Published in final edited form as:
Nat Rev Genet. 2014 November ; 15(11): 709–721. doi:10.1038/nrg3788.

Identifying and Mitigating Bias in Next-Generation Sequencing Methods for Chromatin Biology

Clifford A. Meyer[1,2] and **X. Shirley Liu**[1,2]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, Massachusetts 02115, USA

[2]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA

Abstract

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

Introduction

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 ones, to enable genome-wide investigations of a broad spectrum of chromatin phenomena (Figure 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 bias. 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 has 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 artifacts that have been observed in chromatin profiling NGS experiments, and we describe the analytical strategies that have been developed to handle them. Although RNA also plays 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

Address correspondence and requests for reprints to: Xiaole Shirley Liu, Professor, Ph.D, Dept of Biostats and Comp Bio, Dana-Farber Cancer Institute, Harvard School of Public Health, 450 Brookline Ave, Mail CLS-11007, Boston, MA 02215, Tel: (617) 632-2472, xsliu@jimmy.harvard.edu; Clifford A Meyer, Ph.D, Dept of Biostats and Comp Bio, Dana-Farber Cancer Institute, Harvard School of Public Health, 450 Brookline Ave, Mail CLS-11007, Boston, MA 02215, Tel: (617) 632-5337, cliff@research.dfci.harvard.edu.

computational biologists alike, and central to experimental design, protocol selection and data analyses. We first describe common sources of bias that arise in NGS chromatin profiling experiments before discussing 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 we outline how they can be used to handle bias 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 artifacts. Here 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 where we seek to quantify the protein-DNA interactions of a specific protein, DNA fragmentation, usually via sonication, is required before protein bound fragments are isolated by immunoprecipitation[14]. The mechanical characteristics of chromatin vary across the genome creating fluctuations in DNA fragility. Heterochromatin, not generally associated with transcription factor 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, sample-specific chromatin configuration induced biases. 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 produced from the same cell line under consistent conditions and using the same protocol may be combined as a control.

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) making this enzyme particularly useful for nucleosome occupancy characterization via MNase-seq. MNase induces single stranded 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, stably bound transcription factor[17] or refractory DNA sequence[18]. In MNase-seq studies, fragments of approximately one nucleosome length (~147bp) are typically selected for sequencing[6]. Different size ranges of MNase digested fragments have been shown to reveal different patterns of enrichment[19]. MNase-seq data therefore ought to be interpreted relative to fragment length distribution. Studies have found nucleosomes to occupy regions that are more GC rich than their neighboring regions[20–22], and nucleosomes to be 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 effected by the cleavage reaction temperature[18].

DNase I is a nuclease that, like MNase, generates double stranded breaks by nicking complementary strands of DNA one strand at a time[25]. Unlike MNase, DNase I has not been reported to have significant exonuclease activity, operating in a ‘hit and run’ mode instead of ‘nibbling’ at the ends of DNA until an obstruction is reached. The efficiency of DNase-seq in identifying transcription factor (TF) binding sites is highly dependent on fragment size, short fragments (<100bp) being more efficient than longer ones for several transcription factors. In contrast, longer fragments (>150bp) tend to span entire nucleosomes[26,27], and less likely to cluster around open chromatin regions (Figure 2).

Sites of DNase I cleavage are strongly affected by the precise DNA sequence in the 3 nucleotides 5′ and 3′ of the cleavage site, and this bias is strand specific[28]. Intrinsic DNase I cleavage bias is particularly evident when conducting an analysis of a set of sites in aggregate, where the genomic loci are aligned by the transcription factor 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[30] is also known to cleave DNA in a sequence dependent fashion.

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 commonly used to separate nucleic acids from protein[31]. 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 the enzyme proteinase K. Incomplete digestion can, however, result in DNA binding proteins carrying a fraction of DNA into the phenol phase leading to uneven genome coverage due to chromatin effects[31]. A similar differential solubility phenomenon has been used in Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE)[32] as an alternate to DNase I to determine regions of open chromatin.

PCR amplification biases and duplications

Multiple instances of the same sequence read in an NGS dataset can originate from mistaking one feature for two in sequencing image analysis, sequencing PCR amplicons derived from the same original fragment, or multiple fragments occurring in the original sample. This issue is particularly troublesome with small amounts of starting material[33].

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[34], which could be exacerbated with increasing PCR cycles. This is often manifest as a bias towards GC rich fragments, although not necessarily on extremely high GC rich regions[35]. While 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[35]. 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[36], including paralogous genes, and differences between the sequenced genome and the reference introduce coverage bias between different regions of the genome. Efficient mapping algorithms, including MAQ[37], BWA[38], Bowtie[39], mrFAST[40] and SOAP2[41], that take advantage of the short read length to align NGS reads with the reference genome, 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[42].

The proportion of a genome to which a sequence read may be uniquely assigned depends on 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 a greater uniformity of coverage[42]. Regions to which reads cannot be mapped have often be considered as less likely to be functional; they are often repetitive elements associated with transposon activity. Although most investigators ignore such regions, analysis of repeats using specialized methods[43,44] has revealed significant associations between chromatin marks[45] and transcriptions factors[46] 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 artifact-derived ‘sticky’ regions are frequently observed as ChIP-seq and DNase-seq peaks[47], sometimes as the ‘strongest’ peaks, 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 SNPs, 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[48,49]. 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[50]. Simply masking known SNP positions in the genome can lead to other artifacts due to a combination of factors including multiple SNPs in close proximity, unknown SNPs and similar sequences in other regions of the genome[51].

TF binding characteristics

The characteristics of TF binding to DNA differ significantly between factors[52]. Nucleosome positioning relative to the TF binding site, strength of binding, binding kinetics, and the tendency of a factor to bind in conjunction with other factors or potentially through the recognition of histone posttranslational modifications could influence observed signals. Some transcription factors are therefore more readily detected by TF binding inference techniques based on ATAC-seq, DNase-seq or MNase-seq.

Classical DNase I footprinting studies have shown that TF binding often modulates the pattern of DNase I cleavage on and flanking the nucleotides at the site of protein-DNA interaction, 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 for Bias Mitigation

In attempting to maximize discovery using limited research budgets, investigators tend to carry out minimal controls and replicates in NGS experiments. Controls, nevertheless, are required to accurately assess the effects of bias and replicates are needed to make an assessment of data variability. In experiments involving the comparison of multiple samples, bias effects often produce observable differences between sample batches. Success in correcting for such batch effects hinges on good experimental design. In particular, it is suggested that biologically distinct treatment groups 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 way for all samples (Figure 3). Here 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 arising from PCR amplification can often be confused with multiple fragments having one end in common in the original sample. Paired-end sequencing can help distinguish 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[53].

The number of informative reads produced from an NGS experiment depends on sample quality, sequencing technology and protocol, amongst other factors. As a result, NGS datasets can differ substantially in *read count* and the observed number and distribution of different DNA species, which reflect *library complexity*. Deep sequencing of low complexity libraries produces repeated observations of some species yielding less information than high complexity libraries, making methods to characterize library complexity useful diagnostic tools for NGS analysis[54]. In addition, the ENCODE consortium[55] PBC metric, which is the ratio between 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 bias

Controls 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 transcription factor binding signals may be observed because regions of transcription factor binding also tend to be regions where chromatin is more amenable to fragmentation[15]. Although, for the sake of economy, input controls are often sequenced to a lower depth than the ChIP samples, this is not recommended. The broader genomic distribution of signal in chromatin input DNA requires this input be sequenced deeper than ChIP-seq for accurate results[56,57].

Another issue is the potential difference in bias between the samples of interest and the controls. Although ChIP-seq input controls can tell us about mappability, copy number effects, broad chromatin accessibility, and other sources of bias have been found to vary substantially between control and ChIP samples[49,57]. 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 demonstrated by spiking human Hela cell chromatin into mouse samples for ChIP targeting subunits of RNA polymerases II and III[58]. This control may be especially useful in ChIP-seq studies of histone modifications. Although we believe 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.

When conducting a ChIP-seq experiment using an untested antibody it is crucial to perform a number of control experiments to establish the specificity of the antibody in genome-wide experiments[14,59]. Such experiments include the use of different antibodies as well as knockdown or knockout of the target protein. Antibody effects such as epitope masking can result in antibody-specific biases for the same transcription factor[59].

Controls for enzymatic cleavage assays: Genomic assays, including ATAC-seq, DNase-seq, and MNase-seq, that are based on the selection of fragments produced from enzymatic DNA cleavage 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 by DNA sequence 5′ and 3′ 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 the latter 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 for nucleosome positioning signals[24].

Analytical Techniques for Bias Correction

In this section we discuss issues that are generally applicable in NGS chromatin profiling analyses as well as methods implemented as software for specific analytical tasks. The general issues include identifying biases most likely to confound results, characterizing bias, adjusting for sequencing depth, handling duplicate reads, and modeling 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 artifacts that might effect a variety of analysis types as well as ways of diagnosing and correcting these effects.

Length scales of bias and biological features

Genomic analyses are carried out over length scales from 1bp SNP variant analyses to ~10bp DNase I footprint analyses, ~100bp TF ChIP-seq peak calling, and ~100bp to ~100kb 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 effect fragments of ~100bp. The biases most likely to confound results are those that are manifest 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 ~100bp 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 (http://liulab.dfci.harvard.edu/ChiLin). ChiLin reports quality control characteristics of reads as well as genome level measures that reflect the tendency of reads to appear in clusters or in peak-like patterns[55]. 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. Quality control measures often depend on sequencing depth, therefore a fixed number of reads need to be sampled when comparing the QC measures of different data sets. NGS read characteristics can also be quantified using alternative software packages such as, SAMstat[60], RNA-SeQC[61], RSeQC[62], and htSeqTools[63]. CHANCE[64] and HOMER[65] software evaluate alternative enrichment quality control characteristics.

In most chromatin profiling applications it is better to characterize bias from the genomic instead of the read perspective. A commonly used approach for characterizing a single source of bias is as follows: first, partition the genome into elements such as genes or genomic intervals and compute the bias parameters such as GC content in each element; second, group elements into bins according to these parameters; third, count reads in each element and calculate robust estimates of bias within each bin. 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[66] or loess[67] can be used instead of dividing data into bins. When there are multiple sources of bias and data is insufficient to partition the parameter space into bins, robust estimates of parameters can be calculated using techniques such as quantile regression[68]. Although it may be relatively easy to measure the relationship between NGS read counts and genomic features, further interpretation is complicated as different sources of bias may be correlated with each other and with biological factors. In addition, reducing the influence of bias requires that read count *variability* be taken into consideration.

Adjusting for sequencing depth

ChIP-seq studies usually involve the comparison of immunoprecipitate (IP) and input control samples, and sometimes ChIP-seq of one condition is compared with another. Although sequence depth represented as total read count is commonly used to normalize ChIP-seq data, this ignores differences in the proportion of IP to background reads. In PeakSeq, the genome is partitioned into 10kb bins and linear regression is used to compute the scaling constant between input control and IP samples[69]. SES is an alternative global scaling method for ChIP-seq that separates reads in IP samples into signal and background components and uses the background estimates for scaling[64]. This method partitions the genome into bins of equal size (1kb) and uses the lower tail of the cumulative distribution function of counts within each of these bins to estimate the background signal. NCIS employs a similar strategy[70] selecting both window size and background read cutoff in an adaptive yet robust manner.

When comparing ChIP-seq between treatment groups, normalization schemes appropriate for normalizing input and IP samples may not be appropriate for normalization amongst IP 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 is based on the specific assumption that the proportion of reads mapping to the enriched portion of the genome is consistent between samples. Instead of scaling based on total read counts, under the assumption that levels of TF binding is 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 so instead of scaling on total counts scaling can be based on the median read count within peak regions or using more sophisticated scaling factors implemented in DESeq[71], or Trimmed Mean of M values (TMM)[72] implemented in the edgeR[73]. These methods calculate normalization factors after a feature-wise comparison between samples and the exclusion of outliers[72].

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 in many chromatin profiling applications, especially when the TF of interest has different expression levels between conditions. Quantile normalization might also be adversely impacted by bias and outlier effects, and could perform poorly in cases when some samples contain a higher proportion of features with counts of zero relative to others[74].

MAnorm[75], developed for differential analysis of ChIP-seq data, assumes 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 log ratio of reads in the two samples relative to the average log read counts.

Choosing an appropriate normalization scheme requires prior knowledge of the system, important considerations being the expected enriched fraction of the genome and the degree of consistency in signals between samples. We recommend assessing if consistent results can be obtained using different normalization schemes. Normalization assumptions can also be evaluated using alternative technologies such as qPCR on selected regions. Finally, chromatin spike-in controls can be included in genomic experiments for normalization purposes[58]. 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 a significant detrimental impact on specificity[56]. Instead of either filtering all duplicates or retaining them, a threshold of duplication can be used, above which additional copies are discarded. In ChIP-seq, DNase-seq or ATAC-seq, where 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 MACS it may be prudent to use the option of discarding duplicates so as not to call false peaks[56]. On the other hand 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.

Modeling variation in NGS profiling data

In addition to variability due to stochastic counting processes, NGS data inevitably displays greater than expected variation (over-dispersion) due to bias. The nature and severity of bias and over-dispersion is strongly dependent on the scale of the genomic interval being analyzed. Cleavage biases and sequencing errors may be observed at the single nucleotide scale, PCR amplification biases become manifest at the ~100bp scale, and chromatin structure effects are manifest across a broad range of scales from ~100bp to greater than ~100kb. Statistical power can be increased through the explanation of some of the bias induced variation, and several distributions have been usefully employed for NGS analysis. The Poisson distribution, a simple single parameter model suitable for modeling count data, tends to underestimate the variance in NGS data but can be used to model bias by allowing the parameter to vary as a function of genome position[76]. FIXSEQ, a preprocessing method for mitigating read count over-dispersion effects, can improve the performance of analyses that are based on Poisson assumptions[77]. Alternatively, NGS data can be described using more complex distributions such as the negative binomial[71,73,78], zero-inflated negative binomial[49], and beta negative binomial distributions[79], that allow the variance to be estimated separately from the mean. When replicates are insufficient to allow robust estimates of variance to be made, simplifying assumptions about the relation between the mean and variance can be used to allow variance to be estimated by pooling regions with a similar mean[71,73,78]. Standard statistical diagnostics including comparisons of theoretical and empirical distributions, analysis of residuals and simulations are important to check 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[2,69,76,80–85]. For ChIP-seq, in cases where input controls are available and representative of the bias in IP 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[69], PICS[85] and MOSAiCS[84] take mappability into consideration, although PeakSeq considers mappability on a much larger scale than the peak scale (~100bp). 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[57]. The MACS[76] peak detection algorithm takes neither GC content nor mappability explicitly into account, and instead 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, leading to consistently good performance across many datasets.

In ChIP-seq and MNase-seq, peak shape is another concept that can be used to identify peaks. Reads mapping to the forward and reverse strand form characteristic patterns near TF binding sites and positioned nucleosomes[4,86,87]. In ChIP-seq, the fragmentation of DNA associated with a TF bound at a single isolated locus and the subsequent sequencing of fragment ends leads to clusters of forward strand tags 5′ of the binding site and reverse strand tags 3′ of the binding sites. The distance separating these clusters is dependent on the size distribution of sequenced fragments and the size of the local open chromatin region[76]. Algorithms that are designed to recognize the shape of ChIP-seq signal can be helpful in distinguishing chromatin and PCR induced effects from transcription factor binding events. Similarly in MNase-seq, well positioned nucleosomes are bracketed by 5′ and 3′ reads. Transcription factors or modified histones that bind across broad regions rather than at precise loci will, however, 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 amongst other biological factors[26]. Interpreting these read patterns can help us improve chromatin accessibility protocols and yield insight into ways in which chromatin is modified[52]. 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 Analysis

Although none of the DNase I footprinting algorithms developed so far explicitly take biases such as nucleosome occupancy, DNA sequence dependent cleavage, and transcription factor binding (which can affect the patterns of DNase I cleavage) into account, the way in which footprint significance is calculated and interpreted acknowledge bias effects to different degrees.

The first algorithms developed for DNase-seq footprint identification reduce sensitivity to sequence and other bias effects by ranking the read counts at each position in the central and flanking regions[8,88]. 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[89]. A recent method[9] estimates footprint significance based on 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, leading to high false discovery rates[26,90]. Analysis of DNase I cleavage patterns or evidence of transcription factor binding at a nucleotide resolution requires statistical modeling that accurately represents the intrinsic variability of DNase I cleavage.

Another way of distinguishing *bona fide* TF induced footprints from bias induced artifacts 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 transcription factor binding sites, and computes statistical significance based on the numbers of strand specific reads observed in a single flank relative to the footprint region[90].

While the occupancy of some transcription factors like CTCF is associated with DNase-seq footprint patterns, for many transcription factors these patterns are weak or, in cases like the androgen receptor, non-identifiable using current methods[26]. Transcription factors interact with chromatin in various ways resulting in diverse chromatin landscapes near transcription factor binding sites. Some TFs, like CTCF, bind in regions that are nucleosome free, flanked by well organized nucleosome arrays[91], while others bind in a such way that nucleosome occupancy is dependent on binding orientation[52]. Yet others, like the estrogen receptor, bind in a way that does not strongly depend on nucleosome occupancy[92]. CENTIPEDE[93] and more recently PIQ[52] analyze the shape and magnitude of DNase-seq profiles, together with TF position weight matrixes (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 remodeling.

Domain calling from ChIP-seq

ChIP-seq targeting certain histone modifications, including H3K9me3, H3K27me3 and H3K36me3, tends to produce diffuse regions of enrichment rather than the sharp peaks typically observed in TF ChIP-seq. These broad signals are challenging to analyze because the signal is diffuse and at times difficult to distinguish from the confounding effects of bias. 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 a variety of 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[94–96]. SICER[94] identifies broad intervals by first identifying bins with read counts above a predefined threshold, and subsequently computing a statistic for the aggregate of several such bins, possibly separated by small numbers of low read bins[94]. RSEG uses the Hidden Markov Model framework to specifically identify the boundaries of broad domains[95]. In this approach single sample read counts in genomic intervals are modeled using a negative binomial distribution, and the relationship between the read counts in an IP and an input sample is modeled using a difference of negative binomials distributions. Combinations of histone modifications are often observed together in chromatin states, patterns indicative of distinct modes of biological activity. These patterns may be identified by integrating multiple histone modification ChIP-seq data sets using the ChromHMM[97] or SegWay[98] 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[99], GPS[100] and PICS[85] de-convolute 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 while GPS can control for bias 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[101] 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 non-specific binding into account and allows shift distributions to be non-uniform across all binding sites.

Differential region identification

In a population of cells, transcription factor binding at a given locus may involve heterogeneous protein-DNA interaction strengths, with differing occupancies amongst cells and over time. TF binding is therefore better described by a continuous rather than binary variable as changes in binding can be as biologically relevant as the apparent loss or gain of binding sites. While 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 for detection of more subtle differences. Methodologies for identifying differential count enrichment, including DEseq[71,78] and edgeR[73,78], model count data in a way that is consistent with the counting process. These methods allow the use of offsets, parameters capturing artifacts[102] such as GC content, that are taken into account in the computation of differential enrichment. Such offsets can be computed using methods such as Conditional Quantile Normalization (CQN)[66]. It has been suggested that input controls be used to distinguish TF binding signal from background levels before comparisons can be made[70]. A procedure for comparative analysis of ChIP-seq peaks is carried out in DBChIP[70] making use of negative binomial modeling to estimate the over-dispersion of reads between samples. Comparisons of ChIP-seq 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[103]. Statistical techniques may be used to model effects arising from observable batch groupings such as date of sequencing[104]. Sometimes these effects cannot be associated with any particular batch annotation but may still be observed in clustering analyses that reveal clusters of samples that are inconsistent with any biological treatment groupings. Analyses such as surrogate variable analysis maybe used to mitigate these batch effects of unknown origin[105,106].

Chromatin Interaction Analysis

In Hi-C experiments, to quantify the interaction frequency between chromatin loci, pairs of DNA sequence fragments that are in close 3-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 modeled explicitly[107,108], an effective alternative perspective eliminates the need to explicitly account for these factors[109]. 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[109]. The bias identified in this way agrees to a remarkable degree with the bias detected through explicit modeling, adding confidence to both approaches. Hi-C interaction analysis in gigabase scale genomes like the human genome requires extremely high sequencing depths even for ~50kb scale resolution of interaction frequencies. Targeted approaches can be used to produce higher resolution interaction maps at selected genomic regions. Carbon Copy Chromatin Conformation Capture (5C) experiments[110] target specific regions of the genome using PCR primers, and ChIA-PET[12] uses chromatin immunoprecipitation to pull down loci that interact with particular proteins. In the analysis of data from 5C and ChIA-PET, bias and noise introduced in the selection step also need to be taken into consideration in the calculation of interaction frequencies.

Conclusion and Future Directions

NGS sequencing 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 phenotype. Investigators should be cautious in NGS data analysis to avoid interpreting biases and technical artifacts 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 the next. ChIP-seq studies of factors with good antibodies in cell lines are now ubiquitous and lists of several thousand binding sites can reliably be detected by several available algorithms. Challenges remain in the analysis of tissue and low cell number samples and 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. The interpretation of transcription factor occupancy in relation to chromatin accessibility profiling methods such as DNase-seq and ATAC-seq is an important emerging field. As the uses of NGS technologies and technologies themselves evolve, the detection and normalization of biases will require the development of effective and flexible methods implemented in efficient modular computational packages.

Acknowledgments

The authors thank members of the laboratories of Drs Shirley Liu and Myles Brown for their helpful discussions. This work is supported by the NIH grant R01GM099409.

Biographies

Clifford Meyer

Clifford Meyer received his PhD in Chemical Engineering from Princeton University, and is currently a Research Scientist in the Department of Biostatistics and Computational Biology at the Dana-Farber Cancer Institute and Harvard School of Public Health. His research focuses on understanding gene expression regulation through the computational analysis of genomic data.

X. Shirley Liu

X. Shirley Liu is Professor of Biostatistics and Computational Biology at the Dana-Farber Cancer Institute and Harvard School of Public Health. The Liu laboratory is developing algorithms and data integration approaches for high throughput data in transcriptional and epigenetic gene regulation.

Glossary

ChIP-seq	Chromatin immunoprecipitation (ChIP) followed by next-generation DNA sequencing (NGS) for the identification of DNA-associated protein binding sites
DNase-seq	DNase I digestion of chromatin combined with NGS for identifying regulatory regions of the genome, including enhancers and promoters
MNase-seq	MNase digestion of chromatin followed by NGS to identify loci of high nucleosome occupancy
Hi-C	Extension of chromosome conformation capture that uses NGS to observe long-range interaction frequencies between different regions of the genome
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing combines chromatin immunoprecipitation (ChIP)-based enrichment, chromatin proximity ligation, with paired-end NGS to determine chromatin interactions genome-wide
ATAC-seq	Assay to determine Transposase-Accessible Chromatin combining NGS with *in vitro* transposition of sequencing adapters into native chromatin
FAIRE-seq	Formaldehyde-Assisted Isolation of Regulatory Elements used for determining. regulatory regions of the genome
Spline	A flexible smooth nonlinear function defined piece-wise by polynomials for fitting nonlinear trends
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 is robust against outliers
Random barcoding	The ligation of a diverse assortment of short random DNA sequences to the unamplified DNA sample, a technique that can be used to distinguish duplicates produced by PCR from those originating from the unamplified DNA
Spike-in	Spike-in controls are known quantities of readily identifiable nucleic acids that are added to a sample prior to critical steps in an experimental protocol. Such controls may be used for bias assessment and calibration purposes
Surrogate variable analysis	Statistical analysis for the identification and modeling of variables that are not explicitly annotated yet have measureable effects

References

1. Barski A, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007; 129:823–37. [PubMed: 17512414]

2. Johnson D, Mortazavi A, Myers R, Wold B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science. 2007; (80):1497–1502. [PubMed: 17540862]
3. Mikkelsen TS, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007; 448:553–60. [PubMed: 17603471]
4. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. 2008; 26:1351–9. [PubMed: 19029915]
5. Schones DE, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell. 2008; 132:887–98. [PubMed: 18329373]
6. He HH, et al. Nucleosome dynamics define transcriptional enhancers. Nat Genet. 2010; 42:343–7. [PubMed: 20208536]
7. Boyle AP, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011; 21:456–64. [PubMed: 21106903]
8. Hesselberth JR, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. 2009; 6:283–289.
9. Neph S, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012; 489:83–90. [PubMed: 22955618]
10. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326:289–93. [PubMed: 19815776]
11. Dixon JR, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485:376–80. [PubMed: 22495300]
12. Fullwood MJ, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009; 462:58–64. [PubMed: 19890323]
13. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013:1–15.10.1038/nmeth.2688 [PubMed: 23547284]
14. Landt SG, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012; 22:1813–31. [PubMed: 22955991]
15. Teytelman L, et al. Impact of chromatin structures on DNA processing for genomic analyses. PLoS One. 2009; 4:e6700. [PubMed: 19693276]
16. Modak SP, Beard P. Nucleic Acids Research. 1980; 8:2665–2678. [PubMed: 6253888]
17. Zentner GE, Henikoff S. Surveying the epigenomic landscape, one base at a time. Genome Biol. 2012; 13:250. [PubMed: 23088423]
18. Telford DJ, Stewart BW. MICROCOCCAL NUCLEASE : ITS SPECIFICITY ANALYSIS AND USE FOR CHROMATIN. 1989; 21:127–137.
19. Henikoff JG, Belsky JA, Krassovsky K, Macalpine DM, Henikoff S. Epigenome characterization at single base-pair resolution. 2011
20. Tillo D, et al. High nucleosome occupancy is encoded at human regulatory sequences. PLoS One. 2010; 5:e9129. [PubMed: 20161746]
21. Valouev A, et al. Determinants of nucleosome organization in primary human cells. Nature. 2011; 474:516–20. [PubMed: 21602827]
22. Gaffney DJ, et al. Controls of nucleosome positioning in the human genome. PLoS Genet. 2012; 8:e1003036. [PubMed: 23166509]
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 U S A. 2010; 107:17945–50. [PubMed: 20921369]
24. Chung HR, et al. The effect of micrococcal nuclease digestion on nucleosome positioning data. PLoS One. 2010; 5:e15754. [PubMed: 21206756]
25. Campbell VW, Jackson DA. The Effect. 1980; 255:3726–3735.
26. He HH, et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nat Methods. 2013
27. Vierstra J, Wang H, John S, Sandstrom R, Stamatoyannopoulos Ja. Coupling a Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. Nat Methods. 2014; 11:66–72. [PubMed: 24185839]

28. Lazarovici A, et al. Probing DNA shape and methylation state on a genomic scale with DNase I. Proc Natl Acad Sci U S A. 2013; 110:6376–81. [PubMed: 23576721]
29. Grøntved L, et al. Rapid genome-scale mapping of chromatin accessibility in tissue. Epigenetics Chromatin. 2012; 5:10. [PubMed: 22734930]
30. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013:1–8. [PubMed: 23547284]
31. Van Heesch S, et al. Systematic biases in DNA copy number originate from isolation procedures. Genome Biol. 2013; 14:R33. [PubMed: 23618369]
32. Giresi PG, Lieb JD. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). Methods. 2009; 48:233–9. [PubMed: 19303047]
33. Gilfillan GD, et al. Limitations and possibilities of low cell number ChIP-seq. BMC Genomics. 2012; 13:645. [PubMed: 23171294]
34. 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. 2012; 52:87–94. [PubMed: 22313406]
35. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012; 40:e72. [PubMed: 22323520]
36. Wheeler TJ, et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. 2013; 41:D70–82. [PubMed: 23203985]
37. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008; 18:1851–8. [PubMed: 18714091]
38. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–60. [PubMed: 19451168]
39. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]
40. Alkan C, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009; 41:1061–7. [PubMed: 19718026]
41. Li R, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009; 25:1966–7. [PubMed: 19497933]
42. Derrien T, et al. Fast computation and applications of genome mappability. PLoS One. 2012; 7:e30377. [PubMed: 22276185]
43. Kunarso G, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010; 42:631–4. [PubMed: 20526341]
44. 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. 2011; 7:e1002111. [PubMed: 21779159]
45. Day DS, Luquette LJ, Park PJ, Kharchenko PV. Estimating enrichment of repetitive elements from high-throughput sequence data. Genome Biol. 2010; 11:R69. [PubMed: 20584328]
46. Wang T, et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc Natl Acad Sci U S A. 2007; 104:18613–8. [PubMed: 18003932]
47. Pickrell JK, Gaffney DJ, Gilad Y, Pritchard JK. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. Bioinformatics. 2011; 27:2144–6. [PubMed: 21690102]
48. Vogelstein B, et al. Cancer genome landscapes. Science. 2013; 339:1546–58. [PubMed: 23539594]
49. Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. Genome Biol. 2011; 12:R67. [PubMed: 21787385]
50. Degner JF, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics. 2009; 25:3207–12. [PubMed: 19808877]

51. Rozowsky J, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011; 7:522. [PubMed: 21811232]

52. Sherwood RI, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol. 2014

53. König J, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol. 2010; 17:909–15. [PubMed: 20601959]

54. Daley T, Smith AD. Predicting the molecular complexity of sequencing libraries. Nat Methods. 2013; 10:325–7. [PubMed: 23435259]

55. Marinov GK, Kundaje A, Park PJ, Wold BJ. Large-scale quality analysis of published ChIP-seq data. G3 (Bethesda). 2014; 4:209–23. [PubMed: 24347632]

56. Chen Y, et al. Systematic evaluation of factors influencing ChIP-seq fidelity. Nat Methods. 2012; 9:609–14. [PubMed: 22522655]

57. Ho JWK, et al. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. BMC Genomics. 2011; 12:134. [PubMed: 21356108]

58. Bonhoure N, et al. Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome Res. 201410.1101/gr.168260.113

59. Kidder BL, Hu G, Zhao K. ChIP-Seq: technical considerations for obtaining high-quality data. Nat Immunol. 2011; 12:918–22. [PubMed: 21934668]

60. Lassmann T, Hayashizaki Y, Daub CO. SAMStat: monitoring biases in next generation sequencing data. Bioinformatics. 2010; 27:130–131. [PubMed: 21088025]

61. DeLuca DS, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012; 28:1530–2. [PubMed: 22539670]

62. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinforma Oxford Engl. 2012; 28:2184–2185.

63. Planet E, Attolini CSO, Reina O, Flores O, Rossell D. htSeqTools: high-throughput sequencing quality control, processing and visualization in R. Bioinformatics. 2012; 28:589–90. [PubMed: 22199381]

64. Diaz A, Nellore A, Song JS. CHANCE: comprehensive software for quality control and validation of ChIP-seq data. Genome Biol. 2012; 13:R98. [PubMed: 23068444]

65. Heinz S, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38:576–89. [PubMed: 20513432]

66. Hansen KD, Irizarry Ra, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012; 13:204–16. [PubMed: 22285995]

67. Cleveland WS. Robust Locally and Smoothing Weighted Regression Scatterplots. 2013; 74:829–836.

68. Koenker R, Hallock KF. Quantile Regression. 2013; 15:143–156.

69. Rozowsky J, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009; 27:66–75. [PubMed: 19122651]

70. Liang K, Keles S. Detecting differential binding of transcription factors with ChIP-seq. Bioinformatics. 2012; 28:121–2. [PubMed: 22057161]

71. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11:R106. [PubMed: 20979621]

72. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010; 11:R25. [PubMed: 20196867]

73. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–40. [PubMed: 19910308]

74. Dillies MA, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2012

75. Shao Z, Zhang Y, Yuan GC, Orkin SH, Waxman DJ. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biol. 2012; 13:R16. [PubMed: 22424423]

76. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137. [PubMed: 18798982]

77. Hashimoto TB, Edwards MD, Gifford DK. Universal Count Correction for High-Throughput Sequencing. 2014; 10:14–18.

78. Anders S, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc. 2013; 8:1765–86. [PubMed: 23975260]

79. McVicker G, et al. Identification of Genetic Variants That Affect Histone Modifications in Human Cells. Science. 2013

80. Robertson G, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. 2007; 4:651–657.

81. Ji H, et al. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol. 2008; 26:1293–300. [PubMed: 18978777]

82. Nix DA, Courdy SJ, Boucher KM. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. 2008; 9:1–9.

83. Valouev A, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. 2008; 5:829–834.

84. Sun G, Chung D, Liang K. Statistical Analysis of ChIP-seq Data with MOSAiCS. 2013; 1038

85. Zhang X, et al. PICS: probabilistic inference for ChIP-seq. Biometrics. 2011; 67:151–63. [PubMed: 20528864]

86. Kornacker K, Rye MB, Håndstad T, Drabløs F. The Triform algorithm : improved sensitivity and specificity in ChIP-Seq peak finding. 2012

87. Kumar V, et al. Uniform, optimal signal processing of mapped deep-sequencing data. Nat Biotechnol. 2013; 31:615–22. [PubMed: 23770639]

88. Chen X, Hoffman MM, Bilmes Ja, Hesselberth JR, Noble WS. A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. Bioinformatics. 2010; 26:i334–42. [PubMed: 20529925]

89. Boyle AP, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. 2011:456–464.

90. Piper J, et al. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Res. 2013; 41:e201. [PubMed: 24071585]

91. Fu Y, Sinha M, Peterson CL, Weng Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet. 2008; 4:e1000138. [PubMed: 18654629]

92. He HH, et al. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. Genome Res. 2012; 22:1015–25. [PubMed: 22508765]

93. Pique-Regi R, et al. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2011; 21:447–55. [PubMed: 21106904]

94. Zang C, et al. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics. 2009; 25:1952–8. [PubMed: 19505939]

95. Song Q, Smith AD. Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics. 2011; 27:870–1. [PubMed: 21325299]

96. Wang J, Lunyak VV, Jordan IK. BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets. Bioinformatics. 2013; 29:492–3. [PubMed: 23300134]

97. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010; 28:817–25. [PubMed: 20657582]

98. Hoffman MM, et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012; 9:473–6. [PubMed: 22426492]

99. Lun DS, Sherrid A, Weiner B, Sherman DR, Galagan JE. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. 2009; 12:1–12.

100. Guo Y, et al. Discovering homotypic binding events at high spatial resolution. Bioinformatics. 2010; 26:3028–34. [PubMed: 20966006]

101. Chung D, et al. dPeak : High Resolution Identification of Transcription Factor Binding Sites from PET and SET ChIP-Seq Data. 2013; 9:9–11.

102. Li J, Jiang H, Wong WH. Modeling non-uniformity in short-read rates in RNA-Seq data. 2010:1–11.

103. Leek JT, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010; 11:733–9. [PubMed: 20838408]

104. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8:118–27. [PubMed: 16632515]

105. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007; 3:1724–35. [PubMed: 17907809]

106. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012; 28:882–3. [PubMed: 22257669]

107. Hu M, et al. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012; 28:3131–3. [PubMed: 23023982]

108. Hu M, et al. Bayesian inference of spatial organizations of chromosomes. PLoS Comput Biol. 2013; 9:e1002893. [PubMed: 23382666]

109. Imakaev M, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. 2012

110. Dostie J, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006; 16:1299–309. [PubMed: 16954542]

111. Degner JF, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012; 482:390–4. [PubMed: 22307276]

112. McVicker G, et al. Identification of Genetic Variants That Affect Histone Modifications in Human Cells. Science. 201310.1126/science.1242429

113. He HH, et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nat Meth. 2014; 11:73–78.

114. Zeng W, Mortazavi A. Technical considerations for functional sequencing assays. Nat Immunol. 2012; 13:802–7. [PubMed: 22910383]

115. Jung YL, et al. Impact of sequencing depth in ChIP-seq experiments. Nucleic Acids Res. 2014:1–10.

116. Landt SG, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012; 22:1813–31. [PubMed: 22955991]

117. Zhang Y, et al. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat Struct Mol Biol. 2009; 16:847–852. [PubMed: 19620965]

118. Bravo HC, Irizarry Ra. Model-based quality assessment and base-calling for second-generation sequencing data. Biometrics. 2010; 66:665–74. [PubMed: 19912177]

119. Pickrell JK, Gilad Y, Pritchard JK. Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”. Science. 2012; 335:1302. author reply 1302. [PubMed: 22422963]

120. Teytelman L, Thurtle DM, Rine J, Oudenaarden A Van. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. 2013:2–7.

121. Wang J, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 2012; 22:1798–812. [PubMed: 22955990]

122. Hansen KD, Irizarry Ra, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012; 13:204–16. [PubMed: 22285995]

123. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009; 10:669–80. [PubMed: 19736561]

124. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–72. [PubMed: 20220758]

125. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer New York USA HuberW. 2001; 18:764.

Online summary

- In NGS chromatin profiling experiments technical artifacts may be introduced at any stage, most importantly in fragmenting DNA, selecting the fragment population of interest, DNA amplification, DNA sequencing itself and read mapping to a reference genome.
- The effect of technical biases on experimental results will depend to a large extent on the genomic scale of the feature being analyzed and the scale on which the bias is manifest. Bias will have the greatest effect when the length scale of the bias is similar to the scale of the feature.
- Genomic experiments should be planned recognizing the potential confounding effects of bias and the limits of the technology. Proper controls to understand and characterize the potential bias in chromatin profiling should be included and sequenced to sufficient depth in such experiments.
- Nuclease induced fragmentation is usually biased by DNA sequence in ways that can produce patterns that might appear to have biological significance.
- Basic principles of statistical analysis should be applied to the analysis of chromatin profiling experiments: variability and bias should be accounted for and the fit of statistical models to observed data should be characterized.

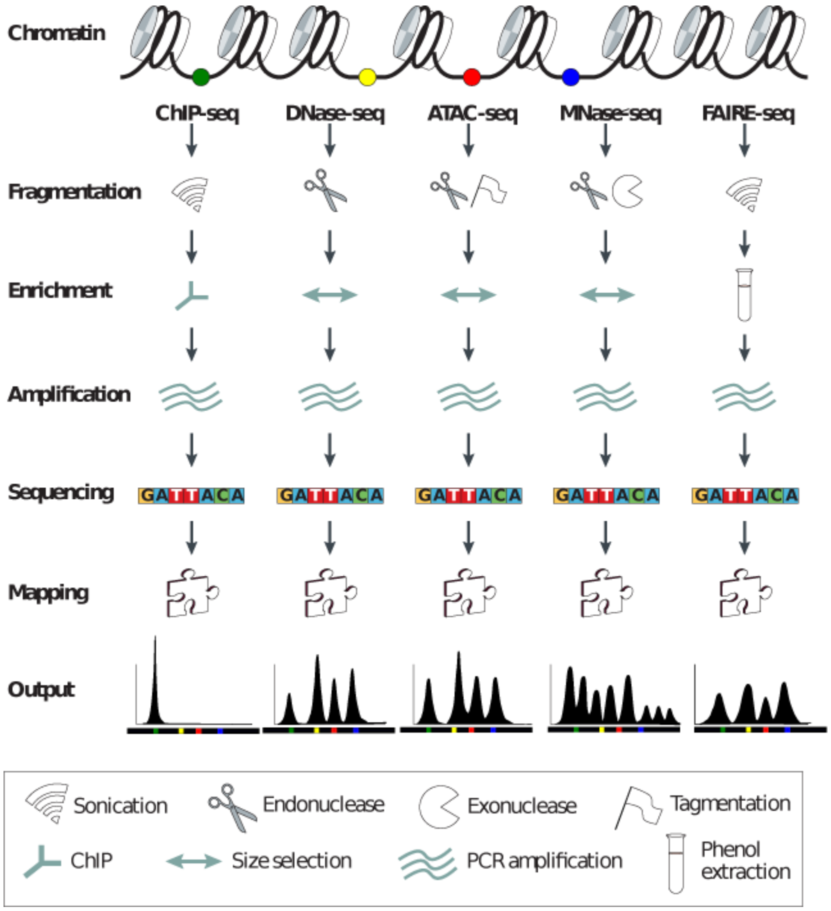

Figure 1. Overview of ChIP-seq, DNase-seq, ATAC-seq and MNase-seq experiments

A genomic locus analyzed by complementary chromatin profiling experiments reveals different facets of chromatin structure; ChIP-seq reveals binding sites of specific transcription factors, DNase-seq and ATAC-seq reveal regions of open chromatin while MNase-seq identifies well-positioned nucleosomes. In ChIP-seq chromatin immunoprecipitation (ChIP) is used to extract DNA fragments that are bound to the target protein, either directly or via other proteins in a complex containing 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 to DNase-seq that uses an engineered Tn5 transposase to cleave DNA and to integrate primer DNA sequences into the cleaved genomic DNA. Micrococcal nuclease (MNase) is an endo-exo- nuclease that processively digests DNA until an obstruction such as a nucleosome is reached.

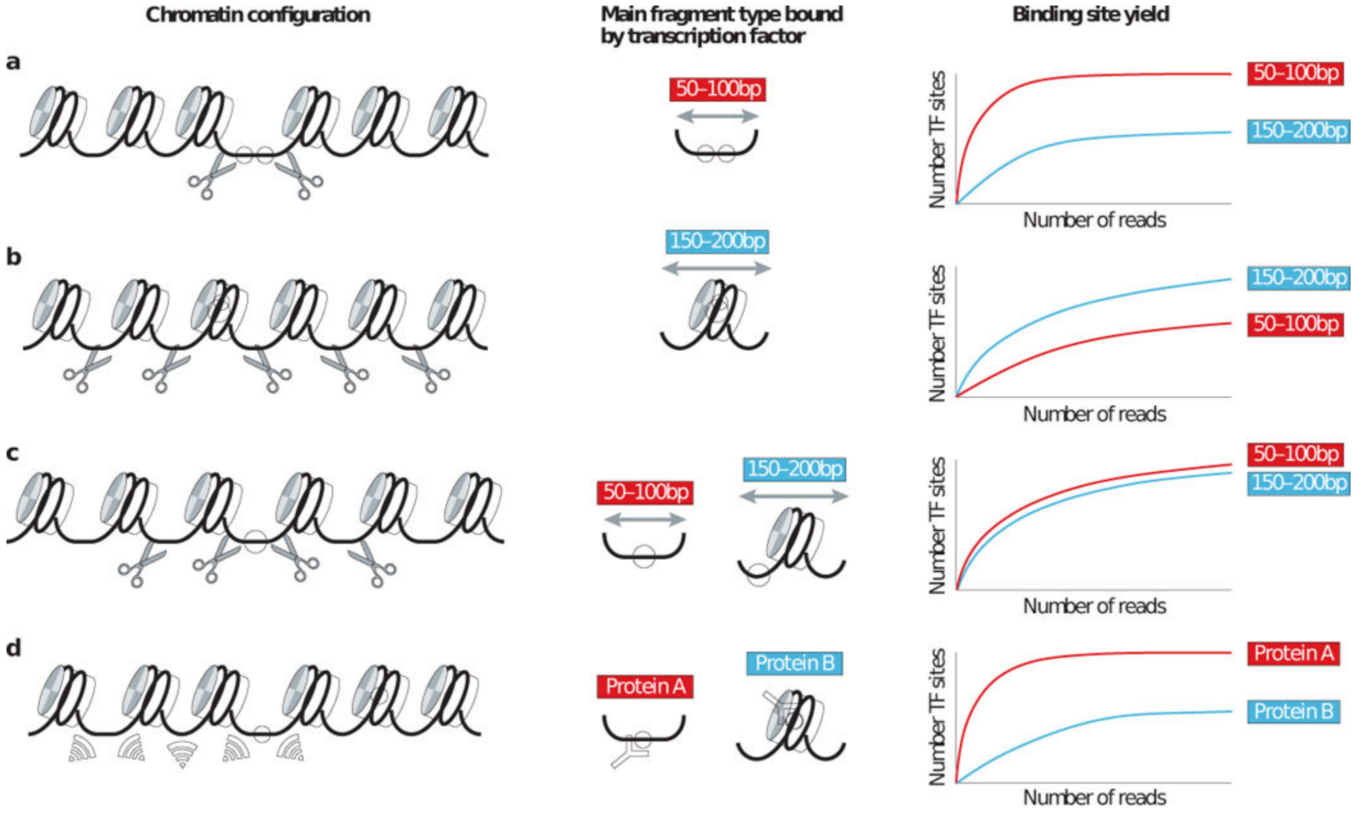

Figure 2. Fragmentation Effects in DNase-seq and ChIP-seq

Chromatin structure, fragmentation, and enrichment, interact to produce biased patterns of enrichment across the genome. **(a)** Some transcription factors, such as CTCF, typically bind in short nucleosome-depleted regions that are flanked by arrays of nucleosomes. When carrying out DNase-seq short fragments are far more efficient than longer ones for identifying such sites. **(b)** Histones and other factors that associate with 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 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. Like DNase-seq sonication is more efficient in regions of open chromatin. Factors bound in open chromatin contexts are more likely to be identified by ChIP-seq.

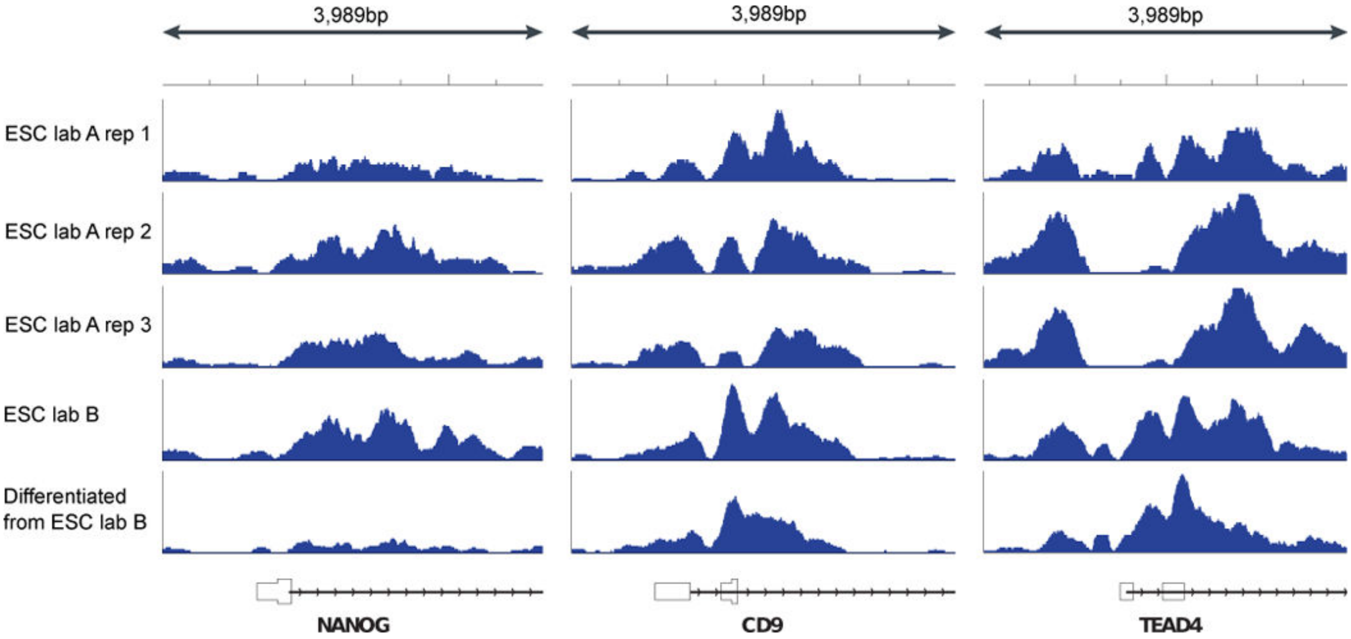

Figure 3. Variability of H3K4me3 ChIP-seq in human embryonic stem (ES) and differentiated cell lines

Several factors including fragmentation, immunoprecipitation conditions, and PCR biases can lead to different patterns of H3K4me3 enrichment at gene promoters in the same cell line. Coarse characteristics of H3K4me3 enrichment are consistent between samples, for example the depletion of H3K4me3 immediately upstream of the transcription start site of a core set of genes. Closer inspection reveals clear qualitative and quantitative differences between samples. For example, some samples show sharper peaks, perhaps due to differences in MNase digestion conditions and fragment selection. Regions that appear to be different between ES cells and differentiated cells in ChIP-seq samples produced by laboratory B also show variability in ES cell 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 way to produce data with comparable characteristics. Modeling bias can help reduce the amount of unexplained variability and increase sensitivity in detecting true differences between sample groups.

Table 1

Considerations in designing chromatin profiling NGS experiments.

Factor	Common Options	Considerations
Chromatin profiling assay	▪ ChIP-seq and antibody ▪ DNase-seq ▪ ATAC-seq ▪ MNase-seq ▪ MNase-ChIP-seq and antibody	▪ ChIP-seq requires a good specific antibody[14,59]. ▪ Differences in ChIP-seq data quality using different antibodies prevents all but the roughest comparisons of inter-antibody data sets. ▪ DNase-seq requires careful calibration of digestion conditions and fragment selection[26]. ▪ DNase-seq samples or ChIP-seq samples with the same antibody may be compared provided protocols are followed consistently, and bias effects and variability are taken into account[111,112]. ▪ ATAC-seq requires fewer cells, and less experimental calibration[30] but bias characteristics are not as well understood as DNase-seq bias[113]. ▪ MNase-ChIP-seq using antibodies associated with enhancers such as H3K4me and H3K4me2, or promoters 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 25-150bp ▪ Single-end ▪ Paired-end	▪ Read length is less important for chromatin profiling assays than studies of genomic or RNA transcript assembly[114]. ▪ Longer reads are suggested for studies that seek to identify allele specific chromatin events[50]. ▪ In highly specialized studies of chromatin, for example investigations of transposable elements[45], longer reads and paired-end reads would be useful in improving mappability[44,56]. ▪ Paired-end reads have three advantages over single-end reads: they increase the mappable proportion of the genome, they allow PCR duplicates to be more easily identified, and they allow the precise ends of fragments to be identified[26,56] ▪ Sequencing costs of longer reads and paired-end sequencing needs to be balanced against the value of more informative reads.
Read depth	▪ Multiplex ▪ Number of lanes ▪ Sequencing machine	▪ Multiplexing allows several samples to be sequenced in a single lane to a lower read depth[114]. ▪ Sequencing multiple biological replicates or sample replicates to a lower sequencing depth ispreferable to sequencing a single sample to a great depth. ▪ Information per read decreases as a function of read number. ChIP-seq that targets transcription factors that bind with high specificity reaches saturation at a lower read depth than more broadly bound histone modifications[115]. DNase-seq also requires greater sequencing depths with fragments longer than 147bp saturating at much higher levels than shorter reads[26]. ▪ Even at low sequence depth chromatin profiling should be informative for regions with strong signal and it is suggested that pilot studies at low coverage be carried out before sequencing deeper. ▪ It is important to examine library complexity in sequenced libraries as deep low complexity data sets can

Factor	Common Options	Considerations
		be less informative than shallower high complexity ones[54,116]. ▪ It is better to sequence a high quality sample at low depth than a low quality sample to a high depth. ▪ Sample QC can be carried out rapidly on Mi-seq.
Replicates	▪ Biological replicates ▪ Technical replicates starting from same biological material ▪ Sequencing replicates	▪ Many technical bias effects accumulate prior to library preparation and sequencing, therefore sequencing the same library multiple times is, in general, not informative. ▪ Biological replicates are essential to characterize variability between samples. ▪ Technical replicates starting from the same biological material can help understand the degree to which technical bias contributes to variability. ▪ When processing samples it is important to not process replicates of the same treatment condition in the same batch, as this would result in batch effects confounding treatment effects[103,104].
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 artifacts. These input controls ought to be sequenced to a greater depth than IP samples to obtain adequate coverage[56]. ▪ Input controls are preferred to IgG as they produce more complex libraries[116]. ▪ Conditions under which a transcription factor is not induced may be used as a control for ChIP-seq in the induced condition. Induction can, however, lead to chromatin state changes in places were the transcription factor binds, as well as elsewhere[92]. ▪ Spike-in controls have rarely been used in ChIP experiments so their value is 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[58].
DNase-seq and ATAC-seq and MNase controls	▪ Naked DNA ▪ Condition controls	▪ In DNase-seq or ATAC-seq footprinting studies as well as MNase nucleosome positioning studies naked DNA controls are useful for characterizing the DNA sequence bias of enzymatically induced cleavage[28,113,117]. To be informative such experiments need to be done at high levels of coverage. Although analysis of DNase-seq in chromatin are already highly informative for predicting bias effects[113] naked DNA data could provide additional information about sequence bias effects that are not considered in current models.

Table 2

Diagnosis and mitigation of bias in common analyses of chromatin profiling NGS experiments.

Analysis type	Examples	Biases that are likely to influence result	Diagnosis and mitigation
Allele specificity	▪ ChIP-seq, DNase-seq, ATAC-seq or MNase-seq read counts are associated with a SNP variant.	▪ 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 modeled on sequence characteristics and use error estimates to account for these error rates.[118] ▪ Check for association with the read rather than the genome, for example check if the allelic imbalance predominate at 5′ or 3′ end of reads[119]. ▪ Use special purpose mapping sofware[51,111,112]. ▪ Model nuclease induced cleavage bias or discard DNase-seq or ATAC-seq reads with 5′ ends close to the SNP[113].
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 using the same genomics technology[15,120,121]. ▪ Model effect of G/C or A/T DNA sequence content[122]. ▪ Examine whether spatial characteristics of read distributions look peak-like[123]. 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. ▪ Conduct analysis for different numbers of reads and examine trend of enrichment as a function of total read count[115].
Read enrichment relative to genomic feature	▪ Histone mark ChIP-seq read distributions relative to transcription start sites.	▪ Chromatin effects ▪ PCR amplification ▪ Ratio of background read counts relative to specific ChIP ▪ Read depth.	▪ Compare with controls and other data sets using the same genomics technology[9]. ▪ Examine quality control metrics related to specific versus nonspecific read quality[64]. If QC metrics differ substantially between samples repeat the experiment to get more consistent data quality[116]. ▪ Examine spatial distribution of G/C or A/T 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].

Analysis type	Examples	Biases that are likely to influence result	Diagnosis and mitigation
			▪ Carry out analysis using genomic control loci[26]. Exons, for example, tend to be GC-rich and are surrounded by less GC-rich sequence. Controls for exons might be intronic sequence with similar DNA sequence characteristics.
Differential abundance between conditions.	▪ In ChIP-seq, DNase-seq or ATAC-seq read level enrichment or depletion in treatment relative to control	▪ Batch effects. ▪ PCR amplification. ▪ 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[103,105,106]. ▪ Analyze dependence of fragment abundance on DNA sequence composition including GC content[35,122,124]. ▪ Include known quantitative factors in differential abundance analysis[103]. Batch variables such as date of sequencing are examples of such factors. ▪ Use unsupervised techniques such as surrogate variable analysis to remove systematic effects of unknown origin[106].
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[106]. Contrast data from general assays such as DNase-seq and ATAC-seq with ChIP-seq targeting specific protein.
Association of one biological phenomenon with another.	▪ Overlap of ChIP-seq peaks of two transcription factors. ▪ Claims of significant association between factor binding or differences in factor binding.	▪ Antibody quality. ▪ Relative immunoprecipiate enrichment. ▪ Chromatin effects. ▪ PCR amplification. ▪ Read depth.	▪ Check if common sources of technical bias underlie observations. ▪ Conduct analyses using different levels of read sampling. Sites with the strongest biological signal will be detected at a shallow read depth while weaker sites will be detected as the read depth increases[56]. ▪ Choose meaningful background models to discover associations. Transcription factor ChIP-seq peaks of different factors in the same cell line will often overlap significantly relative to a background of random genomic loci. Most transcription factor binding sites fall inside 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[125].

Analysis type	Examples	Biases that are likely to influence result	Diagnosis and mitigation
DNA motif analysis.	▪ Identification of transcription factor binding sites in ChIP-seq.	▪ Chromatin and fragmentation effects. ▪ PCR amplification. ▪ Nucleic acid isolation	▪ Evaluate bias and signal variability in controls[26]. ▪ Compare data with controls and data from other systems[121]. ▪ Evaluate results using independent data types.