **13**, 36–46 (2011).
70. Wang, J., Huda, A., Lunyak, V. V. & Jordan, I. K. A. Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics* **26**, 2501–2508 (2010).
71. Chung, D. *et al.* Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. *PLoS Comput. Biol.* **7**, e1002111 (2011).
72. Bell, O., Tiwari, V. K., Thoma, N. H. & Schubeler, D. Determinants and dynamics of genome accessibility. *Nature Rev. Genet.* **12**, 554–564 (2011).
73. Wu, C., Bingham, P. M., Livak, K. J., Holmgren, R. & Elgin, S. C. The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. *Cell.* **16**, 797–806 (1979).
74. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).
75. Cockerill, P. N. Structure and function of active chromatin and DNase I hypersensitive sites. *FEBS J.* **278**, 2182–2210 (2011).
76. Crawford, G. E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16**, 123–131 (2006).
This paper describes the first DNaseI hypersensitivity experiments that used high-throughput sequencing technology.
77. Wei, G., Hu, G., Cui, K. & Zhao, K. Genome-wide mapping of nucleosome occupancy, histone modifications, and gene expression using next-generation sequencing technology. *Methods Enzymol.* **513**, 297–313 (2012).
78. Wal, M. & Pugh, B. F. Genome-wide mapping of nucleosome positions in yeast using high-resolution MNase ChIP-Seq. *Methods Enzymol.* **513**, 233–250 (2012).
79. Galas, D. J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).
80. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
81. Bryne, J. C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* **36**, D102–D106 (2008).
82. Newburger, D. E. & Bulyk, M. L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **37**, D77–D82 (2009).

83. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
84. Fullwood, M. J. & Ruan, Y. ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell. Biochem.* **107**, 30–39 (2009).
85. Stormo, G. D. & Zhao, Y. Determining the specificity of protein–DNA interactions. *Nature Rev. Genet.* **11**, 751–760 (2010).
86. Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnol.* **24**, 1429–1435 (2006).
87. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
88. Guertin, M. J., Martins, A. L., Siepel, A. & Lis, J. T. Accurate prediction of inducible transcription factor binding intensities in vivo. *PLoS Genet.* **8**, e1002610 (2012).
This paper describes a method that showed the importance of chromatin state dynamics, in addition to sequence preferences, in the DNA-binding intensities of proteins.
89. Dion, M. F. *et al.* Dynamics of replication-independent histone turnover in budding yeast. *Science* **315**, 1405–1408 (2007).
90. van Werven, F. J., van Teeffelen, H. A., Holstege, F. C. & Timmers, H. T. Distinct promoter dynamics of the basal transcription factor TBP across the yeast genome. *Nature Struct. Mol. Biol.* **16**, 1043–1048 (2009).
91. Lickwar, C. R., Mueller, F., Hanlon, S. E., McNally, J. G. & Lieb, J. D. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature* **484**, 251–255 (2012).
This paper provides evidence for a model of transcription factor binding in which factors are either stably bound and promote consistent transcription, or are 'treadmilling' through bound and unbound states resulting in lower transcription rates.
92. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
This paper describes a short-read sequence aligner that can simultaneously align to multiple DNA sequence variants. This removes the bias that results from using a single reference genome, in which sequences containing alleles present in the reference genome are better-aligned, whereas sequences containing non-reference alleles are penalized.
93. Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M. & Snyder, M. Genetic analysis of variation in transcription factor binding in yeast. *Nature* **464**, 1187–1191 (2010).
94. Marks, H. *et al.* High-resolution analysis of epigenetic changes associated with X inactivation. *Genome Res.* **19**, 1361–1373 (2009).
95. Motallebipour, M. *et al.* Differential binding and co-binding pattern of FOXA1 and FOXA3 and their relation to H3K4me3 in HepG2 cells revealed by ChIP-seq. *Genome Biol.* **10**, R129 (2009).
96. Yildirim, E., Sadreyev, R. I., Pinter, S. F. & Lee, J. T. X-chromosome hyperactivation in mammals via nonlinear relationships between chromatin states and transcription. *Nature Struct. Mol. Biol.* **19**, 56–61 (2011).
97. Gaulton, K. J. *et al.* A map of open chromatin in human pancreatic islets. *Nature Genet.* **42**, 255–259 (2010).
98. Bischof, J. M. *et al.* A genome-wide analysis of open chromatin in human tracheal epithelial cells reveals novel candidate regulatory elements for lung function. *Thorax* **67**, 385–391 (2011).
99. Waki, H. *et al.* Global mapping of cell type-specific open chromatin by FAIRE-seq reveals the regulatory role of the NFI family in adipocyte differentiation. *PLoS Genet.* **7**, e1002311 (2011).
100. Wu, W. *et al.* Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome Res.* **21**, 1659–1671 (2011).
101. Stitzel, M. L. *et al.* Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell. Metab.* **12**, 443–455 (2010).
102. Magnani, L., Ballantyne, E. B., Zhang, X. & Lupien, M. PBX1 genomic pioneer function drives ERα signaling underlying progression in breast cancer. *PLoS Genet.* **7**, e1002368 (2011).
103. Parker, S. C. *et al.* Mutational signatures of de-differentiation in functional non-coding regions of melanoma genomes. *PLoS Genet.* **8**, e1002871 (2012).
104. He, H. H. *et al.* Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Res.* **22**, 1015–1025 (2012).
105. Shibata, Y. *et al.* Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genet.* **8**, e1002789 (2012).
106. Cheng, C. *et al.* Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput. Biol.* **7**, e1002190 (2011).
107. Muino, J. M., Angenent, G. C. & Kaufmann, K. Visualizing and characterizing in vivo DNA-binding events and direct target genes of plant transcription factors. *Methods Mol. Biol.* **754**, 293–305 (2011).
108. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
109. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **40**, D13–D25 (2012).
110. Drouin, R. *et al.* Structural and functional characterization of the human FMR1 promoter reveals similarities with the hnRNP-A2 promoter region. *Hum. Mol. Genet.* **6**, 2051–2060 (1997).
111. Essien, K. *et al.* CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome Biol.***10**, R131 (2009).

Acknowledgements
I gratefully acknowledge support from the US National Institutes of Health grants U54-HG004563, R21-DA027040 and U01 CA157703, the Department of Defense grant W81XWH-10-1-0772, and the University Cancer Research Fund from the University of North Carolina at Chapel Hill.

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

FURTHER INFORMATION
Terrence Furey's homepage: http://fureylab.web.unc.edu
Gene Expression Omnibus (GEO): http://www.ncbi.nlm.nih.gov/geo
Picard sequence analysis tools: http://picard.sourceforge.net
UCSC Genome Browser: http://genome.ucsc.edu

ALL LINKS ARE ACTIVE IN THE ONLINE PDF