APPLICATIONS OF NEXT-GENERATION SEQUENCING

ChIP–seq: advantages and challenges of a maturing technology

Peter J. Park

Abstract | Chromatin immunoprecipitation followed by sequencing (ChIP–seq) is a technique for genome-wide profiling of DNA-binding proteins, histone modifications or nucleosomes. Owing to the tremendous progress in next-generation sequencing technology, ChIP–seq offers higher resolution, less noise and greater coverage than its array-based predecessor ChIP–chip. With the decreasing cost of sequencing, ChIP–seq has become an indispensable tool for studying gene regulation and epigenetic mechanisms. In this Review, I describe the benefits and challenges in harnessing this technique with an emphasis on issues related to experimental design and data analysis. ChIP–seq experiments generate large quantities of data, and effective computational analysis will be crucial for uncovering biological mechanisms.

Nucleosome
The basic structural subunit of chromatin. A nucleosome consists of approximately 147 base pairs of DNA and an octamer of histone proteins.

Epigenome
The chromatin states that are found along the genome, defined for a given time point and cell type. Thus, for a given genome there may be hundreds or thousands of epigenomes, depending on the stability of the chromatin states.

DNase I hypersensitive site
A chromosomal region that is highly accessible to cleavage by DNase I. Such sites are associated with open chromatin conformations and transcriptional activity.

Harvard Medical School, 10 Shattuck Street, Boston, Massachusetts 02115, USA.
e-mail:
peter_park@harvard.edu
doi:10.1038/nrg2641
Published online
8 September 2009

Genome-wide mapping of protein–DNA interactions and epigenetic marks is essential for a full understanding of transcriptional regulation. A precise map of binding sites for transcription factors, core transcriptional machinery and other DNA-binding proteins is vital for deciphering the gene regulatory networks that underlie various biological processes[1]. The combination of nucleosome positioning and dynamic modification of DNA and histones has a key role in gene regulation[2–4] and guides development and differentiation[5]. Chromatin states can influence transcription directly by altering the packaging of DNA to allow or prevent access to DNA-binding proteins, or they can modify the nucleosome surface to enhance or impede recruitment of effector protein complexes. Recent advances suggest that this interplay between chromatin and transcription is dynamic and more complex than previously appreciated[6], and there is a growing recognition that systematic profiling of the epigenomes in multiple cell types and stages may be needed for understanding developmental processes and disease states[7].

The main tool for investigating these mechanisms is chromatin immunoprecipitation (ChIP), which is a technique for assaying protein–DNA binding *in vivo*[8]. In ChIP, antibodies are used to select specific proteins or nucleosomes, which enriches for DNA fragments that are bound to these proteins or nucleosomes. The introduction of microarrays allowed the fragments obtained from ChIP to be identified by hybridization to a microarray (ChIP–chip), therefore enabling a genome-scale view of DNA–protein interactions[9,10]. On high-density tiling arrays, oligonucleotide probes can now be placed across an entire genome or across selected regions of a genome — for instance, promoter regions, specific chromosomes or gene families — at a preferred resolution.

Owing to the rapid technological developments in next-generation sequencing (NGS), the arsenal of genomic assays available to the biologist has been transformed[11–13]. The ability to sequence tens or hundreds of millions of short DNA fragments in a single run is enabling increasingly large experiments that could only be imagined a few years ago. Next-generation sequencing has already been applied in many areas, including the sequencing of whole genomes[14,15], the sequencing of mRNA for gene expression profiling (RNA–seq)[16–18], the characterization of structural variation[19], the profiling of DNase I hypersensitive sites[20], the detection of fusion genes from mRNA transcripts[21] and the discovery of new classes of small RNAs[22]. If the 'third-generation' sequencing technologies that are under development deliver as promised, they will lead to another epoch of genome-scale investigations[23].

Chromatin immunoprecipitation followed by sequencing (ChIP–seq) was one of the early applications of NGS, and the first studies to use it were published in 2007 (REFS 24–27). In ChIP–seq, the DNA fragments of interest are sequenced directly instead of being hybridized on an array. ChIP–seq has higher resolution, fewer artefacts, greater coverage and a larger dynamic range

Box 1 | The contribution of ChIP–seq to epigenome mapping

The enhanced spatial resolution afforded by next-generation sequencing improves the characterization of binding sites for transcription factors and other DNA-binding proteins and enables the identification of sequence motifs. The increased precision is especially important for profiling nucleosome-level features, and it allows the systematic cataloguing of patterns of histone modifications, histone variants and nucleosome positioning. Here, I briefly describe recent chromatin immunoprecipitation (ChIP) studies that have enabled progress in the characterization of epigenomes.

Histone modification maps

The first comprehensive genome-wide maps produced through ChIP followed by sequencing (ChIP–seq) were created in 2007. Twenty histone methylation marks, as well as the histone variant H2A.Z, RNA polymerase II and the DNA-binding protein CTCF (CCCTC-binding factor), were profiled using the Solexa 1G platform in human T cells[25] with an average of ~8 million tags per sample. This was followed by a map of 18 histone acetylation marks in the same cell type[90]. These studies suggested novel functions for histone modification and the importance of combinatorial patterns of modifications. To examine the role of histone modifications in differentiation, embryonic stem cells have also been profiled. Several histone lysine trimethylation modifications were profiled in mouse embryonic stem cells and two types of differentiated cells in 2007 (REF. 27). This study showed that bivalent domains[91] have a role in lineage potential and identified marks for imprinting control. Before ChIP–seq, genome-wide modification profiles were available for yeast using tiling arrays[92–94], but only selected regions had been profiled for mice and humans (see REF. 35 for further descriptions of the techniques used).

Nucleosome maps

Using ChIP followed by microarray (ChIP–chip), nucleosome depletion at active promoters in yeast was described in 2004 (REF. 95). This was followed by a high-resolution study[96] in 2005 and a complete map of nucleosome positioning[97] in 2007. In *Caenorhabditis elegans*, micrococcal nuclease digestion followed by sequencing was used in 2006 to map core nucleosomes[98]. ChIP–seq with Roche 454 pyrosequencing was used to generate a map of the histone variant H2A.Z in yeast[99] in 2007 and in flies[100] in 2008. For human cells, epigenetically modified and bulk mono-nucleosome positions were profiled for T cells in 2007 and 2008[25,30,90] with >140 million reads per experiment using the Illumina Solexa platform (reviewed in REF. 2). These studies have revealed the role of nucleosomes in transcriptional regulation and hint at the principles that guide nucleosome positioning.

than ChIP–chip and therefore provides substantially improved data. Although the short reads (~35 bp) generated by NGS platforms pose serious difficulties for certain applications — for example, *de novo* genome assembly — they are acceptable for ChIP–seq. The more precise mapping of protein-binding sites provided by ChIP–seq allows for a more accurate list of targets for transcription factors and enhancers, in addition to better identification of sequence motifs[24,28]. Enhanced spatial resolution is particularly important for profiling histone variants, post-translational modifications of chromatin and nucleosome positioning, and ChIP–seq has enabled tremendous progress in these areas (BOX 1).

In this Review, I describe the advantages and challenges in applying ChIP–seq. I discuss various issues in experimental design, including sample quality, controls, depth of sequencing and the number of replicates. Given the large quantities of data generated by ChIP–seq, computational analysis — including the identification of binding sites and their subsequent analysis — poses a substantial challenge for most laboratories; therefore I also discuss the main issues in data processing and statistical analysis.

Bivalent domain
A region of chromatin marked by a histone modification associated with active transcription (histone H3 lysine 4 trimethylation) and a modification associated with repression (histone H3 lysine 27 trimethylation). It is postulated to mark genes that are silent but poised for transcription.

Imprinting
The differential expression of genes depending on whether they were inherited maternally or paternally.

ChIP–seq basics

In a ChIP experiment for DNA-binding proteins, DNA fragments associated with a specific protein are enriched (FIG. 1). The DNA-binding protein is crosslinked to DNA *in vivo* by treating cells with formaldehyde and the chromatin is sheared by sonication into small fragments, which are generally in the 200–600 bp range. An antibody specific to the protein of interest is used to immunoprecipitate the DNA–protein complex. Finally, the crosslinks are reversed and the released DNA is assayed to determine the sequences bound by the protein.

In ChIP experiments that aim to map nucleosome positions or histone modifications, micrococcal nuclease (MNase) digestion without crosslinking is most often used to fragment the chromatin. Although sonication has also been used in this context[29], MNase treatment is generally preferred because it removes linker DNA more efficiently than sonication and therefore allows more precise mapping of each nucleosome[30]. However, MNase digestion has a more pronounced sequence bias than sonication[31], and the solubility of chromatin also creates bias[32]. There may also be changes in nucleosome positions and histone modifications during the course of the experiment in the absence of crosslinking. ChIP with and without crosslinking is sometimes referred to as X-ChIP[33] and N-ChIP[34], respectively, in which X denotes 'crosslinking' and N denotes 'native'.

During the construction of a sequencing library, the immunoprecipitated DNA is subjected to size selection (typically in the ~150–300 bp range, although there seems to be a bias towards shorter fragments in sequencing). Nearly all ChIP–seq data have been generated through the Illumina Genome Analyzer, although other platforms, such as Applied Biosystems' SOLiD and the Helicos platform, are now available (FIG. 1). The Genome Analyzer and SOLiD platforms currently generate 100–400 million reads in a single run, and ~60–80% of reads can be aligned uniquely to the genome.

Advantages and disadvantages of ChIP–seq

ChIP–seq offers many advantages over ChIP–chip, as summarized in TABLE 1 (see also REF. 35). First, its base pair resolution is perhaps the greatest improvement over ChIP–chip, as shown in FIG. 2a. Although arrays can be tiled at a high density, this requires a large number of probes and remains expensive for mammalian genomes[36]. Arrays also have fundamental limitations in resolution due to the uncertainties in the hybridization process. Second, ChIP–seq does not suffer from the noise generated by the hybridization step in ChIP–chip. Nucleic acid hybridization is complex and dependent on many factors, including the GC content, length, concentration and secondary structure of the target and probe sequences. Therefore, cross-hybridization between imperfectly matched sequences frequently occurs and contributes to the noise. Third, the intensity signal measured on arrays might not be linear over its entire range, and its dynamic range is limited below and above saturation points. In a recent study, distinct and biologically meaningful peaks seen in ChIP–seq were obscured when the same experiment was conducted with ChIP–chip[37]. Finally, in

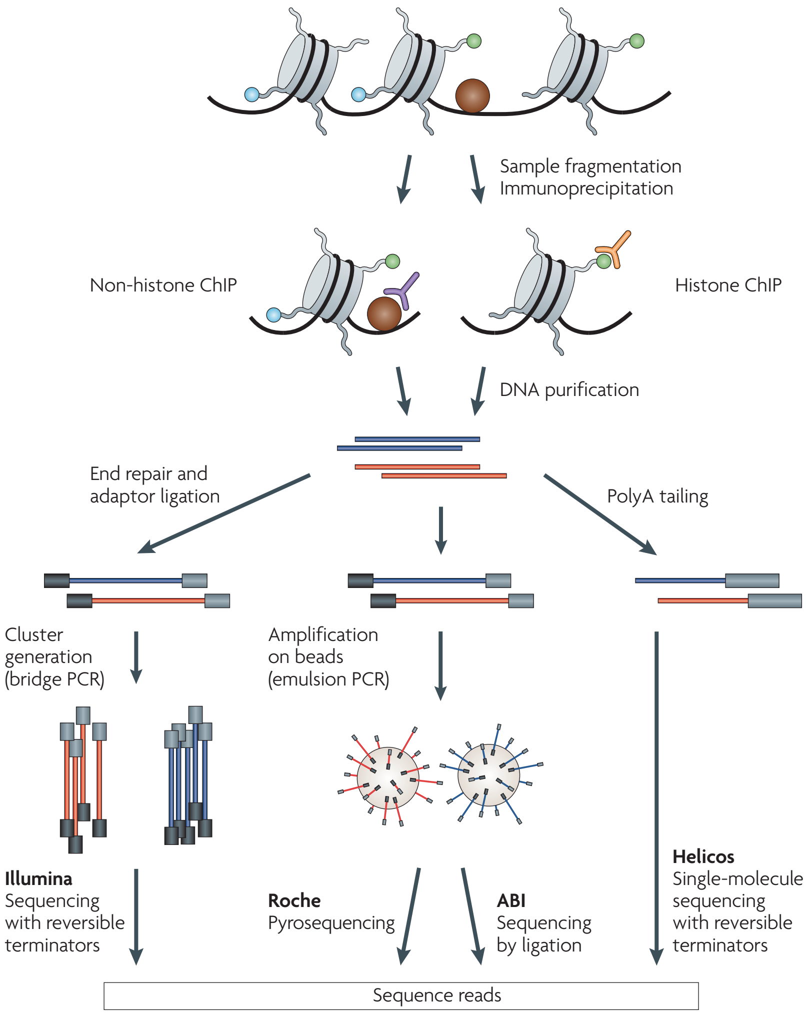

Figure 1 | **Overview of a ChIP–seq experiment.** Using chromatin immunoprecipitation (ChIP) followed by massively parallel sequencing, the specific DNA sites that interact with transcription factors or other chromatin-associated proteins (non-histone ChIP) and sites that correspond to modified nucleosomes (histone ChIP) can be profiled. The ChIP process enriches the crosslinked proteins or modified nucleosomes of interest using an antibody specific to the protein or the histone modification. Purified DNA can be sequenced on any of the next-generation platforms[12]. The basic concepts are similar for different platforms: common adaptors are ligated to the ChIP DNA and clonally clustered amplicons are generated. The sequencing step involves the enzyme-driven extension of all templates in parallel. After each extension, the fluorescent labels that have been incorporated are detected through high-resolution imaging. On the Illumina Solexa Genome Analyzer (bottom left), clusters of clonal sequences are generated by bridge PCR, and sequencing is performed by sequencing-by-synthesis. On the Roche 454 and Applied Biosystems (ABI) SOLiD platforms (bottom middle), clonal sequencing features are generated by emulsion PCR and amplicons are captured on the surface of micrometre-scale beads. Beads with amplicons are then recovered and immobilized to a planar substrate to be sequenced by pyrosequencing (for the 454 platform) or by DNA ligase-driven synthesis (for the SOLiD platform). On single-molecule sequencing platforms such as the HeliScope by Helicos (bottom right), fluorescent nucleotides incorporated into templates can be imaged at the level of single molecules, which makes clonal amplification unnecessary.

Heterochromatin
A region of highly compact chromatin. Constitutive heterochromatin is largely composed of repetitive DNA.

ChIP–seq the genome coverage is not limited by the repertoire of probe sequences fixed on the array. This is particularly important for the analysis of repetitive regions of the genome, which are typically masked out on arrays. Studies involving heterochromatin or microsatellites, for instance, can be done much more effectively by ChIP–seq. Sequence variations within repeat elements can be captured by sequencing and used to map reads to the genome; unique sequences that flank repeats are also helpful in aligning the reads to the genome. For example, only 48% of the human genome is non-repetitive, but 80% is mappable with 30 bp reads and 89% is mappable with 70 bp reads[38].

All profiling technologies produce unwanted artefacts, and ChIP–seq is no exception. Although sequencing errors have been reduced substantially as the technology has improved, they are still present, especially towards the end of each read. This problem can be ameliorated by improvements in alignment algorithms (see below) and computational analysis. There is also bias towards GC-rich content in fragment selection, both in library preparation and in amplification before and during sequencing[14,39], although notable improvements have been made recently. In addition, when an insufficient number of reads is generated, there is loss of sensitivity or specificity in detection of enriched regions. There are also technical issues in performing the experiment, such as loading the correct amount of sample: too little sample will result in too few tags; too much sample will result in fluorescent labels that are too close to one another, and therefore lower quality data.

However, the main disadvantage with ChIP–seq is its current cost and availability. Several groups have successfully developed and applied their own protocols for library construction, which has lowered that cost substantially. But the overall cost of ChIP–seq, which includes machine depreciation and reagent cost, will have to be lowered further for it to be comparable with the cost of ChIP–chip in every case. For high-resolution profiling of an entire large genome, ChIP–seq is already less expensive than ChIP–chip, but depending on the genome size and the depth of sequencing needed, a ChIP–chip experiment on carefully selected regions using a customized microarray may yield as much biological understanding. The recent decrease in sequencing cost per base pair has not affected ChIP–seq as substantially as other applications, as the decrease has come as much from increased read lengths as from the number of sequenced fragments. The gain in the fraction of reads that can be uniquely aligned to the genome decreases noticeably after ~25–35 bp and is marginal beyond 70–100 nucleotides[40]. However, as the cost of sequencing continues to decline and institutional support for sequencing platforms continues to grow, ChIP–seq is likely to become the method of choice for nearly all ChIP experiments in the near future.

Issues in experimental design

Antibody quality. The value of any ChIP data, including ChIP–seq data, depends crucially on the quality of the antibody used. A sensitive and specific antibody will give a high level of enrichment compared with the background, which makes it easier to detect binding events. Many antibodies are commercially available, and some are noted as ChIP grade, but the quality of different antibodies is highly variable and can also vary among batches

Microsatellite
A class of repetitive DNA that is made up of repeats that are 2–8 nucleotides in length.

RNA interference
The process by which the introduction or expression within cells of single- or double-stranded RNA leads to the degradation of mRNA and therefore to gene suppression.

of a specific antibody. Rigorous validation is a laborious process: for histone modifications for instance, the reactivity of the antibody with unmodified histones or nonhistone proteins should be checked by western blotting. Furthermore, cross-reactivity with similar histone modifications (for example, dimethylation compared with trimethylation at the same residue) should be checked by using two independent antibodies in combination with RNA interference against enzymes that are predicted to add the modifying group or with mass spectrometry of the precipitated peptides. As part of the model organism ENCyclopedia Of DNA Elements (modENCODE) project[41], I have been involved in the large-scale profiling of histone modifications for *Drosophila melanogaster*, and the antibody validation procedure for this project (which uses the steps described above) has resulted in the finding that 20–35% of the commercially produced antibodies tested were unsatisfactory.

Sample quantity. One advantage of ChIP–seq over ChIP–chip is the smaller amount of sample material needed. A typical ChIP experiment requires ~10^7 cells and yields 10–100 ng of DNA. Several ChIP protocols have been developed that use smaller numbers of cells — for example, 10^4–10^5 cells for genome-wide profiling[42] or 10^2–10^3 cells for PCR quantification at specific loci[43–45] — but to work they require abundant transcription factors or histone modifications (such as RNA polymerase II or histone H3 trimethylated at lysine 27 (H3K27me3)) and a high-quality antibody. For ChIP–chip, the ChIP sample is usually amplified to generate >2 μg of DNA per array. By contrast, for ChIP–seq on the Illumina platform, 10–50 ng of DNA is recommended. Furthermore, fewer rounds of amplification are required for ChIP–seq, so the potential for artefacts due to PCR bias is lower. The precise amount of ChIP DNA and the number of cells needed depend on the abundance of the chromatin-associated protein targets or histone modifications, in addition to the quality of the antibody. ChIP–seq without amplification is possible on the Helicos True Single Molecule Sequencing platform[46] and other 'third-generation' platforms that are in development (FIG. 1).

Control experiment. The experimental steps in ChIP involve several potential sources of artefacts. Shearing of DNA, for example, does not result in uniform fragmentation of the genome: open chromatin regions tend to be fragmented more easily than closed regions, which creates an uneven distribution of sequence tags across the genome. Also, repetitive sequences might seem to be enriched because of inaccuracies in the number of copies of the repeats in the assembled genome. Therefore, a peak in the ChIP–seq profile should be compared with the same region in a matched control sample to determine its significance. There are three commonly used types of control sample: input DNA (a portion of the DNA sample removed prior to immunoprecipitation (IP)); mock IP DNA (DNA obtained from IP without antibodies); and DNA from nonspecific IP (IP performed using an antibody, such as immunoglobulin G, against a protein that is not known to be involved in DNA binding or chromatin modification). These types of control sample test for different types of artefacts, and there is no consensus on which is the most appropriate. Input DNA has been used as the control sample in nearly all ChIP–seq studies; comparison with input DNA corrects for bias related to the variable solubility of different regions, the shearing of DNA and amplification. One problem with using a mock IP sample is that very little material can be pulled down in the absence of an antibody and therefore the results of multiple mock IPs may not be consistent. In one set of ChIP–chip experiments, the mock IP control was found to contribute little to the overall result when the data were properly normalized[47]. When analysing histone modifications, using the ratio between data from the ChIP sample and from the bulk nucleosomes is

Table 1 | **Comparison of ChIP–chip and ChIP–seq**

	ChIP–chip	**ChIP–seq**
Maximum resolution	Array-specific, generally 30–100 bp	Single nucleotide
Coverage	Limited by sequences on the array; repetitive regions are usually masked out	Limited only by alignability of reads to the genome; increases with read length; many repetitive regions can be covered
Cost	US$400–800 per array (1–6 million probes); multiple arrays may be needed for large genomes	Currently US$1,000–2,000 per lane (using the Illumina Genome Analyzer); 6–15 million reads before alignment
Source of platform noise	Cross-hybridization between probes and nonspecific targets	Some GC bias can be present
Experimental design	Single- or double-channel, depending on the platform	Single channel
Cost-effective cases	Profiling of selected regions; when a large fraction of the genome is enriched for the modification or protein of interest (broad binding)	Large genomes; when a small fraction of the genome is enriched for the modification or protein of interest (sharp binding)
Required amount of ChIP DNA	High (a few micrograms)	Low (10–50 ng)
Dynamic range	Lower detection limit; saturation at high signal	Not limited
Amplification	More required	Less required; single-molecule sequencing without amplification is available
Multiplexing	Not possible	Possible

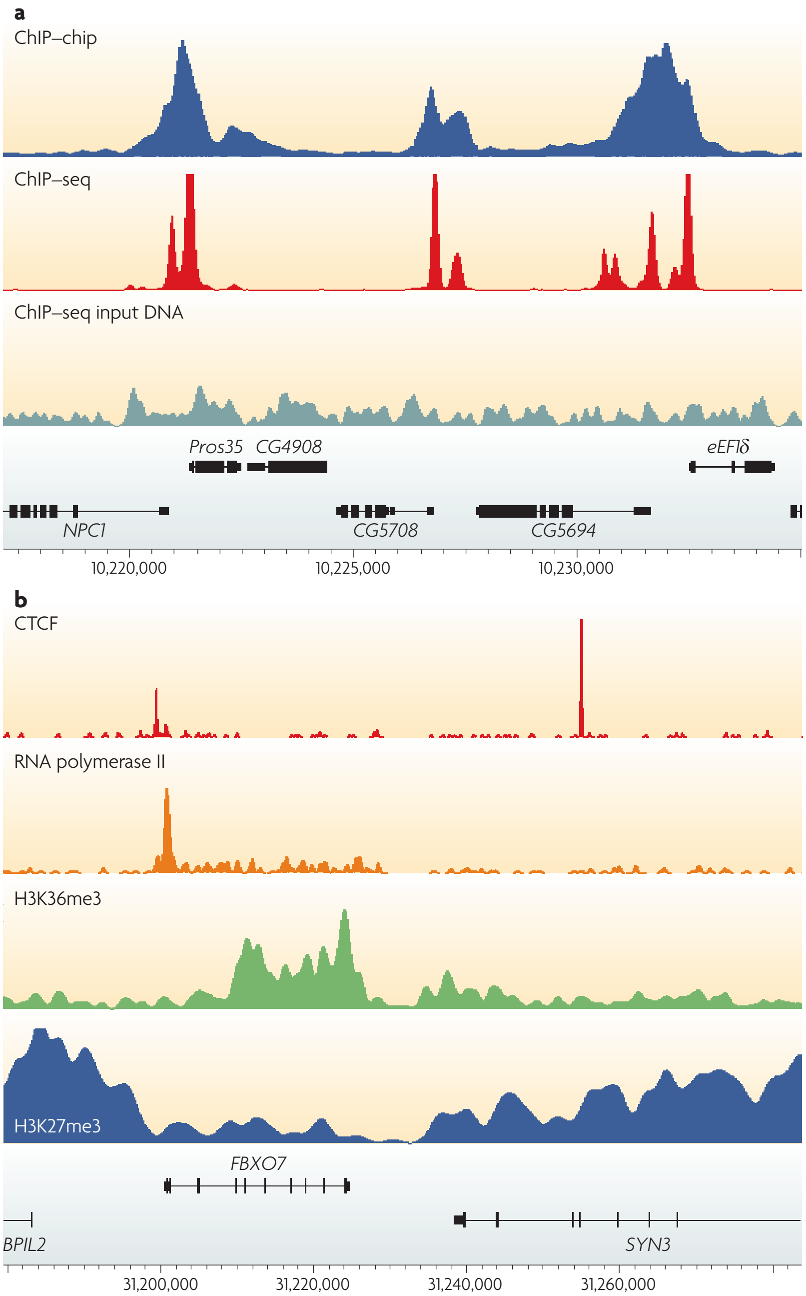

Figure 2 | **ChIP profiles. a** | Examples of the profiles generated by chromatin immunoprecipitation followed by sequencing (ChIP–seq) or by microarray (ChIP–chip). Shown is a section of the binding profiles of the chromodomain protein Chromator, as measured by ChIP–chip (unlogged intensity ratio; blue) and ChIP–seq (tag density; red) in the *Drosophila melanogaster* S2 cell line. The tag density profile obtained by ChIP–seq reveals specific positions of Chromator binding with higher spatial resolution and sensitivity. The ChIP–seq input DNA (control experiment) tag density is shown in grey for comparison. **b** | Examples of different types of ChIP–seq tag density profiles in human T cells. Profiles for different types of proteins and histone marks can have different types of features, such as: sharp binding sites, as shown for the insulator binding protein CTCF (CCCTC-binding factor; red); a mixture of shapes, as shown for RNA polymerase II (orange), which has a sharp peak followed by a broad region of enrichment; medium size broad peaks, as shown for histone H3 trimethylated at lysine 36 (H3K36me3; green), which is associated with transcription elongation over the gene; or large domains, as shown for histone H3 trimethylated at lysine 27 (H3K27me3; blue), which is a repressive mark that is indicative of Polycomb-mediated silencing. *BPIL2*, bactericidal/permeability-increasing protein-like 2; *FBXO7*, F box only 7; *NPC1*, Niemann-Pick disease, type C1; *Pros35*, proteasome 35 kDa subunit; *SYN3*, synapsin III. Data for part **b** are from REF. 25.

also informative, as this ratio corresponds to the fraction of nucleosomes with the particular modification at that location, averaged over all the cells assayed.

One of the difficulties in conducting a ChIP–seq control experiment is the large amount of sequencing that may be necessary. For input DNA and bulk nucleosomes, many of the sequenced tags are spread evenly across the genome. To obtain accurate estimates throughout the genome, sufficient numbers of tags are needed at each point; otherwise fold enrichment at the peaks will result in large errors due to sampling bias. Therefore, the total number of tags to be sequenced is potentially very large. Alternatively, it is possible to avoid sequencing a control sample if one is only interested in differential binding patterns between conditions or time points and if the variation in chromatin preparations is small.

Depth of sequencing. One crucial difference between ChIP–chip and ChIP–seq is that the number of tiling arrays that is used in a ChIP–chip experiment is fixed regardless of the protein or modification of interest, whereas the number of fragments that is sequenced in a ChIP–seq experiment is determined by the investigator. In published ChIP–seq experiments, a single lane of the Illumina Genome Analyzer was the basic unit of sequencing. When it was introduced, a single lane generated 4–6 million reads before alignment but, owing to improvements in the system, a single lane now generates 8–15 million reads or more. Given the cost of each experiment, many early data sets contained reads from a single lane regardless of what the specific experiment was. Intuitively, one expects that when a large number of binding sites are present in the genome for a DNA-binding protein or when a histone modification covers a large fraction of the genome, a correspondingly large number of tags will be needed to cover each bound region at the same tag density. One reasonable criterion for determining sufficient sequencing depth would be that the results of a given analysis do not change when more reads are obtained. In terms of the number of binding sites, this criterion translates to the presence of a 'saturation point' after which no further binding sites are discovered with additional reads.

The issue of saturation points has been examined in a recent paper through simulation studies[48]. In three example data sets, a reference set of sites was generated based on the full set of sequencing reads in each case. Then, a wide range of different read counts was sampled from the complete data set, with multiple random selections for each sample size. Binding sites were determined for each sample with a threshold probability (p value), and the results for each sample size were averaged. The fraction of the reference set that was recovered as a function of the number of reads is shown in FIG. 3A. If there was a saturation point, the number of sites found would increase up to a certain point and then plateau, which would indicate that the rate at which new sites were being discovered had slowed down to the point where any further increase in the number of reads would be inefficient at yielding new sites. When the simulation was performed, however, the results indicated that

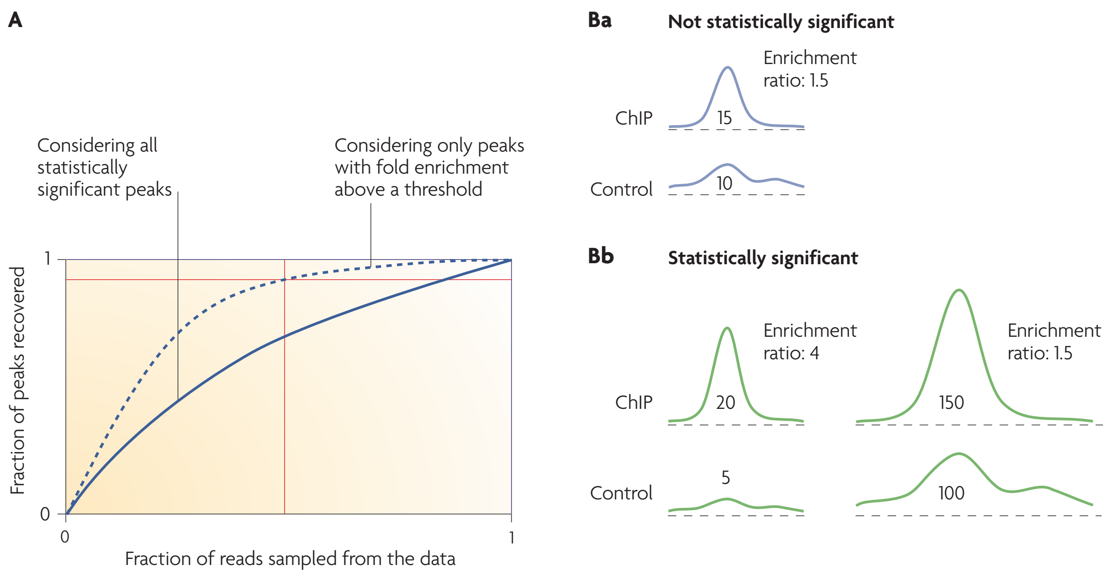

Figure 3 | **Depth of sequencing. A** | To determine whether enough tags have been sequenced, a simulation can be carried out to characterize the fraction of the peaks that would be recovered if a smaller number of tags had been sequenced. In many cases, new statistically significant peaks are discovered at a steady rate with an increasing number of tags (solid curve) — that is, there is no saturation of binding sites. However, when a minimum threshold is imposed for the enrichment ratio between chromatin immunoprecipitation (ChIP) and input DNA peaks, the rate at which new peaks are discovered slows down (dashed curve) — that is, saturation of detected binding sites can occur when only sufficiently prominent binding positions are considered. For a given data set, multiple curves corresponding to different thresholds can be examined to identify the threshold at which the curve becomes sufficiently flat to meet the desired saturation criteria (defined by the intersection of the orange lines on the graph). We refer to such a threshold as the minimum saturation enrichment ratio (MSER). The MSER can serve as a measure for the depth of sequencing achieved in a data set: a high MSER, for example, might indicate that the data set was undersampled, as only the more prominent peaks were saturated (see REF. 48 for details). **Ba** | A peak that is not statistically significant — the enrichment ratio between the ChIP and control experiments is low (1.5) and the number of tag counts (shown under the peaks) is also low. **Bb** | Two ways in which a peak can be statistically significant. On the left, although the number of tag counts is low, the enrichment ratio between the ChIP and control experiments is high (4). On the right, the peaks have the same enrichment ratio as those in **a** but have a larger number of tag counts; this example shows that continued sequencing might lead to less prominent peaks becoming statistically significant and that there might not necessarily be a saturation point after which no further binding sites are discovered.

more and more sites continued to be found at a steady pace with additional sequencing (FIG. 3A, lower curve). In another study[38], human RNA polymerase II targets were shown to saturate quickly, but for signal transducer and activator of transcription 1 (STAT1), the number of targets continued to rise steadily. This suggests that, at least in some cases, there might not be a saturation point that can be used to determine the number of tags to be sequenced if peaks are found based on statistical significance.

However, a saturation point does exist if a fixed threshold is imposed on the fold enrichment between the peaks in the ChIP experiment and the peaks in the control experiment — that is, saturation occurs when only prominent peaks (as defined by minimum fold enrichment) are considered. When all peaks are considered, even peaks with small enrichment can become statistically significant as more tags accumulate (FIG. 3B) and therefore the number of significant peaks may continue to rise with more sequencing. This is similar to what happens in genome-wide association studies and other genomic investigations in which a large sample size increases the statistical power and causes features that have small effect sizes to attain statistical significance. In the study discussed above[48], we proposed that each ChIP–seq data set could be annotated with a minimal saturated enrichment ratio (MSER) — a point at which saturation occurs — to give a sense of the sequencing depth achieved. We also found that there is a linear relationship between the number of reads and the MSER, when properly scaled. This makes it possible to predict how many more reads are needed when a particular level of MSER is desired. Although these concepts and tools should be tested on more data sets, they provide a framework for understanding depth-of-sequencing issues in ChIP–seq experiments.

Multiplexing. For small genomes, including those of *Saccharomyces cerevisiae, Caenorhabditis elegans* and *D. melanogaster*, the number of reads generated in a sequencing unit (for example, one of eight lanes on an Illumina Genome Analyzer) may be several times greater than the number of reads needed to provide sufficient coverage of the genome at a suitable depth

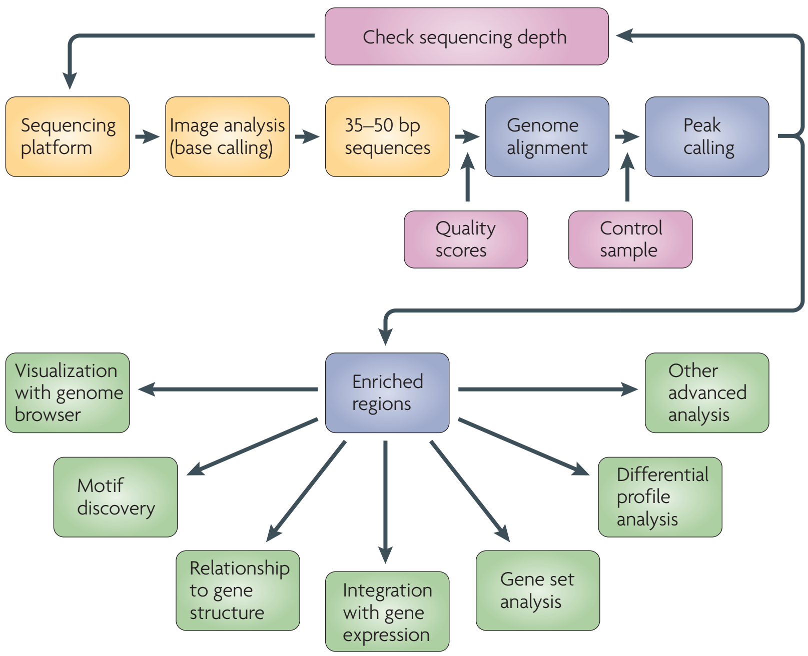

Figure 4 | **Overview of ChIP–seq analysis.** The raw data for chromatin immunoprecipitation followed by sequencing (ChIP–seq) analysis are images from the next-generation sequencing platform (top left). A base caller converts the image data to sequence tags, which are then aligned to the genome. On some platforms, they are aligned with the aid of quality scores that indicate the reliability of each base call. Peak calling, using data from the ChIP profile and a control profile (which is usually created from input DNA), generates a list of enriched regions that are ordered by false discovery rate as a statistical measure. Subsequently, the profiles of enriched regions are viewed with a browser and various advanced analyses are performed.

for the ChIP–seq experiment. As the number of reads per run continues to increase, the ability to sequence multiple samples at the same time (referred to as 'multiplexing') becomes important for cost effectiveness. In theory, multiplexing of samples is not difficult and only requires different barcode adaptors to be ligated to different samples during sample preparation. Even allowing for sequencing errors, a few bases are sufficient to serve as unique identifiers for many samples. In practice, however, multiplexing has not been widely used so far on the Illumina platform owing to uneven coverage of the samples and other technical problems. However, some recent protocols show promise[49], and multiplexing is likely to be used frequently in the future.

Additional considerations. Although ChIP fragments are generally sequenced at the 5′ ends, they can also be sequenced at both ends, as is frequently done for detection of structural variations in the genome[19]. Paired-end sequencing can be used in conjunction with ChIP to provide additional specificity (especially when mapping repetitive regions) and to map long-range chromatin interactions[50].

ChIP experiments should be replicated to ensure reproducibility of the data. For microarrays, platforms and protocols have improved substantially so that replicate experiments using the same samples are generally no longer needed. Although this is likely to become the case for ChIP–seq[51], replicate experiments are still recommended to account for variation between samples and to verify the fidelity of experimental steps. Assuming that they are sequenced deeply, two concordant replicate experiments are usually sufficient, as a third replicate seems to add little value[38].

Challenges in data analysis

As NGS platforms and ChIP–seq protocols mature, data generation is gradually becoming routine, and the limiting factor in a study is shifting to computational analysis of the data and to validation experiments. In this section, I discuss the key issues and concepts involved in data analysis. These concepts underpin a much wider range of ChIP–seq analysis techniques, which are too varied and complex to be discussed in this review. A flow chart of the steps involved in ChIP–seq analysis is shown in FIG. 4.

Data management. Next-generation sequencing produces an unprecedented amount of data. Raw data and images are on the order of terabytes per machine run, which makes data storage a challenge even for facilities with considerable expertise in the management of genomic data. Data can be stored at three levels: image data, sequence tags and alignment data. Ideally, the raw image data should be kept so that if a new base caller is developed the raw data can be reprocessed. Sequence tags can be used to map the data when an improved aligner is available or when a reference genome assembly is updated. Alignment data can be useful for generating summary statistics and can be used to generate SNP or copy number variation calls. There is no consensus in the community with regard to which data types should be stored, but many argue that the image data are too expensive to maintain and that a reasonable approach is to discard the raw data after a short period of time and keep only the sequence-level data.

In microarray-based studies, investigators are encouraged, and often required, to submit their data upon publication to a public database, such as Gene Expression Omnibus[52]. For NGS data, data transfer and maintenance are more complicated owing to the large file sizes. Depositing data through standard FTP or HTTP protocols, for instance, is likely to fail when many gigabytes are to be uploaded. To meet this challenge, the National Center for Biotechnology Information in the US, the European Bioinformatics Institute and the DNA Databank of Japan have developed the Sequence Read Archive[53,54]. To ensure that the archive is useful to the community, meta-data describing the details of each experiment should be submitted to the repositories at the same time as the sequencing data.

Genome alignment. Image processing and base calling are platform specific and are mostly done using the software provided by the sequencing platform manufacturer, although some new base callers have been proposed recently[55,56] for the Illumina platform. More important is the choice of strategy for genome

Poisson model
A probability distribution that is often used to model the number of random events in a fixed interval. Given an average number of events in the interval, the probability of a given number of occurrences can be calculated.

alignment, as all subsequent results are based on the aligned reads. Owing to the large number of reads, the use of conventional alignment algorithms can take hundreds or thousands of processor hours; therefore, a new generation of aligners has been developed[57], and more are expected soon. Every aligner is a balance between accuracy, speed, memory and flexibility, and no aligner can be best suited for all applications. Alignment for ChIP–seq should allow for a small number of mismatches due to sequencing errors, SNPs and indels or the difference between the genome of interest and the reference genome. This is simpler than in RNA–seq, for example, in which large gaps corresponding to introns must be considered. Popular aligners include: Eland, an efficient and fast aligner for short reads that was developed by Illumina and is the default aligner on that platform; Mapping and Assembly with Qualities (MAQ)[58], a widely used aligner with a more exhaustive algorithm and excellent capabilities for detecting SNPs; and Bowtie[59], an extremely fast mapper that is based on an algorithm that was originally developed for file compression. These methods use the quality score that accompanies each base call to indicate its reliability. For the SOLiD di-base sequencing technology, in which two consecutive bases are read at a time, modified aligners have been developed[60,61]. Many current analysis pipelines discard non-unique tags, but studies involving the repetitive regions of the genome[27,62–64] require careful handling of these non-unique tags.

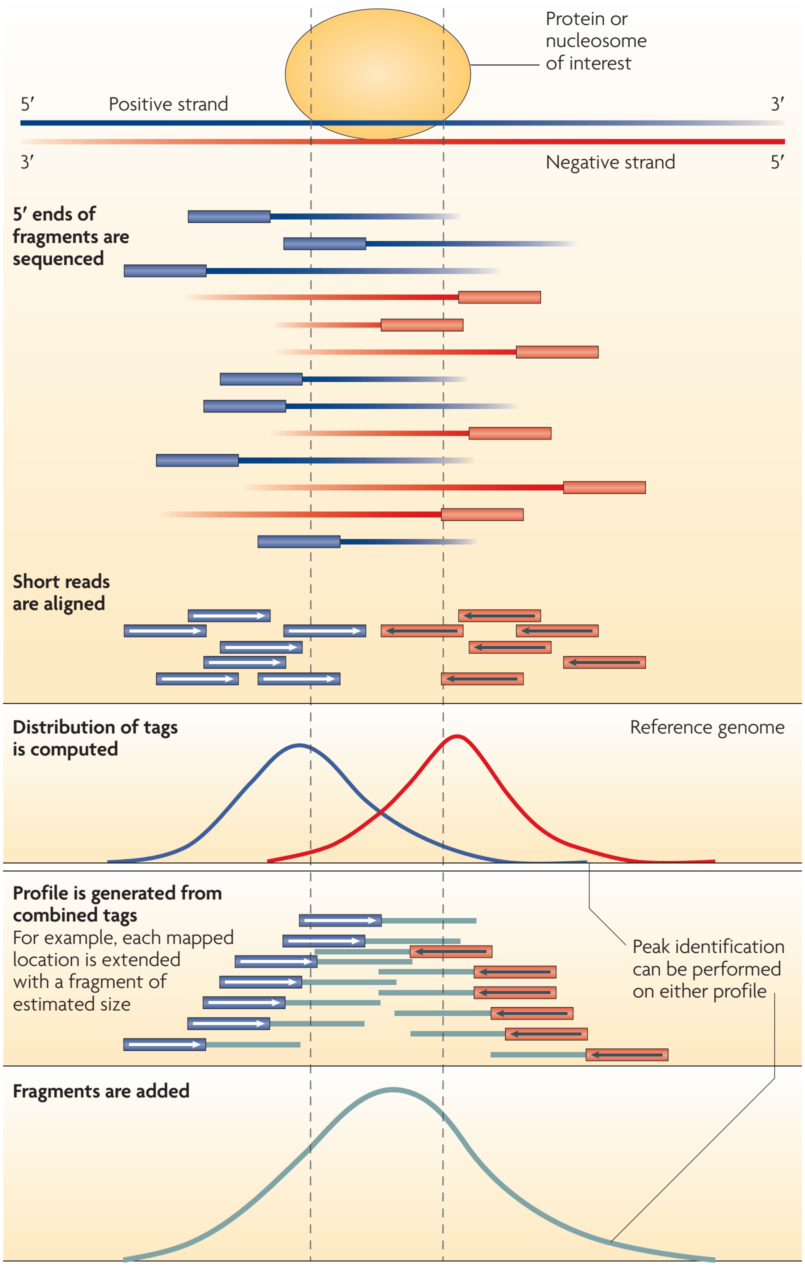

Figure 5 | **Strand-specific profiles at enriched sites.** DNA fragments from a chromatin immunoprecipitation experiment are sequenced from the 5′ end. Therefore, the alignment of these tags to the genome results in two peaks (one on each strand) that flank the binding location of the protein or nucleosome of interest. This strand-specific pattern can be used for the optimal detection of enriched regions. To create an approximate distribution of all fragments, each tag location can be extended by an estimated fragment size in the appropriate orientation and the number of fragments can be counted at each position.

Identification of enriched regions. After sequenced reads are aligned to the genome, the next step is to identify regions that are enriched in the ChIP sample relative to the control with statistical significance.

Several 'peak callers' that scan along the genome to identify the enriched regions are currently available[24,26,38,48,65–70]. In early algorithms, regions were scored by the number of tags in a window of a given size and then assessed by a set of criteria based on factors such as enrichment over the control and minimum tag density. Subsequent algorithms take advantage of the directionality of the reads[71]. As shown in FIG. 5, the fragments are sequenced at the 5′ end, and the locations of mapped reads should form two distributions, one on the positive strand and the other on the negative strand, with a consistent distance between the peaks of the distributions. In these methods, a smoothed profile of each strand is constructed[65,72] and the combined profile is calculated either by shifting each distribution towards the centre or by extending each mapped position into an appropriately oriented 'fragment' and then adding the fragments together. The latter approach should result in a more accurate profile with respect to the width of the binding, but it requires an estimate of the fragment size as well as the assumption that fragment size is uniform.

Given a combined profile, peaks can be scored in several ways. A simple fold ratio of the signal for the ChIP sample relative to that of the control sample around the peak (FIG. 3B) provides important information, but it is not adequate. A fold ratio of 5 estimated from 50 and 10 tags (from the ChIP and control experiments, respectively) has a different statistical significance to the same ratio estimated from, for example, 500 and 100 tags. A Poisson model for the tag distribution is an effective approach that accounts for the ratio as well as the absolute tag numbers[27], and it can also be modified to account for regional bias in tag density due to the chromatin structure, copy number variation or amplification bias[67].

A binomial distribution or other models can also be used[38]. In another approach, the peaks are scored before a combined profile is generated by considering how well the tag distributions on the two strands resemble each other and whether the distance between the peaks is close to the expected number of base pairs[48]. Another important local correction, regardless of the peak detection method, is to adjust for sequence alignability. Depending on how the non-uniquely mapped reads are processed, regions of the genome containing repetitive elements will have a different expected tag count. By keeping track of how many times each sequence of given length along a segment appears in the rest of the genome, one can correct for the variation in mappability among segments[27,38].

A major difficulty in identifying enriched regions is that there are three types: sharp, broad and mixed (FIG. 2b). Sharp peaks are generally found for protein–DNA binding or histone modifications at regulatory elements, whereas broad regions are often associated with histone modifications that mark domains — for example, transcribed or repressed regions. Most current algorithms have been designed for sharp peaks, and adjacent peaks are coalesced *post hoc* for broad regions. However, many peak detection techniques used in ChIP–chip and DNA copy number analysis[73] will soon be modified for ChIP–seq, and new approaches are being developed[74,75]. A powerful method would integrate an approach for sharp peaks with an approach for broad peaks and be able to apply an appropriate technique for the features found without previous knowledge of the type of enrichment.

The performance of a peak caller can be tested by validating a large set of sites using quantitative PCR or by computing the distribution of distances from each peak to a nearby known protein-binding sequence motif. A careful comparison of the algorithms is still being carried out, but it is clear that the best methods should at least take advantage of the strand-specific pattern that is expected at a binding location, adjust for local variation as measured by input DNA and, to a lesser extent, correct for sequence alignability. The statistical significance of enriched sites is generally measured by the false discovery rate (FDR)[76,77], which is the expected proportion of incorrectly identified sites among those that are found to be significant. Determining significance for a multitude of features in the data results in a 'multiple hypothesis problem', in which features that seem to be significant arise owing to the large number of features being considered. The q value of a peak is the minimum FDR at which the peak is deemed significant and is analogous to the p value in a single hypothesis test setting. As in analysis of other genomic data types, it is important to note that the accuracy of the statistical significance computed in these algorithms depends on how realistic the underlying null distribution is. For ChIP–seq, an FDR derived from a null distribution based on randomization of ChIP reads can be off by an order of magnitude[48] because tags in the same or neighbouring positions are not completely independent even without true binding, as can be seen in the input control profile.

Downstream analysis. There are many approaches that can be taken to analyse the biological implications of ChIP–seq data. Owing to space restrictions I do not discuss these extensively, but some important aspects can be highlighted. For protein–DNA binding, the most common follow-up analysis is discovery of binding sequence motifs[78]. The sequences of the top-scoring sites can be entered into motif-finding algorithm programs such as MEME[79], MDScan[80], Weeder[81] and WebMOTIFS[82], and potential motifs are returned along with their statistical significance. In some cases, a single motif clearly stands out with much higher statistical significance than the subsequent matches and is largely insensitive to the number of the sites used to search. In other cases, there is a series of motifs with a gradual decrease in statistical significance, and further analysis of combinatorial occurrences of the motifs may be informative in identifying cooperative interactions among transcription factors or other more complex relationships among the motifs. The process of computing statistical significance is not straightforward, and the available algorithms use different null models and multiple-testing adjustment; therefore, it is important to functionally validate any motifs that are found. ChIP–chip has been used successfully on numerous occasions for motif discovery, but analyses have shown that the distances between the peaks of transcription factor binding and the nearby motifs are smaller for ChIP–seq, which indicates that ChIP–seq data are superior for this application[48,65]. For some factors, most of the ChIP–seq peaks are within 10–30 bp of the known motif[48]. After a motif is found, searching for the sequence in the genome generally reveals that there are many more sites with the motif than those identified by ChIP–seq. Why some occurrences of a motif are functional and others are not is at least partially related to the presence or absence of nucleosomes or a specific histone modification; this can be explored with nucleosome profiles that are obtained by sequencing[29,37].

Another basic analysis that can be performed using ChIP–seq data is to annotate the location of the peaks on the genome in relation to known genomic features, such as the transcriptional start site, exon–intron boundaries and the 3′ ends of genes. The transcriptional start sites of active genes, for instance, are known to be enriched with histone H3 trimethylated at lysine 4 (H3K4me3), and enhancers are enriched with histone H3 monomethylated at lysine 4 (H3K4me1)[25,83]. It is informative to view this type of data at a relative scale — for example, by rescaling all genes to have the same length so that the average profile over the gene body can be viewed — as well as absolute scale. To find relationships between the profiles, a correlation analysis can be performed, as well as more advanced clustering methods[84]. ChIP–chip and ChIP–seq data from the same experiments are generally similar but have subtle differences; therefore, combining both platforms requires careful attention, especially to the amount of smoothing applied to profiles. Incorporating other data types into the analysis is also necessary for biological interpretation. Classifying ChIP–seq patterns by their relationship to expression data, for example, is an important first step — if the expression level of a gene correlates with the binding status of a transcriptional activator, this might indicate that the gene is a target of that activator, or if a chromatin mark is enriched at the promoters of genes with high expression, it can be inferred to be related to transcriptional activation. For a group of genes with a common feature — for example, genes that bind the same transcription factor or have the same modifications — Gene Ontology analysis[85] can be performed to see whether a particular molecular function or biological process is over-represented in those genes[86]. More advanced analysis includes the discovery of novel elements based on ChIP–seq data. For example, the locations of H3K4me3 and histone H3 trimethylated at lysine 36 (H3K36me3), which are known to be found at promoters and across transcribed regions, respectively, can be used to identify large non-coding RNAs[87]. Combined with SNP information, ChIP–seq data can also be used to investigate allele-specific binding and modification[27].

Available software. Many of the algorithms for alignment and peak detection discussed earlier are accompanied by software. Some are available as a plug-in package for the statistical language R, a powerful system for data analysis that is popular among bioinformaticians[88]; others are based on standard compiled languages such as C or C++. In addition to the binding profile, most programs generate a list of enriched sites, which are viewed on a genome browser. One program with a menu-driven user interface is CisGenome[69], which features a ChIP–chip and ChIP–seq analysis pipeline with support for interactive analysis and visualization. More user-friendly software tools designed for biologists will be developed in the future, but it is unlikely that the tools available in a single software package will meet all analysis needs. This is particularly the case when the experimental design is more complicated or when advanced analysis that involves the integration of other data types is needed. Therefore, in most genomics projects it is imperative that a bioinformatics expert is a member of the research team.

Conclusion and future directions

ChIP has become a principal tool for understanding transcriptional cascades and deciphering the information encoded in chromatin and, owing to the recent remarkable progress in high-throughput sequencing platforms, ChIP–seq is poised to become the dominant profiling approach. The high cost of sequencing and the lack of easy access to platforms are still the limiting factors for most investigators, but the situation is expected to improve in the near future. ChIP–seq already offers higher resolution and cleaner data at a lower cost than the array-based alternatives for genome-wide profiling of large genomes. Improved spatial resolution has already resulted in substantial progress in several areas, most notably in genome-wide characterization of chromatin modifications at the nucleosome level and in accurate identification of the DNA sequence elements involved in transcriptional regulation. In the future, improved sequencing capabilities will allow the profiling of a large number of DNA-binding proteins, as well as a more complete set of chromatin marks in a myriad of epigenomes across multiple tissues, cell types, conditions and developmental stages. The ENCODE project[89], the modENCODE project[41] and the NIH Roadmap Epigenomics Program are a first step in large-scale profiling, and lessons from these projects will spur more detailed characterizations in specific systems. To extract the most information from ChIP–seq data, integrative analysis with other data types will be essential. For example, the integration of ChIP–seq data with RNA–seq data may result in the elucidation of gene regulatory networks and the characterization of the interplay between the transcriptome and the epigenome. Experimental challenges for the future include the careful validation of antibodies, the development of methods for working with a small number of cells and single-cell-level characterization. Even greater challenges for many laboratories are likely to be the effective management and analysis of the immense amount of sequencing data. This will require the development of user-friendly and robust software tools for data analysis and closer interaction between laboratory scientists and bioinformaticians.

1. Farnham, P. J. Insights from genomic profiling of transcription factors. *Nature Rev. Genet.* **10**, 605–616 (2009).
2. Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nature Rev. Genet*. **10**, 161–172 (2009).
3. Henikoff, S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nature Rev. Genet.* **9**, 15–26 (2008).
4. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128**, 707–719 (2007).
5. Allis, C. D., Jenuwein, T. & Reinberg, D. (eds) *Epigenetics* (Cold Spring Harb. Lab. Press, New York, 2007).
6. Berger, S. L. The complex language of chromatin regulation during transcription. *Nature* **447**, 407–412 (2007).
7. Bernstein, B. E., Meissner, A. & Lander, E. S. The mammalian epigenome. *Cell* **128**, 669–681 (2007).
8. Solomon, M. J., Larsen, P. L. & Varshavsky, A. Mapping protein–DNA interactions *in vivo* with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* **53**, 937–947 (1988).
9. Blat, Y. & Kleckner, N. Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell* **98**, 249–259 (1999).
10. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
11. Bentley, D. R. Whole-genome re-sequencing. *Curr. Opin. Genet. Dev*. **16**, 545–552 (2006).
12. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature Biotech*. **26**, 1135–1145 (2008).
13. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).
14. Hillier, L. W. *et al.* Whole-genome sequencing and variant discovery in *C. elegans. Nature Methods* **5**, 183–188 (2008).
15. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
16. Kim, J. B. *et al.* Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481–1484 (2007).
17. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
18. Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243 (2008).
19. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
20. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
21. Maher, C. A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97–101 (2009).
22. Lau, N. C. *et al.* Characterization of the piRNA complex from rat testes. *Science* **313**, 363–367 (2006).
23. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nature Biotech*. **26**, 1146–1153 (2008).
24. Johnson, D. S. *et al.* Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316**, 1497–1502 (2007).

This study is an early demonstration of the increased sensitivity and specificity of ChIP–seq for genome-wide mapping of transcription factor binding sites.

25. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
The first large-scale profiling of chromatin marks using ChIP–seq. Histone H2A.Z, RNA polymerase II, CTCF and 20 histone methylations were profiled for human T cells.
26. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**, 651–657 (2007).
Another early demonstration of the increased sensitivity and specificity of ChIP–seq.
27. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
The first study to examine in a genome-wide manner how chromatin states change as cells move from immature to adult states.
28. Visel, A. *et al.* ChIP–seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
29. Robertson, A. G. *et al.* Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res*. **18**, 1906–1917 (2008).
30. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).
31. Tolstorukov, M. Y., Kharchenko, P. V., Goldman, J. A., Kingston, R. E. & Park, P. J. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. *Genome Res*. **19**, 967–977 (2009).
32. Henikoff, S., Henikoff, J. G., Sakai, A., Loeb, G. B. & Ahmad, K. Genome-wide profiling of salt fractions maps physical properties of chromatin. *Genome Res*. **19**, 460–469 (2009).
33. Orlando, V. Mapping chromosomal proteins *in vivo* by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem. Sci*. **25**, 99–104 (2000).
34. O'Neill, L. P. & Turner, B. M. Immunoprecipitation of native chromatin: NChIP. *Methods* **31**, 76–82 (2003).
35. Schones, D. E. & Zhao, K. Genome-wide approaches to studying chromatin modifications. *Nature Rev. Genet*. **9**, 179–191 (2008).
36. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).
37. Alekseyenko, A. A. *et al.* A sequence motif within chromatin entry sites directs MSL establishment on the *Drosophila* X chromosome. *Cell* **134**, 599–609 (2008).
38. Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP–seq experiments relative to controls. *Nature Biotech*. **27**, 66–75 (2009).
This paper proposes a peak-scoring approach that emphasizes the need for input control and sequence alignability.
39. Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nature Methods* **5**, 1005–1010 (2008).
40. Whiteford, N. *et al.* An analysis of the feasibility of short read sequencing. *Nucleic Acids Res*. **33**, e171 (2005).
41. Celniker, S. E. *et al.* Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
42. Acevedo, L. G. *et al.* Genome-scale ChIP–chip analysis using 10,000 human cells. *Biotechniques* **43**, 791–797 (2007).
43. Dahl, J. A. & Collas, P. MicroChIP — a rapid micro chromatin immunoprecipitation assay for small cell samples and biopsies. *Nucleic Acids Res*. **36**, e15 (2008).
44. Wu, A. R. *et al.* Automated microfluidic chromatin immunoprecipitation from 2,000 cells. *Lab Chip* **9**, 1365–1370 (2009).
45. O'Neill, L. P. *et al.* Epigenetic characterization of the early embryo with a chromatin immunoprecipitation protocol applicable to small cell populations. *Nature Genet*. **38**, 835–841 (2006).
46. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106–109 (2008).
47. Peng, S., Alekseyenko, A. A., Larschan, E., Kuroda, M. I. & Park, P. J. Normalization and experimental design for ChIP–chip data. *BMC Bioinformatics* **8**, 219 (2007).
48. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP–seq experiments for DNA-binding proteins. *Nature Biotech*. **26**, 1351–1359 (2008).
This study develops peak callers based on strand-specific patterns and examines the issue of sequencing depth.
49. Lefrançois, P. *et al.* Efficient yeast ChIP–seq using multiplex short-read DNA sequencing. *BMC Genomics* **10**, 37 (2009).
50. Fullwood, M. J., Wei, C. L., Liu, E. T. & Ruan, Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res*. **19**, 521–532 (2009).
51. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA–seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. **18**, 1509–1517 (2008).
52. Barrett, T. *et al.* NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*. **37**, D885–D890 (2009).
53. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. **37**, D5–D15 (2009).
54. Cochrane, G. *et al.* Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res*. **37**, D19–D25 (2009).
55. Erlich, Y., Mitra, P. P., delaBastide, M., McCombie, W. R. & Hannon, G. J. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nature Methods* **5**, 679–682 (2008).
56. Rougemont, J. *et al.* Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics* **9**, 431 (2008).
57. Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech*. **27**, 455–457 (2009).
58. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. **18**, 1851–1858 (2008).
This study introduces a popular short-read aligner for NGS platforms.
59. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. **10**, R25 (2009).
60. Ondov, B. D., Varadarajan, A., Passalacqua, K. D. & Bergman, N. H. Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. *Bioinformatics* **24**, 2776–2777 (2008).
61. Rumble, S. M. *et al.* SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol*. **5**, e1000386 (2009).
62. Bourque, G. *et al.* Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res*. **18**, 1752–1762 (2008).
63. Pauler, F. M. *et al.* H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res*. **19**, 221–233 (2009).
64. Zheng, D. Asymmetric histone modifications between the original and derived loci of human segmental duplications. *Genome Biol*. **9**, R105 (2008).
65. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP–seq data. *Nature Methods* **5**, 829–834 (2008).
This paper proposes a peak-calling method that accounts for the directionality of reads and the size of sequenced fragments.
66. Fejes, A. P. *et al.* FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24**, 1729–1730 (2008).
67. Zhang, Y. *et al.* Model-based analysis of ChIP–seq (MACS). *Genome Biol*. **9**, R137 (2008).
68. Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of *in vivo* protein–DNA binding sites from ChIP–seq data. *Nucleic Acids Res*. **36**, 5221–5231 (2008).
69. Ji, H. *et al.* An integrated software system for analyzing ChIP–chip and ChIP–seq data. *Nature Biotech*. **26**, 1293–1300 (2008).
This article introduces a software system that has a graphical user interface for data analysis and includes tools for data visualization and motif discovery.
70. Nix, D. A., Courdy, S. J. & Boucher, K. M. Empirical methods for controlling false positives and estimating confidence in ChIP–seq peaks. *BMC Bioinformatics* **9**, 523 (2008).
71. Schmid, C. D. & Bucher, P. ChIP–seq data reveal nucleosome architecture of human promoters. *Cell* **131**, 831–832 (2007); author reply **131**, 832–833 (2007).
72. Boyle, A. P., Guinney, J., Crawford, G. E. & Furey, T. S. F–seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537–2538 (2008).
73. Lai, W. R., Johnson, M. D., Kucherlapati, R. & Park, P. J. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763–3770 (2005).
74. Xu, H., Wei, C. L., Lin, F. & Sung, W. K. An HMM approach to genome-wide identification of differential histone modification sites from ChIP–seq data. *Bioinformatics* **24**, 2344–2349 (2008).
75. Zang, C. *et al.* A clustering approach for identification of enriched domains from histone modification ChIP–seq data. *Bioinformatics* **25**, 1952–1958 (2009).
76. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
77. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
78. Tompa, M. *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotech*. **23**, 137–144 (2005).
79. Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*. **34**, W369–W373 (2006).
80. Liu, X. S., Brutlag, D. L. & Liu, J. S. An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotech*. **20**, 835–839 (2002).
81. Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res*. **32**, W199–W203 (2004).
82. Romer, K. A. *et al.* WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches. *Nucleic Acids Res.* **35**, W217–W220 (2007).
83. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet*. **39**, 311–318 (2007).
84. Hon, G., Ren, B. & Wang, W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput. Biol*. **4**, e1000201 (2008).
85. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genet*. **25**, 25–29 (2000).
86. Orford, K. *et al.* Differential H3K4 methylation identifies developmentally poised hematopoietic genes. *Dev. Cell* **14**, 798–809 (2008).
87. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
88. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. **5**, R80 (2004).
89. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
90. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genet*. **40**, 897–903 (2008).
This paper examines the correlations among 39 histone modification patterns and their relationship to transcriptional activation.
91. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
92. Kurdistani, S. K., Tavazoie, S. & Grunstein, M. Mapping global histone acetylation patterns to gene expression. *Cell* **117**, 721–733 (2004).
93. Liu, C. L. *et al.* Single-nucleosome mapping of histone modifications in *S. cerevisiae. PLoS Biol*. **3**, e328 (2005).
94. Pokholok, D. K. *et al.* Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**, 517–527 (2005).

95. Lee, C. K., Shibata, Y., Rao, B., Strahl, B. D. & Lieb, J. D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genet*. **36**, 900–905 (2004).
96. Yuan, G. C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae. Science* **309**, 626–630 (2005).
97. Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genet*. **39**, 1235–1244 (2007).
98. Johnson, S. M., Tan, F. J., McCullough, H. L., Riordan, D. P. & Fire, A. Z. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res*. **16**, 1505–1516 (2006).
99. Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446**, 572–576 (2007).
100. Mavrich, T. N. *et al.* Nucleosome organization in the *Drosophila* genome. *Nature* **453**, 358–362 (2008).

Acknowledgements
I thank P. Kharchenko, M. Tolstorukov, A. Alekseyenko and other members of the Park and the Kuroda laboratories for their insights. I gratefully acknowledge support from the National Institutes of Health grants R01GM082798, U01HG004258 and RL1DE019021.

FURTHER INFORMATION
Peter J. Park's homepage: http://compbio.med.harvard.edu
A community forum for discussion of issues related to NGS: http://seqanswers.com
The European Molecular Biology Laboratory Nucleotide Sequence Database: http://www.ebi.ac.uk/embl
The Short Read Archive: http://www.ncbi.nlm.nih.gov/Traces/sra

ALL LINKS ARE ACTIVE IN THE ONLINE PDF