STUDY DESIGNS

Identifying and mitigating bias in next-generation sequencing methods for chromatin biology

Clifford A. Meyer and X. Shirley Liu

Abstract | Next-generation sequencing (NGS) technologies have been used in diverse ways to investigate various aspects of chromatin biology by identifying genomic loci that are bound by transcription factors, occupied by nucleosomes or accessible to nuclease cleavage, or loci that physically interact with remote genomic loci. However, reaching sound biological conclusions from such NGS enrichment profiles requires many potential biases to be taken into account. In this Review, we discuss common ways in which biases may be introduced into NGS chromatin profiling data, approaches to diagnose these biases and analytical techniques to mitigate their effect.

ChIP–seq
(Chromatin immunoprecipitation followed by next-generation DNA sequencing). A method to identify DNA-associated protein-binding sites.

MNase-seq
A method in which micrococcal nuclease (MNase) digestion of chromatin is followed by next-generation sequencing to identify loci of high nucleosome occupancy.

Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, Massachusetts 02115, USA; and Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA.
e-mails: cliff@research.dfci.harvard.edu; xsliu@jimmy.harvard.edu
doi:10.1038/nrg3788
Published online
16 September 2014

Technologies such as ChIP–seq[1–4], MNase-seq[1,5,6], FAIRE–seq, DNase-seq[7–9], Hi-C[10,11], ChIA-PET[12] and ATAC-seq[13] combine next-generation sequencing (NGS) with new biochemical techniques or modifications of established methods to enable genome-wide investigations of a broad range of chromatin phenomena (FIG. 1). Inevitably, the understanding of data produced by these techniques lags behind their development, and sometimes phenomena observed through newly minted techniques are later understood to result from biases. In the initial excitement over NGS technologies themselves, there was a common misconception that the digital readout of read counts could give unbiased results. However, it is now clear from data that have been produced from increasingly sophisticated NGS experiments that substantial biases are indeed common.

In this Review, we summarize the most important lessons learned about the systematic artefacts that have been observed in NGS chromatin profiling experiments and describe the analytical strategies that have been developed to handle such artefacts. Although RNA also has an important role in chromatin structure and function, we have limited the scope of this Review to DNA-centric assays. These considerations are of interest to experimental and computational biologists alike, and are also central to experimental design, protocol selection and data analyses. We first describe common sources of bias that arise in NGS chromatin profiling experiments and continue with a discussion on experimental design considerations, including the use of controls, the need for replicates and methods to mitigate batch effects. Finally, we discuss the emerging methods that have been developed for various analytical tasks and outline how they can be used to handle biases in genome-wide investigations.

Sources of bias

Genomic approaches for chromatin biology are under continual development — protocols are frequently refined, and new questions are constantly being posed. In some cases, applying appropriate software that accounts for bias effects is sufficient to obtain sound results. However, further experiments, controls and analyses are often needed to account for technical artefacts. Below, we describe the main sources of bias, including chromatin structure, enzymatic cleavage, nucleic acid isolation, PCR amplification and read mapping effects.

Chromatin fragmentation and size selection: sonication. Chromatin structure itself is a major source of bias in chromatin profiling experiments. In ChIP–seq in which the aim is to quantify the protein–DNA interactions of a specific protein, DNA fragmentation (usually by sonication) is required before protein-bound fragments are isolated by immunoprecipitation[14]. The mechanical characteristics of chromatin vary across the genome, which creates fluctuations in DNA fragility. Heterochromatin, which is not generally associated with transcription factor (TF) binding, tends to be more resistant to shearing than euchromatin[15]. Moreover, the way in which sonication is carried out can result in different fragment size distributions and consequently

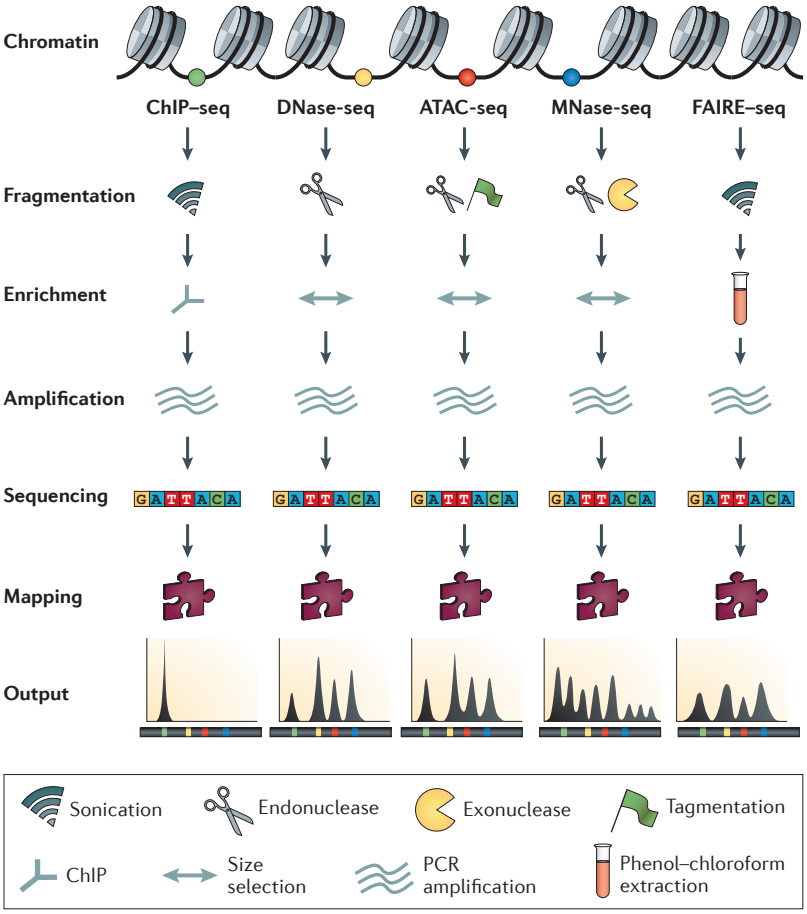

Figure 1 | **An overview of ChIP–seq, DNase-seq, ATAC-seq, MNase-seq and FAIRE–seq experiments.** A genomic locus analysed by complementary chromatin profiling experiments reveals different aspects of chromatin structure: ChIP–seq reveals binding sites of specific transcription factors (TFs); DNase-seq, ATAC-seq and FAIRE–seq reveal regions of open chromatin; and MNase-seq identifies well-positioned nucleosomes. In ChIP–seq, specific antibodies are used to extract DNA fragments that are bound to the target protein, either directly or through other proteins in a complex that contains the target factor. In DNase-seq, chromatin is lightly digested by the DNase I endonuclease. Size selection is used to enrich for fragments that are produced in regions of chromatin where the DNA is highly sensitive to DNase I attack. ATAC-seq is an alternative method to DNase-seq that uses an engineered Tn5 transposase to cleave DNA and to integrate primer DNA sequences into the cleaved genomic DNA (that is, tagmentation). Micrococcal nuclease (MNase) is an endo–exonuclease that processively digests DNA until an obstruction, such as a nucleosome, is reached. In FAIRE–seq, formaldehyde is used to crosslink chromatin, and phenol–chloroform is used to isolate sheared DNA.

FAIRE–seq
(Formaldehyde-assisted isolation of regulatory elements followed by sequencing). A method to determine regulatory regions of the genome.

DNase-seq
A method in which DNase I digestion of chromatin is combined with next-generation sequencing to identify regulatory regions of the genome, including enhancers and promoters.

sample-specific biases that are induced by chromatin configuration. As a result, it is not recommended to use a single input sample as a control for ChIP–seq peak calling if it is not sonicated together with the ChIP sample. Input samples from many different batches of ChIP–seq experiments that are produced from the same cell line under consistent conditions and using the same protocol may be combined as a control.

Chromatin fragmentation and size selection: enzymatic cleavage. Enzymatic cleavage approaches are also strongly influenced by chromatin structure, although the detailed nature of the effect varies between enzymes. For example, nucleosome-associated DNA is particularly insensitive to digestion by micrococcal nuclease (MNase), and this enzyme is thus particularly useful for nucleosome occupancy characterization in MNase-seq. MNase induces single-strand breaks and subsequently double stranded ones by cleaving the complementary strand in close proximity to the first break[16]. MNase continues to digest the exposed DNA ends until it reaches an obstruction, such as a nucleosome, a stably bound TF[17] or a refractory DNA sequence[18]. In MNase-seq studies, fragments of approximately one nucleosome length (~147 bp) are typically selected for sequencing[6]. Different size ranges of MNase-digested fragments have been shown to reveal different patterns of enrichment[19]. Therefore, MNase-seq data ought to be interpreted relative to fragment length distribution. Studies have found that nucleosomes occupy regions that are more GC rich than their neighbouring regions[20–22] and that they are intrinsically depleted at transcription terminator regions[23]. However, bias in MNase digestion towards AT-rich sequences[23,24] suggests that MNase cleavage bias might be at least partially responsible for this effect. As a further complication, the degree to which DNA sequence influences MNase cleavage is affected by the cleavage reaction temperature[18].

Similarly to MNase, the nuclease DNase I generates double-strand breaks by nicking complementary strands of DNA one strand at a time[25]. However, unlike MNase, DNase I has not been reported to have substantial exonuclease activity, and it operates in a 'hit-and-run' mode rather than 'nibbles' at the ends of DNA until an obstruction is reached. The efficiency of DNase-seq in identifying TF binding sites is highly dependent on fragment size and, for several TFs, it is more efficient to use shorter fragments (<100 bp) than longer ones. By contrast, longer fragments (>150 bp) tend to span entire nucleosomes[26,27] and are less likely to cluster around open chromatin regions (FIG. 2).

Sites of DNase I cleavage are strongly affected by the precise sequence of the three nucleotides on either side of the cleavage site, and this bias is strand specific[28]. Intrinsic DNase I cleavage bias is particularly evident when analysing a set of sites in aggregate, in which the genomic loci are aligned by the TF motif on DNase I-hypersensitive sites. This issue is not limited to DNase I; other nucleases, including MNase[22,24], cyanase and benzonase[29], also cleave DNA in a sequence-sensitive way. The Tn5 transposase used in ATAC-seq[13] is also known to cleave DNA in a sequence-dependent manner.

Nucleic acid isolation. Whole-genome sequencing, which should be free of chromatin effects, sometimes produces tissue-specific patterns of high- and low-coverage across the genome. This phenomenon occurs as a result of the phenol–chloroform extraction step that is commonly used to separate nucleic acids from proteins[30]. Differential solubility is the principle of this separation step: nucleic acids are more soluble in the aqueous chloroform phase, whereas proteins tend to be more soluble in the organic phenol phase. Prior to phenol–chloroform extraction, protein is digested using

Hi-C
An extension of chromosome conformation capture that uses next-generation sequencing to observe long-range interaction frequencies between different regions of the genome.

ChIA-PET
(Chromatin interaction analysis by paired-end tag sequencing). A method that combines chromatin immunoprecipitation-based enrichment and chromatin proximity ligation with paired-end next-generation sequencing to determine genome-wide chromatin interactions.

the proteinase K enzyme. However, incomplete digestion can result in DNA-binding proteins carrying a fraction of DNA into the phenol phase, which leads to uneven genome coverage owing to chromatin effects[30]. A similar differential solubility phenomenon has been used in FAIRE–seq[31] as an alternative method to DNase-seq to determine regions of open chromatin.

PCR amplification biases and duplications. Multiple instances of the same sequence read in an NGS data set can originate from mistaking one feature for two in sequencing image analyses, from sequencing PCR amplicons derived from the same original fragment or from the presence of multiple fragments in the original sample. This issue is particularly troublesome with small amounts of starting material[32].

PCR amplification biases arise because DNA sequence content and length determine the kinetics of annealing and denaturing in each cycle of this procedure. The combination of temperature profile, polymerase and buffer used during PCR can therefore lead to differential efficiencies in amplification between different sequences[33], which could be exacerbated with increasing PCR cycles. This is often manifested as a bias towards GC-rich fragments, although not necessarily in regions with extremely high GC levels[34]. Although the sequence read is the end product of sequencing, the fragment of DNA amplified in PCR, which is usually longer than the read itself, is the relevant entity in the analysis of PCR amplification effects[34]. We recommend limited use of PCR amplification because bias increases with every PCR cycle.

Read mapping. The short sequence reads that are produced by NGS experiments are typically mapped onto a reference genome before subsequent analysis steps are carried out. Repetitive elements, duplications of genomic sequences[35] (including paralogous genes) and differences between the sequenced genome and the reference genome can all introduce coverage bias between different regions of the genome. Efficient mapping algorithms that take advantage of the short read length to

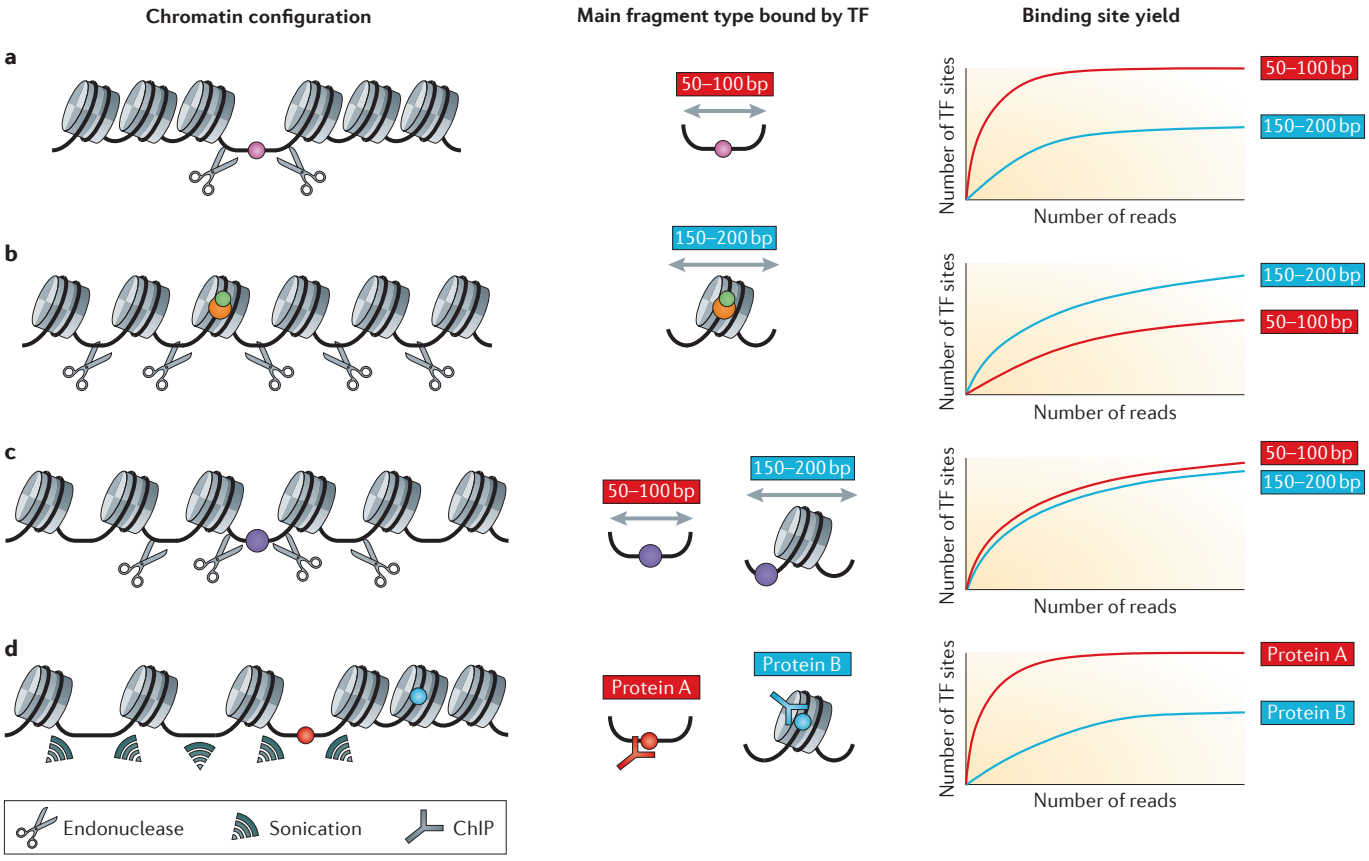

Figure 2 | **Fragmentation effects in DNase-seq and ChIP–seq.** Chromatin structure and fragmentation interact to produce biased patterns of enrichment across the genome. **a** | Some transcription factors (TFs), such as CCCTC-binding factor (CTCF), typically bind in short nucleosome-depleted regions that are flanked by arrays of nucleosomes. When carrying out DNase-seq, shorter fragments are much more efficient than longer ones for identifying such sites. **b** | Histones and other factors that associate with DNA in nucleosomes rather than linker regions may also be located in DNase I-hypersensitive regions. Longer fragments may be more efficient for detecting the binding of such factors. **c** | Some factors bind in linker regions that are flanked by loosely packed and unorganized nucleosomes. Such regions can be enriched in both long and short fragments in DNase-seq. **d** | In ChIP–seq, chromatin is typically fragmented by sonication. Similar to DNase digestion, sonication is more efficient in regions of open chromatin. Factors bound in open chromatin contexts are more likely to be identified by ChIP–seq.

align NGS reads with the reference genome — including MAQ[36], Burrows–Wheeler Alignment (BWA)[37], Bowtie[38], mrFAST[39] and SOAP2 (REF. 40) — introduce algorithm-specific biases when finding imperfect or ambiguous matches to the genome. As a result, there are algorithm-specific 'unmappable' regions of the genome to which no reads can be aligned. These regions may be approximated by systematically attempting to map every possible read in the reference genome back to the entire reference genome[41].

The proportion of a genome to which a sequence read may be uniquely assigned depends on both the length of the sequence reads and the accuracy of the sequencing. Longer reads and paired-end reads with known insert sizes allow read mapping with greater coverage and greater uniformity of coverage[41]. Regions to which reads cannot be mapped have often been considered as less likely to be functional, and they are often repetitive elements associated with transposon activity. Although most investigators ignore such regions, analyses of repeats using specialized methods[42,43] have revealed significant associations between chromatin marks[44] and TFs[45] with particular repeat families.

Incompleteness and inaccuracies in the genome assembly can result in regions of low and high coverage that cannot be explained by an analysis of mappability. For example, a region that is unique in the assembled reference genome may have multiple copies in the genome of the experimental sample. This occurs occasionally in studies of non-cancerous human samples and, to a greater extent, in more recently assembled genomes that are of lower quality than the human reference genome. In the human genome, such artefact-derived 'sticky' regions are frequently observed as ChIP–seq and DNase-seq peaks[46], sometimes as the 'strongest' peaks, and such regions are often close to centromeres and telomeres. We expect that recently updated genome assemblies, such as HG38 and MM10, will mitigate some mappability issues.

Genomic variation — including single-nucleotide polymorphisms (SNPs), insertions and deletions (indels), and rearrangements — may produce sequence reads that cannot be mapped to the reference genome. In cancer cell lines, genomic loci with high copy numbers are more likely to be determined as enriched in ChIP–seq and other chromatin assays[47,48]. When mapping allele-specific reads to a reference genome, there is a greater likelihood of aligning a short SNP-containing read if the SNP variant is consistent with the reference genome. This situation is exacerbated when the read contains sequencing errors[49]. Simply masking known SNP positions in the genome can lead to other artefacts owing to a combination of factors, including the presence of multiple SNPs in close proximity, unknown SNPs and similar sequences in other regions of the genome[50].

TF binding characteristics. The characteristics of TF binding to DNA differ substantially between TFs[51]. The observed signals can be influenced by nucleosome positioning relative to the TF binding site, strength of binding, binding kinetics and the tendency of a TF to bind in conjunction with other factors or potentially through the recognition of histone post-translational modifications. Some TFs are therefore more readily detected by TF binding inference techniques based on ATAC-seq, DNase-seq or MNase-seq.

Classic DNase I footprinting studies have shown that TF binding often modulates the pattern of DNase I cleavage at the site of protein–DNA interaction and at the flanking nucleotides, usually in a way such that DNase I cleavage is impeded at central positions where the DNA–protein interaction occurs and facilitated at the flanking positions. Close examination of DNase-seq read positions within regions of DNase I hypersensitivity reveals highly non-homogeneous patterns. Factors that contribute to these complex patterns include nucleosome occupancy, DNA sequence-dependent cleavage and other biases, as well as the effect of TF binding itself[26].

Experimental design considerations

To maximize discovery using limited research budgets, investigators tend to carry out minimal controls and replicates in NGS experiments. Nevertheless, controls are required to accurately evaluate the effects of biases, and replicates are needed to make an assessment of data variability. In experiments that involve comparison of multiple samples, bias effects often produce observable differences between sample batches. Success in correcting for such batch effects is dependent on good experimental design. In particular, it is suggested that biologically distinct treatment groups need to be distributed evenly over processing batches so that experimental effects and batch effects can be distinguished. In addition, in order to obtain meaningful results from differential analyses between conditions, the experimental protocol needs to be carried out in a highly consistent manner for all samples (FIG. 3). Below, we detail some of the considerations that should be taken into account when designing NGS chromatin profiling experiments to obtain the most meaningful results (TABLE 1).

Sequencing depth and read length. Several sequencing options are available, including selection of read length, single-end or paired-end reads and the expected number of reads. In single-end sequencing, duplicates that arise from PCR amplification can often be confused with multiple fragments that have one end in common in the original sample. Paired-end sequencing can help to distinguish between these, as the probability of sampling two fragments with the exact same start and end is much lower than the probability of identifying a single common end. Some commercial library construction kits, such as the Rubicon ThruPLEX-FD Prep Kit, are more efficient in making sequencing libraries with less duplication bias from very little starting material. Random barcoding is another technique that can be used to distinguish PCR duplicates from duplicates in the unamplified DNA[52].

The number of informative reads produced from an NGS experiment depends on sample quality, sequencing technology and protocol, among other factors. As a result, NGS data sets can differ substantially in read count, as well as in the observed number and distribution

ATAC-seq
(Assay for transposase-accessible chromatin using sequencing). A method that combines next-generation sequencing with *in vitro* transposition of sequencing adapters into native chromatin.

Random barcoding
A technique that ligates a diverse assortment of short random DNA sequences to an unamplified DNA sample, which can be used to distinguish duplicates produced by PCR from those originating from the unamplified DNA.

Spike-in
Controls that are known quantities of readily identifiable nucleic acids, which are added to a sample prior to critical steps in an experimental protocol. Such controls may be used for bias assessment and calibration purposes.

of different DNA species, which reflects library complexity. Deep sequencing of low-complexity libraries produces repeated observations of some DNA species, which yields less information than high-complexity libraries, and methods to characterize library complexity are therefore useful diagnostic tools for NGS analyses[53]. In addition, the Encyclopedia of DNA Elements (ENCODE) consortium[54] PCR bottleneck coefficient (PBC) metric — the ratio of genomic locations with a single uniquely mapped read over the total number of genomic locations with uniquely mapped reads — is an informative measure of library complexity if evaluated at similar sequencing depths.

Controls to detect and correct biases for ChIP–seq. In ChIP–seq it is common to use a chromatin 'input' control, in which sonicated chromatin is assayed without enrichment of specific binding sites through immunoprecipitation. A recurrent issue in the selection and interpretation of controls for bias correction in NGS applications is the occurrence of biological signal in the controls themselves. In input controls, weak TF binding signals may be observed because regions of TF binding also tend to be regions where chromatin is more amenable to fragmentation[15]. Owing to cost considerations, input controls are often sequenced to lower depths than the ChIP samples. However, this is not recommended, as the broader genomic distribution of signal in chromatin input DNA requires this input to be sequenced to a higher coverage than ChIP–seq for accurate results[55,56].

Another issue is the potential difference in bias between the samples of interest and the controls. Although information on mappability can be provided by ChIP–seq input controls, copy-number effects, broad chromatin accessibility and other sources of bias have been found to vary substantially between control and ChIP samples[48,56]. To minimize these sources of technical variation, it is advised to use input controls that are processed together with ChIP samples to correct for background bias.

Addition of a 'spike-in' reference chromatin sample to the study sample before immunoprecipitation provides a reference for quality control and bias characterization, and could enable the identification of global yet uniform TF binding changes. To discriminate spike-in sample reads from those derived from the study sample itself, the spike-in must originate from a different genome. In ChIP–seq, the foreign chromatin material needs to be bound by a homologous protein that is targeted by the antibody as efficiently as the protein in the study sample. The principle of this approach has been shown by spiking chromatin of HeLa cells into mouse samples for ChIP

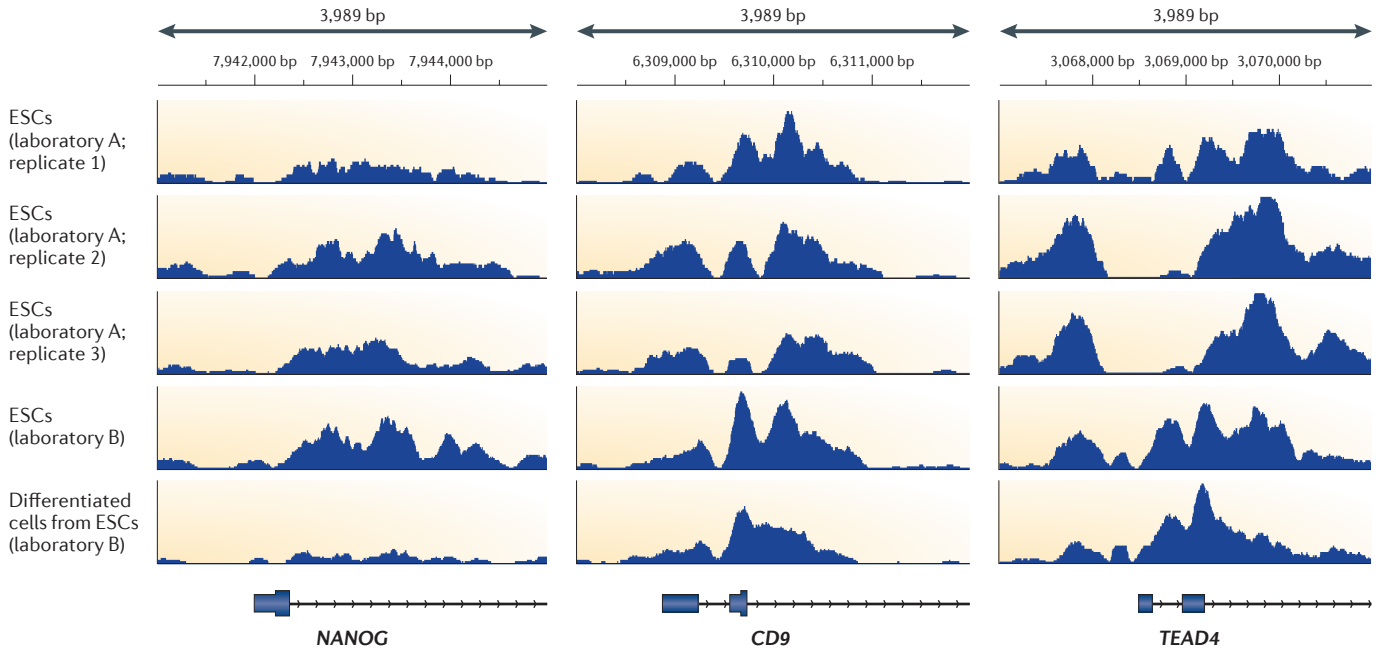

Figure 3 | **Variability of H3K4me3 ChIP–seq in human embryonic stem cells and differentiated cell lines.** Several factors — including fragmentation, immunoprecipitation conditions and PCR biases — can lead to different patterns of histone H3 lysine 4 trimethylation (H3K4me3) enrichment at gene promoters in the same cell line. Coarse characteristics of H3K4me3 enrichment, such as the depletion of H3K4me3 immediately upstream of the transcription start sites of a core set of genes, are consistent between samples. Closer inspection reveals clear qualitative and quantitative differences between samples. For example, some samples show sharper peaks, perhaps owing to differences in micrococcal nuclease (MNase) digestion conditions and fragment selection. Regions that seem to be different between embryonic stem cells (ESCs) and differentiated cells in ChIP–seq samples produced by laboratory B also show variability in ESC ChIP–seq replicates produced by laboratory A. These differences cannot be eliminated simply by scaling read counts to account for differences in read depth, as the effects are not uniform across all genes. Quantitative comparisons of ChIP–seq signal are problematic unless biological replicates are done and protocols are carried out in a highly consistent manner to produce data with comparable characteristics. Modelling biases can help to reduce the amount of unexplained variability and increase sensitivity in detecting true differences between sample groups. *NANOG*, nanog homeobox; *TEAD4*, TEA domain family member 4.

that targets subunits of RNA polymerases II and III[57]. This control may be especially useful in ChIP–seq studies of histone modifications. Although we believe that this type of control may be useful, it has not been extensively tested, and balancing the amount of spike-in relative to the chromatin of interest might still be challenging.

In a ChIP–seq experiment using an untested antibody, it is crucial to carry out various control experiments to establish the specificity of the antibody in genome-wide experiments[14,58]. Such experiments include the use of different antibodies and the knockdown or knockout of the target protein. Antibody

Table 1 | **Considerations in designing next-generation sequencing chromatin profiling experiments**

Factor	Common options	Considerations
Chromatin profiling assay	• ChIP–seq and antibody enrichment • DNase-seq • ATAC-seq • MNase-seq • MNase-ChIP–seq and antibody enrichment	• ChIP–seq requires good and specific antibodies[14,58] • Differences in data quality of ChIP–seq using different antibodies prevent all but the roughest comparisons between data sets • DNase-seq requires careful calibration of digestion conditions and fragment selection[26] • DNase-seq or ChIP–seq samples obtained using the same antibody may be compared, provided that protocols are followed consistently, and that bias effects and variability are taken into account[78,109] • ATAC-seq requires fewer cells and less experimental calibration[13], but bias characteristics are not as well understood as those of DNase-seq[26] • MNase-ChIP–seq using antibodies specific to enhancer features (such as H3K4me and H3K4me2) or promoter features (such as H3K4me3) can be more efficient than global MNase-seq for identifying nucleosome occupancy at regulatory regions of the genome[6]
Sequence length	• Read length of 25–150 bp • Single-end reads • Paired-end reads	• Read length is less important for chromatin profiling assays than studies of genomic or RNA transcript assembly[110] • Longer reads are suggested for studies that seek to identify allele-specific chromatin events[49] • In highly specialized studies of chromatin (for example, investigations of transposable elements[44]), longer reads and paired-end reads would be useful in improving mappability[43,55] • Paired-end reads have three advantages over single-end reads: they increase the mappable proportion of the genome, allow PCR duplicates to be more easily identified and enable the precise ends of fragments to be identified[26,55] • Sequencing costs of generating longer reads and paired-end sequencing need to be balanced against the value of more informative reads
Read depth	• Multiplexing • Number of lanes • Sequencing machine	• Multiplexing allows several samples to be sequenced in a single lane to a lower read depth[110] • Sequencing multiple biological replicates or sample replicates to a lower sequencing depth is preferable to sequencing a single sample to a greater depth • Information per read decreases as a function of read number: ChIP–seq targeting TFs that bind with high specificity reaches saturation at lower read depths than more broadly bound histone modifications[111]; DNase-seq also requires greater sequencing depths, and fragments longer than 147 bp saturate at much higher levels than shorter reads[26] • Even at low sequence depth, chromatin profiling should be informative for regions with strong signals, and pilot studies at low coverage are recommended before sequencing to higher coverage • It is important to examine library complexity in sequenced libraries, as low-complexity data sets that are sequenced to greater depths can be less informative than high-complexity ones sequenced to lower depths[14,53] • It is better to sequence a high-quality sample at low depth than a low-quality sample to high depth • Sample quality control can be carried out rapidly on MiSeq
Replicates	• Biological replicates • Technical replicates starting from the same biological material • Sequencing replicates	• Many technical bias effects accumulate before library preparation and sequencing; therefore, sequencing the same library multiple times is generally not informative • Biological replicates are essential to characterize variability between samples • Technical replicates starting from the same biological material can help to understand the degree to which technical biases contribute to variability • When processing samples, it is important to avoid processing replicates of the same treatment condition in the same batch, as this would result in batch effects confounding treatment effects[101,102]
ChIP–seq controls	• Input control • IgG control • Condition controls • 'Spike-in' controls	• Input controls are suggested in ChIP–seq experiments to distinguish real peak regions from artefacts; they ought to be sequenced to greater depths than immunoprecipitated samples to obtain adequate coverage[55] • Input controls are preferred to IgG, as they produce more complex libraries[14] • Conditions under which a TF is not induced may be used as a control for ChIP–seq in the induced condition; however, induction can lead to chromatin state changes in places where the TF binds and also elsewhere[90] • Spike-in controls have rarely been used in ChIP experiments, and their value is thus not well tested; naked DNA spike-ins would not capture chromatin effects, so for human study samples standardized chromatin spike-ins derived from yeast, fly or mouse may be useful[57]
DNase-seq, ATAC-seq and MNase controls	• Naked DNA • Condition controls	• In DNase-seq or ATAC-seq footprinting studies and MNase nucleosome positioning studies, naked DNA controls are useful for characterizing the DNA sequence bias of enzymatically induced cleavage[26,28,112] • To be informative, such experiments need to be done at high levels of coverage • Although analyses of DNase-seq in chromatin are already highly informative for predicting bias effects[26], naked DNA data could provide additional information about sequence bias effects that are not considered in current models

H3K4me, histone H3 lysine 4 methylation; IgG, immunoglobulin G; TF, transcription factor.

effects, such as epitope masking, can result in antibody-specific biases for the same TF[58].

Controls for enzymatic cleavage assays. Genomic assays that are based on the selection of fragments produced from enzymatic DNA cleavage — including ATAC-seq, DNase-seq and MNase-seq — may be influenced by the tendency of the enzyme to cleave some DNA sequences more efficiently than others. Controlling for such effects is particularly important when considering features at nucleotide resolution. DNase I cleavage bias due to DNA sequence at either end of the cleavage site can be estimated from DNase I digestion of naked genomic DNA, but systematic sequence features of the chromatin sample itself may also be used, as they can capture the sample-specific aspects of this type of bias. It has been shown through yeast naked DNA controls that MNase has cleavage biases that may be mistaken as nucleosome positioning signals[24].

Analytical techniques for bias correction

Below, we discuss issues that are generally applicable in NGS chromatin profiling analyses and methods that are implemented as software for specific analytical tasks. The general issues include identifying biases that are most likely to confound results, characterizing bias, adjusting for sequencing depth, handling duplicate reads and modelling variations in NGS data. Specific analyses include peak detection, DNase-seq footprint and chromatin landscape analyses, domain calling, ChIP–seq peak deconvolution and differential enrichment analysis. TABLE 2 summarizes artefacts that might affect various analysis types, as well as ways of diagnosing and correcting these effects.

Length scales of biases and biological features. Genomic analyses are carried out over length scales from 1 bp (in SNP analyses) to ~10 bp (in DNase I footprint analyses), ~100 bp (in TF ChIP–seq peak calling) and ~100 bp–100 kb (in chromatin domain analyses). Bias effects also occur on different length scales; for example, read errors occur on the single-nucleotide scale, whereas PCR amplification biases affect fragments of ~100 bp. The biases that are most likely to confound results are those manifested on length scales that are similar to the studied biological phenomena while also considering the spatial correlation structure of genomic features. For example, although PCR-amplified fragments tend to be ~100 bp long, GC content can fluctuate across more extensive regions of the genome; therefore, PCR effects would be observable on these broader scales.

Identifying bias. The ChiLin quality control pipeline is a good starting point for understanding the quality and bias characteristics of ChIP–seq, DNase-seq and ATAC-seq samples. ChiLin reports quality control characteristics of reads and genome-level measures that reflect the tendency of reads to appear in clusters or in peak-like patterns[54]. These metrics can be used to identify low-quality samples and to flag data characteristics, such as high read redundancy rates that can lead to poor results. As quality control measures often depend on sequencing depth, a fixed number of reads need to be sampled when comparing the quality control measures of different data sets. NGS read characteristics can also be quantified using alternative software packages such as SAMstat[59], RNA-SeQC[60], RSeQC[61] and htSeqTools[62]. The software CHANCE[63] and HOMER[64] evaluate alternative enrichment quality control characteristics.

In most chromatin profiling applications, it is better to characterize bias from the genomic perspective instead of the read perspective. A commonly used approach for characterizing a single source of bias is as follows. First, the genome is partitioned into elements such as genes or genomic intervals, and the bias parameters such as GC content in each element are computed. Second, elements are grouped into bins according to these parameters. Finally, reads in each element are counted, and robust estimates of bias within each bin are calculated. Genomic length scales of the bias and the biological features should be taken into consideration when partitioning the genome. As the effects of bias are expected to be smooth functions, flexible functions such as splines[65] or locally estimated scatterplot smoothing (LOESS)[66] can be used instead of dividing data into bins. When there are multiple sources of bias and when the data is insufficient to partition the parameter space into bins, robust estimates of parameters can be calculated using techniques such as quantile regression[67]. Although it may be fairly easy to measure the relationship between NGS read counts and genomic features, further interpretation is complicated because different sources of bias may be correlated with each other and with biological factors. In addition, reducing the influence of bias requires read count variability to be taken into consideration.

Adjusting for sequencing depth. ChIP–seq studies usually involve the comparison of immunoprecipitated samples and input control samples, and ChIP–seq of one condition is sometimes compared with that of another condition. Although sequence depth represented as total read count is commonly used to normalize ChIP–seq data, this ignores differences in the proportion of immunoprecipitated reads to background reads. In PeakSeq, the genome is partitioned into 10-kb bins, and linear regression is used to compute the scaling constant between input control and immunoprecipitated samples[68]. Signal extraction scaling (SES) is an alternative global scaling method for ChIP–seq that separates reads in immunoprecipitated samples into signal and background components and that uses the background estimates for scaling[63]. This method partitions the genome into bins of equal size (1 kb) and uses the lower tail of the cumulative distribution function of counts within each of these bins to estimate the background signal. NCIS (normalization of ChIP–seq) uses a similar strategy[69] to select both window size and background read cutoff in an adaptive yet robust manner.

Splines
Flexible smooth nonlinear functions that are defined piecewise by polynomials for fitting nonlinear trends.

Locally estimated scatterplot smoothing
(LOESS). A simple yet robust method for fitting nonlinear trends.

Quantile regression
A statistical regression method that estimates the median or other quantile of the response variables and that is robust against outliers.

When comparing ChIP–seq between treatment groups, normalization schemes that are appropriate for normalizing input and immunoprecipitated samples may not be suitable for normalization among immunoprecipitated samples, especially when the signal-to-noise ratio varies between samples. The simplest approach of scaling read counts by the reciprocal of the total number of mapped reads may not work, as it

Table 2 | **Diagnosis and mitigation of bias in common analyses of next-generation sequencing chromatin profiling experiments**

Analysis type	Examples	Biases that are likely to influence results	Diagnosis and mitigation
Allele specificity	ChIP–seq, DNase-seq, ATAC-seq or MNase-seq read counts are associated with a SNP	• Sequencing errors • Priming efficiency • Reference genome to which reads are mapped • Read mapping algorithm • Differential cleavage bias in DNase-seq, ATAC-seq and MNase-seq	• Estimate sequence error rates modelled on sequence characteristics and use error estimates to account for these error rates[113] • Check for association with the read rather than the genome; for example, check whether the allelic imbalance predominate at 5′ or 3′ end of reads[114] • Use special-purpose mapping software[50,78,109] • Model nuclease-induced cleavage bias, or discard DNase-seq or ATAC-seq reads with 5′ ends close to the SNP[26]
Peak enrichment relative to genomic feature	ChIP–seq peaks are enriched at gene promoters, exons or CpG islands relative to other regions of the genome	• Chromatin effects • PCR amplification bias • Nucleic acid isolation • Read depth	• Collect statistics on enrichment trends in controls and in unrelated data sets that are obtained using the same genomic technology[15,115,116] • Model effects of GC or AT DNA sequence content[65] • Examine whether spatial characteristics of read distributions look like ChIP–seq peaks[117]; in ChIP–seq, a single isolated TF binding site is flanked by mostly positive strand reads upstream and negative strand reads downstream of the site • Carry out analysis for different numbers of reads and examine trend of enrichment as a function of total read count[111]
Read enrichment relative to genomic feature	Histone mark ChIP–seq read distributions relative to transcription start sites	• Chromatin effects • PCR amplification bias • Ratio of background read counts relative to specific ChIP • Read depth	• Compare with controls and other data sets that are obtained using the same genomic technology[9] • Examine quality control metrics related to specific versus nonspecific read quality[63]; if quality control metrics differ substantially between samples, then repeat the experiment to obtain more consistent data quality[14] • Examine spatial distribution of GC or AT DNA sequence content relative to genomic feature[24] • Carry out analysis on 5′ ends of reads separated by strand[26] • When using paired-end data, stratify reads by fragment length[26] • Carry out analysis using genomic control loci[26]; for example, exons tend to be GC-rich and are surrounded by less GC-rich sequence, and controls for exons could be intronic sequences with similar DNA sequence characteristics
Differential abundance between conditions	ChIP–seq, DNase-seq or ATAC-seq read-level enrichment or depletion in treatment relative to control	• Batch effects • PCR amplification bias • Chromatin effects • Nucleic acid isolation • Ratio of background read counts relative to specific ChIP • Read depth	• Test for association with known batch variables and identify unknown effects[101,103,104] • Analyse dependence of fragment abundance on DNA sequence composition, including GC content[34,65,118] • Include known quantitative factors in differential abundance analysis[101], for example, batch variables such as date of sequencing • Use unsupervised techniques, such as surrogate variable analysis, to remove systematic effects of unknown origin[104]
Association of genomic feature with cellular or organismal phenotype	In ChIP–seq, specific binding sites are associated with disease progression	• Batch effects • Cell-type-specific chromatin effects	• Test whether bias-associated variable is related to phenotype using surrogate variable analysis[104] • Contrast data from general assays such as DNase-seq and ATAC-seq with ChIP–seq that targets specific proteins
Association of one biological phenomenon with another	• Overlap of ChIP–seq peaks of two TFs • Claims of significant association between TF binding or differences in TF binding	• Antibody quality • Relative immunoprecipitation enrichment • Chromatin effects • PCR amplification bias • Read depth	• Check whether common sources of technical bias underlie observations • Carry out analyses using different levels of read sampling; sites with the strongest biological signal will be detected at a low read depth, whereas weaker sites will be detected as the read depth increases[55] • Choose meaningful background models to discover associations: ChIP–seq peaks of different TFs in the same cell line will often overlap relative to a background of random genomic loci, and most TF binding sites are found in cell-type-specific DNase-seq peak regions[26] • Use performance statistics, such as receiver–operator characteristic and precision-recall curves, to characterize the trade-off between sensitivity and specificity[119]
DNA motif analyses	Identification of TF binding sites in ChIP–seq	• Chromatin and fragmentation effects • PCR amplification bias • Nucleic acid isolation	• Evaluate bias and signal variability in controls[26] • Compare data with controls and data from other systems[116] • Evaluate results using independent data types

SNP, single-nucleotide polymorphism; TF, transcription factor.

is based on the specific assumption that the proportion of reads that map to the enriched portion of the genome is consistent between samples. Instead of scaling on the basis of total read counts, under the assumption that levels of TF binding are similar between samples, one could scale counts based on the total read count in peak regions. Total read counts may be strongly influenced by outliers; therefore, instead of scaling on the basis of total read counts, scaling can be based on the median read count within peak regions. Alternatively, more sophisticated scaling factors implemented in DESeq[70] or trimmed mean of M values (TMM)[71] implemented in edgeR[72] can be used. These methods calculate normalization factors after a feature-wise comparison between samples and the exclusion of outliers[71].

Quantile normalization equalizes the full distribution of read counts between samples instead of linear scaling. The assumption that enrichment distributions are the same between samples may not hold true in many chromatin profiling applications, especially when the TF of interest has different expression levels between conditions. Quantile normalization might also be adversely affected by bias and outlier effects, and could perform poorly when some samples contain a higher proportion of features with counts of zero than others[73].

MAnorm[74], which was developed for differential analysis of ChIP–seq data, assumes that data sets have a substantial number of peaks in common and that there is no global change in binding at these common peaks. MAnorm normalizes read counts in common peaks using robust linear regression to model the relationship between the logarithmic ratio of reads in the two samples relative to the average logarithmic read counts.

Choosing an appropriate normalization scheme requires prior knowledge of the system, and important considerations include the expected enriched fraction of the genome and the degree of consistency in signals between samples. We recommend assessing whether consistent results can be obtained using different normalization schemes. Normalization assumptions can also be evaluated using alternative technologies such as quantitative PCR on selected regions. Finally, chromatin spike-in controls can be included in genomic experiments for normalization purposes[57]. In many cases, although we would ideally want to study the absolute levels of binding, we have to accept the limitations of ChIP–seq and adapt by designing experiments in such a way that meaningful conclusions can be drawn from relative levels.

Duplicate reads. It is common to filter out duplicate reads in the course of chromatin analysis. Although filtering can have a slight impact on sensitivity, retaining these duplicates can have substantial and detrimental consequences on specificity[55]. Instead of either filtering out all duplicates or retaining all of them, a threshold of duplication can be used, above which additional copies are discarded. In ChIP–seq, DNase-seq and ATAC-seq, in which the coverage of local regions of the genome can be high, duplicates are expected and discarding duplicates is likely to distort quantification. It may be legitimate to handle duplicate reads differently in different analyses of the same data. For example, in ChIP–seq peak detection using model-based analysis of ChIP–seq (MACS), it may be prudent to use the option of discarding duplicates so as to avoid calling false peaks[55]. However, in the comparison of ChIP–seq signal between samples, local coverage may be so high that signal would be truncated without some inclusion of duplicates.

Modelling variation in NGS profiling data. In addition to variability due to stochastic counting processes, NGS data inevitably show variation that is greater than expected (that is, overdispersion) as a result of biases. The nature and severity of biases and overdispersion are strongly dependent on the scale of the genomic interval being analysed. Cleavage biases and sequencing errors may be observed at the single-nucleotide scale, PCR amplification biases become obvious at the ~100-bp scale, and chromatin structure effects are manifested across a broad range of scales from ~100 bp to >100 kb. Statistical power can be increased through the explanation of some of the bias-induced variation, and several distributions have been usefully applied for NGS analyses. The Poisson distribution — a simple single-parameter model that is suitable for modelling count data — tends to underestimate the variance in NGS data but can be used to model biases by allowing the parameter to vary as a function of genome position[75]. FIXSEQ, which is a preprocessing method for mitigating read count overdispersion effects, can improve the performance of analyses that are based on Poisson assumptions[76]. Alternatively, NGS data can be described using more complex distributions that allow the variance to be estimated separately from the mean, for example, the negative binomial[70,72,77], zero-inflated negative binomial[48] and beta negative binomial distributions[78]. When replicates are insufficient to allow robust estimates of variance to be made, simplifying assumptions about the relationship between the mean and the variance can be used to estimate variance by pooling regions with a similar mean[70,72,77]. Standard statistical diagnostics — including comparisons of theoretical and empirical distributions, analyses of residuals and simulations — are important for checking the validity of such models.

Peak detection. In enrichment analyses, when calling peaks in ChIP–seq, DNase-seq and ATAC-seq experiments, genomic regions that are associated with protein binding, histone modifications or open chromatin are determined by read density[2,68,75,79–84]. In cases of ChIP–seq in which input controls are available and representative of the bias in immunoprecipitated samples, peak calling methods can perform well without explicitly taking GC content and mappability into account. GC content and mappability are useful considerations when input control coverage is low or absent. PeakSeq[68], Probabilistic Inference for ChIP–seq (PICS)[84] and MOSAiCS[83] take mappability into consideration, although PeakSeq considers mappability on a much larger scale than the peak scale (~100 bp). Even in analyses that include input controls, adjusting for GC content may still be useful,

as GC bias can vary substantially from one input sample to another[56]. The MACS[75] peak detection algorithm takes neither GC content nor mappability explicitly into account; instead, it makes estimates of background signals from multiple nearby chromatin windows of different scales from the input controls. For TF ChIP–seq data with limited input coverage, the MACS background estimate from multiple windows provides a more robust ChIP enrichment evaluation than single-window estimates, which leads to consistently good performance across many data sets.

In ChIP–seq and MNase-seq, peak shape is another concept that can be used to identify peaks. Reads that map to the forward and reverse strands form characteristic patterns near TF binding sites and positioned nucleosomes[4,85,86]. In ChIP–seq, the fragmentation of DNA associated with a TF bound at a single isolated locus and the subsequent sequencing of fragment ends lead to a cluster of forward-strand tags 5′ of the binding sites and a cluster of reverse-strand tags 3′ of the binding sites. The distance separating these clusters is dependent on the size distribution of sequenced fragments and on the size of the local open chromatin region[75]. Algorithms that are designed to recognize the shape of ChIP–seq signal can be helpful in distinguishing chromatin- and PCR-induced effects from TF binding events. Similarly, in MNase-seq, well-positioned nucleosomes are bracketed by 5′ and 3′ reads. However, TFs or modified histones that bind across broad regions rather than at precise loci will produce a more diffuse distribution of ChIP–seq reads. In DNase-seq and ATAC-seq, patterns of reads in open chromatin regions result from a complex interplay of experimental effects with TF binding and nucleosome occupancy, among other biological factors[26]. The interpretation of these read patterns can help us to improve chromatin accessibility protocols and yield insights into ways in which chromatin is modified[51]. Local DNA sequence and mappability biases can result in read patterns that may be confused with true binding events.

DNase-seq footprint and chromatin landscape analyses. Although none of the DNase I footprinting algorithms developed so far explicitly take into account biases such as nucleosome occupancy, DNA sequence-dependent cleavage and TF binding (which can affect the patterns of DNase I cleavage), the way in which footprint significance is calculated and interpreted acknowledges bias effects to different extents.

The first algorithms developed for DNase-seq footprint identification reduce sensitivity to the effects of sequence and other biases by ranking the read counts at each position in the central and flanking regions[8,87]. Although these approaches do not explicitly model cleavage bias effects, the rank transformation prevents footprints from being identified from outlier signals at a few nucleotides. Another method, as a preprocessing step, uses a polynomial to approximate signal over several nucleotides to reduce the effects of nucleotide-specific bias[7]. A recently developed method[9] estimates footprint significance on the basis of the observed tag count instead of the rank transformation. In this approach, *P* values are computed by shuffling individual reads within local regions. The resulting null distribution severely underestimates the variability of DNase-seq data, and the significance of putative footprint regions are consequently overestimated, which leads to high false discovery rates[26,88]. Analyses of DNase I cleavage patterns or evidence of TF binding at single-nucleotide resolution require statistical modelling that accurately represents the intrinsic variability of DNase I cleavage.

Another way of distinguishing bona fide TF-induced footprints from bias-induced artefacts is to take peak shape into account. The Wellington algorithm makes use of the observation that DNase I cuts tend to occur in a strand-specific way 5′ of the TF binding sites and computes significance based on the numbers of strand-specific reads observed in a single flank relative to the footprint region[88].

Although the occupancy of some TFs, such as CCCTC-binding factor (CTCF), is associated with DNase-seq footprint patterns, for many TFs these patterns are weak or, in cases such as the androgen receptor, non-identifiable using current methods[26]. TFs interact with chromatin in various ways, which results in diverse chromatin landscapes near TF binding sites. Some TFs, such as CTCF, bind in regions that are nucleosome-free and that are flanked by well-organized nucleosome arrays[89], whereas others bind in such a way that nucleosome occupancy is dependent on binding orientation[51]. Yet other TFs, such as the oestrogen receptor, bind in a way that does not strongly depend on nucleosome occupancy[90]. CENTIPEDE[91] and, more recently, protein interaction quantitation (PIQ)[51] analyse the shape and magnitude of DNase-seq profiles together with TF position weight matrices (PWMs). PIQ explores the local chromatin environment surrounding TF binding sites and has been used to classify TFs in terms of their effect on chromatin remodelling.

Domain calling from ChIP–seq. ChIP–seq that targets certain histone modifications — including histone H3 lysine 9 trimethylation (H3K9me3), H3K27me3 and H3K36me3 — tends to produce diffuse regions of enrichment rather than the sharp peaks that are typically observed in ChIP–seq of TFs. These broad signals are challenging to analyse because the signal is diffuse and at times difficult to distinguish from the confounding effects of biases. In addition, these broad regions of enrichment can vary greatly in extent and have undulating profiles across the genome. Although most current analyses summarize these patterns as genomic intervals, other summaries might be more appropriate for describing diverse patterns that could be produced through various biological mechanisms, including co-transcriptional enzymatic activity, local diffusion, nucleosome replacement and looping.

Domain calling algorithms typically segment the genome into bins before grouping bins together as domains[92–94]. SICER[92] identifies broad intervals by first identifying bins with read counts above a predefined threshold and subsequently computing a statistic for the aggregate of several of such bins, which are possibly

separated by small numbers of low-read bins[92]. RSEG uses the hidden Markov model framework to specifically identify the boundaries of broad domains[93]. In this approach, individual sample read counts in genomic intervals are modelled using a negative binomial distribution, and the relationship between the read counts in an immunoprecipitated sample and those in an input sample is modelled using a difference of negative binomial distributions. Combinations of histone modifications are often observed together in chromatin states, which are patterns indicative of distinct modes of biological activity. These patterns may be identified by integrating multiple ChIP–seq data sets on histone modifications using the ChromHMM[95] or SegWay[96] algorithms.

ChIP–seq peak deconvolution. Multiple sites of protein–DNA interaction in close proximity to one another might be identified as a single ChIP–seq-enriched region. CSDeconv[97], Genome Positioning System (GPS)[98] and PICS[84] deconvolute ChIP–seq signal to predict interaction loci using estimates of strand-specific read displacement distributions relative to TF binding sites. PICS explicitly accounts for mappability, whereas GPS can control for biases by including input control data in its deconvolution procedure. Paired-end sequencing in ChIP–seq produces data in which both ends of every fragment are known, and no inference of fragment size is necessary. dPeak[99] resolves complex paired-end ChIP–seq peak regions into multiple loci with a higher accuracy than single-end analyses. The model used in dPeak takes nonspecific binding into account and allows shift distributions to be non-uniform across all binding sites.

Differential region identification. In a population of cells, TF occupancies at a given locus might differ between cells and over time. TF binding is therefore better described by a continuous variable rather than a binary variable, as changes in binding can be as biologically relevant as the apparent loss or gain of binding sites. Although strong changes in TF binding may be observed from single-replicate ChIP–seq comparisons, few studies have included the replicates that are required to quantify signal variability and to allow detection of more subtle differences. Methodologies for identifying differential count enrichment, including DEseq[70,77] and edgeR[72,77], model count data in a way that is consistent with the counting process. These methods allow the use of offsets — parameters that capture artefacts[100] such as GC content — which are taken into account in the computation of differential enrichment. Such offsets can be computed using methods such as conditional quantile normalization (CQN)[65]. The use of input controls has been suggested to distinguish TF binding signal from background levels before comparisons can be made[69]. A procedure for comparative analysis of ChIP–seq peaks is carried out in DBChIP[69], which uses negative binomial modelling to estimate the overdispersion of reads between samples. Comparisons of ChIP–seq data obtained using different antibodies or from different laboratories are problematic, as differential TF binding could be confounded by systematic biases such as differences in antibodies and ChIP conditions.

In studies involving the comparison of multiple samples, it is important to look out for batch effects, which often arise from unknown sources of technical variation[101]. Statistical techniques may be used to model effects that arise from observable batch groupings, such as date of sequencing[102]. Sometimes, these effects cannot be associated with any particular batch annotation but may still be observed in clustering analyses that reveal clusters of samples which are inconsistent with any biological treatment groupings. Analyses such as surrogate variable analysis maybe used to mitigate these batch effects of unknown origin[103,104].

Chromatin interaction analyses. In Hi-C experiments, to quantify the interaction frequency between chromatin loci, pairs of DNA sequence fragments that are in close three-dimensional proximity to one another *in vivo* are ligated together and sequenced[10,11]. Although many of the biases that arise in this experiment may be modelled explicitly[105,106], an effective alternative perspective eliminates the need to explicitly account for these factors[107]. This new analysis assumes that the observed interaction frequency between fragments can be factored into a product of the visibility of each of the individual fragments and an interaction frequency term that is the variable of interest[107]. The bias identified in this way agrees to a remarkable degree with the bias detected through explicit modelling, which adds confidence to both approaches. Hi-C interaction analyses in gigabase-scale genomes, such as the human genome, require extremely high sequencing depths even for ~50 kb-resolution of interaction frequencies. Targeted approaches can be used to produce higher-resolution interaction maps at selected genomic regions. Chromatin conformation capture carbon copy (5C) experiments[108] target specific regions of the genome using PCR primers, and ChIA-PET[12] uses ChIP to pull down loci that interact with particular proteins. In the analysis of data from 5C and ChIA-PET, biases and noise introduced in the selection step also need to be taken into consideration in the calculation of interaction frequencies.

Conclusion and future directions

The use of NGS technologies in combination with adaptations of established experimental protocols is deepening our understanding of chromatin biology, including epigenetic and post-transcriptional gene regulation, mechanisms underlying developmental differentiation and cell reprogramming, and the impact of genetic variation on phenotypes. Investigators should be cautious in analysing NGS data to avoid interpreting biases and technical artefacts as biological phenomena. The lack of standard protocols is a major challenge in the analysis of such data, as a source of bias that is negligible in one laboratory might be large enough to distort results in another. ChIP–seq studies of TFs with good antibodies in cell lines are now ubiquitous, and lists of several thousand TF binding sites can be reliably detected by several available algorithms. Challenges

Surrogate variable analysis
A statistical analysis to identify and model variables that are not explicitly annotated but that have measureable effects.

remain in the analysis of tissue samples and of samples with very few cells, as well as in the representation of broad signals. Better methods are also needed to compare chromatin profiles between treatment groups and to account for variability in sample quality, enrichment level, batch effect and read depth. An important emerging field is the interpretation of TF occupancy in relation to chromatin accessibility profiling methods such as DNase-seq and ATAC-seq. As the use of NGS technologies and the technologies themselves evolve, the detection and normalization of biases will require the development of effective and flexible methods that are implemented in efficient modular computational packages.

1. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
This paper reports the first use of MNase digestion followed by ChIP–seq to characterize genome-wide patterns of 20 varieties of histone lysine and arginine methylation. It identifies common modifications that are associated with active and repressed regions of the genome, transcription start sites, enhancers and insulator elements.
2. Johnson, D., Mortazavi, A., Myers, R. & Wold, B. Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **80**, 1497–1502 (2007).
3. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
4. 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 proposes using the distribution of oriented reads to discriminate between real TF binding sites and artefacts.
5. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).
6. He, H. H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nature Genet.* **42**, 343–347 (2010).
7. Boyle, A. P. *et al.* High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells. *Genome Res.* **21**, 456–464 (2011).
8. Hesselberth, J. R. *et al.* Global mapping of protein–DNA interactions *in vivo* by digital genomic footprinting. *Nature Methods* **6**, 283–289 (2009).
9. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
10. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
11. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
12. Fullwood, M. J. *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
13. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**, 1213–1218 (2013).
14. Landt, S. G. *et al.* ChIP–seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
15. Teytelman, L. *et al.* Impact of chromatin structures on DNA processing for genomic analyses. *PLoS ONE* **4**, e6700 (2009).
16. Modak, S. P. & Beard, P. Analysis of DNA double- and single-strand breaks by two dimensional electrophoresis: action of micrococcal nuclease on chromatin and DNA, and degradation in vivo of lens fiber chromatin. *Nucleic Acids Res.* **8**, 2665–2678 (1980).
17. Zentner, G. E. & Henikoff, S. Surveying the epigenomic landscape, one base at a time. *Genome Biol.* **13**, 250 (2012).
18. Telford, D. J. & Stewart, B. W. Micrococcal nuclease: its specificity and use for chromatin analysis. *Int. J. Biochem.* **21**, 127–137 (1989).
19. Henikoff, J. G., Belsky, J. A., Krassovsky, K., Macalpine, D. M. & Henikoff, S. Epigenome characterization at single base-pair resolution. *Proc. Natl Acad. Sci. USA* **108**, 18318–18323 (2011).
20. Tillo, D. *et al.* High nucleosome occupancy is encoded at human regulatory sequences. *PLoS ONE* **5**, e9129 (2010).
21. Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516–520 (2011).
22. Gaffney, D. J. *et al.* Controls of nucleosome positioning in the human genome. *PLoS Genet.* **8**, e1003036 (2012).
23. Fan, X. *et al.* Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3′-end formation. *Proc. Natl Acad. Sci. USA* **107**, 17945–17950 (2010).
24. Chung, H.-R. *et al.* The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS ONE* **5**, e15754 (2010).
25. Campbell, V. W. & Jackson, D. A. The effect of divalent cations on the mode of action of DNase I. The initial reaction products produced from covalently closed circular DNA. *J. Biol. Chem.* **255**, 3726–3735 (1980).
26. He, H. H. *et al.* Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature Methods* **11**, 73–78 (2014).
This study shows how fragment size selection in DNase-seq can have a large impact on peak identification and that intrinsic DNase I cleavage bias can be mistaken as TF binding footprints.
27. Vierstra, J. Wang, H., John, S., Sandstrom, R. & Stamatoyannopoulos, J. A. Coupling transcription factor occupancy to nucleosome architecture with DNase–FLASH. *Nature Methods* **11**, 66–72 (2014).
28. Lazarovici, A. *et al.* Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl Acad. Sci. USA* **110**, 6376–6381 (2013).
29. Grøntved, L. *et al.* Rapid genome-scale mapping of chromatin accessibility in tissue. *Epigenetics Chromatin* **5**, 10 (2012).
30. Van Heesch, S. *et al.* Systematic biases in DNA copy number originate from isolation procedures. *Genome Biol.* **14**, R33 (2013).
31. 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).
32. Gilfillan, G. D. *et al.* Limitations and possibilities of low cell number ChIP–seq. *BMC Genomics* **13**, 645 (2012).
33. Dabney, J. & Meyer, M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* **52**, 87–94 (2012).
34. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72 (2012).
This study shows the importance of selecting the correct genomic interval for bias analysis, as some sources of bias are best modelled using properties of DNA fragments rather than DNA reads.
35. Wheeler, T. J. *et al.* Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.* **41**, D70–D82 (2013).
36. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
38. 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).
39. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genet.* **41**, 1061–1067 (2009).
40. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
41. Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS ONE* **7**, e30377 (2012).
42. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genet.* **42**, 631–634 (2010).
43. 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).
44. Day, D. S., Luquette, L. J., Park, P. J. & Kharchenko, P. V. Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome Biol.* **11**, R69 (2010).
45. Wang, T. *et al.* Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl Acad. Sci. USA* **104**, 18613–18618 (2007).
46. Pickrell, J. K., Gaffney, D. J., Gilad, Y. & Pritchard, J. K. False positive peaks in ChIP–seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics* **27**, 2144–2146 (2011).
47. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
48. Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W. & Lieb, J. D. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol.* **12**, R67 (2011).
49. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).
50. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7**, 522 (2011).
51. Sherwood, R. I. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotech.* **32**, 171–178 (2014).
52. König, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Struct. Mol. Biol.* **17**, 909–915 (2010).
53. Daley, T. & Smith, A. D. Predicting the molecular complexity of sequencing libraries. *Nature Methods* **10**, 325–327 (2013).
54. Marinov, G. K., Kundaje, A., Park, P. J. & Wold, B. J. Large-scale quality analysis of published ChIP–seq data. *G3 (Bethesda)* **4**, 209–223 (2014).
55. Chen, Y. *et al.* Systematic evaluation of factors influencing ChIP–seq fidelity. *Nature Methods* **9**, 609–614 (2012).
56. Ho, J. W. K. *et al.* ChIP–chip versus ChIP–seq: lessons for experimental design and data analysis. *BMC Genomics* **12**, 134 (2011).
57. Bonhoure, N. *et al.* Quantifying ChIP–seq data: a spiking method providing an internal reference for sample-to-sample normalization. *Genome Res.* **24**, 1157–1168 (2014).
58. Kidder, B. L., Hu, G. & Zhao, K. ChIP–seq: technical considerations for obtaining high-quality data. *Nature Immunol.* **12**, 918–922 (2011).
59. Lassmann, T., Hayashizaki, Y. & Daub, C. O. SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics* **27**, 130–131 (2010).
60. DeLuca, D. S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).
61. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).

62. Planet, E. & Attolini, C. S., Reina, O., Flores, O. & Rossell, D. htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics* **28**, 589–590 (2012).
63. Diaz, A., Nellore, A. & Song, J. S. CHANCE: comprehensive software for quality control and validation of ChIP–seq data. *Genome Biol.* **13**, R98 (2012).
64. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
65. Hansen, K. D., Irizarry, R. A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).
66. Cleveland, W. S. Robust locally and smoothing weighted regression scatterplots. *J. Am. Stat. Soc.* **74**, 829–836 (2013).
67. Koenker, R. & Hallock, K. F. Quantile regression. *J. Econ. Perspect.***15**, 143–156 (2013).
68. Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP–seq experiments relative to controls. *Nature Biotech.* **27**, 66–75 (2009).
69. Liang, K. & Keles, S. Detecting differential binding of transcription factors with ChIP–seq. *Bioinformatics* **28**, 121–122 (2012).
70. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
71. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
72. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
73. Dillies, M.-A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**, 671–683 (2012).
74. Shao, Z., Zhang, Y., Yuan, G.-C., Orkin, S. H. & Waxman, D. J. MAnorm: a robust model for quantitative comparison of ChIP–seq data sets. *Genome Biol.* **13**, R16 (2012).
75. Zhang, Y. *et al.* Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9**, R137 (2008).
This study introduces the idea of estimating background effects using sliding windows on multiple scales. MACS remains one of the most widely used and best-performing algorithms for ChIP–seq peak calling.
76. Hashimoto, T. B., Edwards, M. D. & Gifford, D. K. Universal count correction for high-throughput sequencing. *PLoS Comput. Biol.* **10**, 14–18 (2014).
77. Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protoc.* **8**, 1765–1786 (2013).
78. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747–749 (2013).
79. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**, 651–657 (2007).
80. Ji, H. *et al.* An integrated software system for analyzing ChIP–chip and ChIP–seq data. *Nature Biotech.* **26**, 1293–1300 (2008).
81. 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**, 1–9 (2008).
82. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP–seq data. *Nature Methods* **5**, 829–834 (2008).
83. Sun, G., Chung, D. & Liang, K. Statistical analysis of ChIP–seq data with MOSAiCS. *Methods Mol. Biol.* **1038**, 193–212 (2013).
84. Zhang, X. *et al.* PICS: probabilistic inference for ChIP–seq. *Biometrics* **67**, 151–163 (2011).
85. Kornacker, K., Rye, M. B., Håndstad, T. & Drabløs, F. The Triform algorithm: improved sensitivity and specificity in ChIP–seq peak finding *BMC Bioinformatics* **13**, 176 (2012).
86. Kumar, V. *et al.* Uniform, optimal signal processing of mapped deep-sequencing data. *Nature Biotech.* **31**, 615–622 (2013).
87. Chen, X., Hoffman, M. M., Bilmes, J. A., Hesselberth, J. R. & Noble, W. S. A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics* **26**, i334–i342 (2010).
88. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41**, e201 (2013).
89. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* **4**, e1000138 (2008).
90. He, H. H. *et al.* Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Res.* **22**, 1015–1025 (2012).
91. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447–455 (2011).
92. Zang, C. *et al.* A clustering approach for identification of enriched domains from histone modification ChIP–seq data. *Bioinformatics* **25**, 1952–1958 (2009).
93. Song, Q. & Smith, A. D. Identifying dispersed epigenomic domains from ChIP–seq data. *Bioinformatics* **27**, 870–871 (2011).
94. Wang, J., Lunyak, V. V. & Jordan, I. K. BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP–seq datasets. *Bioinformatics* **29**, 492–493 (2013).
95. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotech.* **28**, 817–825 (2010).
96. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9**, 473–476 (2012).
97. Lun, D. S., Sherrid, A., Weiner, B., Sherman, D. R. & Galagan, J. E. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP–seq data. **12**, 1–12 (2009).
98. Guo, Y. *et al.* Discovering homotypic binding events at high spatial resolution. *Bioinformatics* **26**, 3028–3034 (2010).
99. Chung, D. *et al.* dPeak: high resolution identification of transcription factor binding sites from PET and SET ChIP–seq data. *PLos Comput. Biol.* **9**, 9–11 (2013).
100. Li, J., Jiang, H. & Wong, W. H. Modeling non-uniformity in short-read rates in RNA-seq data. *Genome Biol.* **11**, 1–11 (2010).
101. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Rev. Genet.* **11**, 733–739 (2010).
This review discusses the importance of modelling batch effects in genome-wide analyses and statistical techniques for such analyses.
102. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
103. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).
104. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
105. Hu, M. *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**, 3131–3133 (2012).
106. Hu, M. *et al.* Bayesian inference of spatial organizations of chromosomes. *PLoS Comput. Biol.* **9**, e1002893 (2013).
107. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization *Nature Methods* **9**, 999–1003 (2012).
This study proposes a novel decomposition scheme for the analysis of Hi-C data that separates visibility and interaction components.
108. 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).
109. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
110. Zeng, W. & Mortazavi, A. Technical considerations for functional sequencing assays. *Nature Immunol.* **13**, 802–807 (2012).
111. Jung, Y. L. *et al.* Impact of sequencing depth in ChIP–seq experiments. *Nucleic Acids Res.* **42**, e74 (2014).
112. Zhang, Y. *et al.* Intrinsic histone–DNA interactions are not the major determinant of nucleosome positions *in vivo*. *Nature Struct. Mol. Biol.* **16**, 847–852 (2009).
113. Bravo, H. C. & Irizarry, R. A. Model-based quality assessment and base-calling for second-generation sequencing data. *Biometrics* **66**, 665–674 (2010).
114. Pickrell, J. K., Gilad, Y. & Pritchard, J. K. Comment on "Widespread RNA & DNA sequence differences in the human transcriptome". *Science* **335**, 1302 (2012).
115. Teytelman, L., Thurtle, D. M., Rine, J. & van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl Acad. Sci. USA* **110**, 18602–18607 (2013).
116. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).
117. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Rev. Genet.* **10**, 669–680 (2009).
118. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
119. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning.* (Springer, 2001).

Acknowledgements
The authors thank members of X.S.L and M. Brown's laboratories for their discussions. This work is supported by the US National Institutes of Health grant R01GM099409.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
ChiLin: http://liulab.dfci.harvard.edu/software

ALL LINKS ARE ACTIVE IN THE ONLINE PDF