ChIPulate : A comprehensive ChIP-seq simulation pipeline

Vishaka Datta*[1], Sridhar Hannenhalli[2], and Rahul Siddharthan[3]

[1]Simons Centre for the Study of Living Machines, National Centre for Biological Sciences, TIFR, Bengaluru 560065, India
[2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, 20742, USA
[3]The Institute of Mathematical Sciences/HBNI, Taramani, Chennai 600 113, India

February 12, 2019

Abstract

ChIP-seq (Chromatin Immunoprecipitation followed by sequencing) is a high-throughput technique to identify genomic regions that are bound in vivo by a particular protein, e.g., a transcription factor (TF). Biological factors, such as chromatin state, indirect and cooperative binding, as well as experimental factors, such as antibody quality, cross-linking, and PCR biases, are known to affect the outcome of ChIP-seq experiments. However, the relative impact of these factors on inferences made from ChIP-seq data is not entirely clear. Here, via a detailed ChIP-seq simulation pipeline, ChIPulate, we assess the impact of various biological and experimental sources of variation on several outcomes of a ChIP-seq experiment, viz., the recoverability of the TF binding motif, accuracy of TF-DNA binding detection, the sensitivity of inferred TF-DNA binding strength, and number of replicates needed to confidently infer binding strength. We find that the TF motif can be recovered despite poor and non-uniform extraction and PCR amplification efficiencies. The recovery of the motif is however affected to a larger extent by the fraction of sites that are either cooperatively or indirectly bound. Importantly, our simulations reveal that the number of ChIP-seq replicates needed to accurately measure in vivo occupancy at high-affinity sites is larger than the recommended community standards. Our results establish statistical limits on the accuracy of inferences of protein-DNA binding from ChIP-seq and suggest that increasing the mean extraction efficiency, rather than amplification efficiency, would better improve sensitivity. The source code and instructions for running ChIPulate can be found at `https://github.com/vishakad/chipulate`.

Introduction

ChIP-seq (Chromatin immunoprecipitation and sequencing) is a popular high-throughput experimental technique to find locations that are bound *in vivo* by a single transcription factor (TF)

*To whom correspondence should be addressed. Email: vishakadatta@gmail.com. The author list is alphabetical.

[1]. Upon mapping of the DNA fragments bound by the TF to the reference genome, the genomic loci bound by the TF are identified as high density mapped regions or *peaks*, where each peak is associated with an *intensity* based on the number of sequenced fragments arising from it. The intensity reflects the *in vivo* occupancy of the TF at that locus.

Several studies of ChIP-seq data have focussed on the biological factors distinguishing the loci bound by the TF. It has been shown that in addition to the affinities of binding sites present at a locus, nucleosome positioning is a strong determinant of TF binding *in vivo* [2, 3, 4, 5]. Other studies have shown that the concentration of the target TF [6, 7], short-range cooperative interactions between the target TF and other TFs [8], and variation in chromatin accessibility [5, 7] explain the variation in intensities across peaks. Some of the variation can arise due to indirect binding, where the target TF binds DNA indirectly via a second DNA-bound TF [9, 10, 11]. The intensity of such peaks is then no longer directly dependent on the affinity of the target TF to sequence at the bound locus.

Since the distribution of ChIP-seq peaks and their intensities depend on factors other than the affinity of the target TF towards sequence at a locus, it impacts two kinds of inferences often made from ChIP-seq data. First, the highest intensity peaks are used to infer position weight matrix (PWM) or motif models of the target TF [12]. While it is known that changes in the concentration of the target TF can distort the inferred PWM [13, 14], the extent to which cooperative and indirect interactions distort the PWM is unclear. Second, along with variations in chromatin accessibility, these interactions weaken the statistical dependence of peak intensity on target TF binding site affinity, which means that a peak with a higher intensity need not necessarily contain a higher affinity target TF binding site.

In addition to these biological factors, ChIP-seq peaks are affected by purely experimental sources of noise. The ChIP-seq protocol broadly consists of three key steps [1, 15, 16] — (i) the extraction of fragments that are bound by the target TF (ii) PCR amplification of these extracted fragments, and (iii) sequencing of these fragments. It is known that fragments can be extracted more easily at some genomic regions than others due to differences in chromatin structure and cross-linking efficiency across the genome [17, 18, 19]. Similarly, certain fragments are more efficiently amplified by PCR than others due to differences in GC content or the presence of nucleotide repeats [20, 21, 22]. The extent to which these factors distort peak intensities and make otherwise identical genomic regions appear differentially occupied has not been quantified. Multiple biological replicates of ChIP-seq are recommended [23] to overcome these issues, but the quantitative improvement in accuracy as more replicates are performed has not been ascertained.

To evaluate the influence of the aforementioned biological and experimental sources of variation on ChIP-seq peak intensities, we have developed a comprehensive pipeline to simulate a ChIP-seq experiment called ChIPulate. In particular, we simulate genome-wide TF-DNA binding based on established biophysical models of binding and also simulate the extraction, amplification and sequencing steps of the ChIP-seq protocol. Our simulation thus associates a set of read counts at each genomic locus with its occupancy by the target TF. Whereas existing statistical models [24, 25, 26, 14] assume that read counts follow a prescribed probability distribution and treat the ChIP-seq protocol as a black box, we explicitly model key steps of the ChIP-seq protocol. Our approach allows us to individually evaluate the impact of variation in extraction and PCR amplification efficiencies, as well as chromatin accessibility, indirect binding, and cooperative binding, on peak intensities and PWM inference. We are also able to quantify the extent to which additional replicates of ChIP-seq

improve its ability to robustly measure the occupancy of a genomic region.

We find that biological factors such as indirectly and cooperatively bound sequences distort inferred PWMs more than experimental sources of variation. Variations in extraction efficiency across the genome distort peak intensities and lower their ability to distinguish between occupied loci. Poor extraction efficiency also increases the number of false positive peaks, which are peak calls that do not contain a binding site for the target TF. In contrast to the effect of variations in extraction efficiency, even drastic variations in PCR amplification efficiency have little impact on peak intensities, and hence do not affect the inferred PWM or increase the number of false positive peaks. Finally, we found that at least two biological replicates of ChIP-seq read counts are necessary to reliably infer the binding energy of a genomic region.

Our work provides a general framework and a software tool for simulating ChIP-seq read counts through a realistic model of TF-DNA binding and the steps of the ChIP-seq protocol. Improvements in the protocol to lower variation in extraction efficiency, rather than PCR amplification efficiency, are more likely to improve our ability to distinguish between genomic regions of differing occupancy. Further, changes in the protocol that allow multiple biological ChIP-seq replicates to be performed, or computational approaches that can reliably combine read counts from ChIP-seq experiments, would allow more accurate inferences of TF-DNA occupancy, especially in regions containing multiple binding sites.

Results

A framework to simulate read counts from TF binding sites in a ChIP-seq experiment

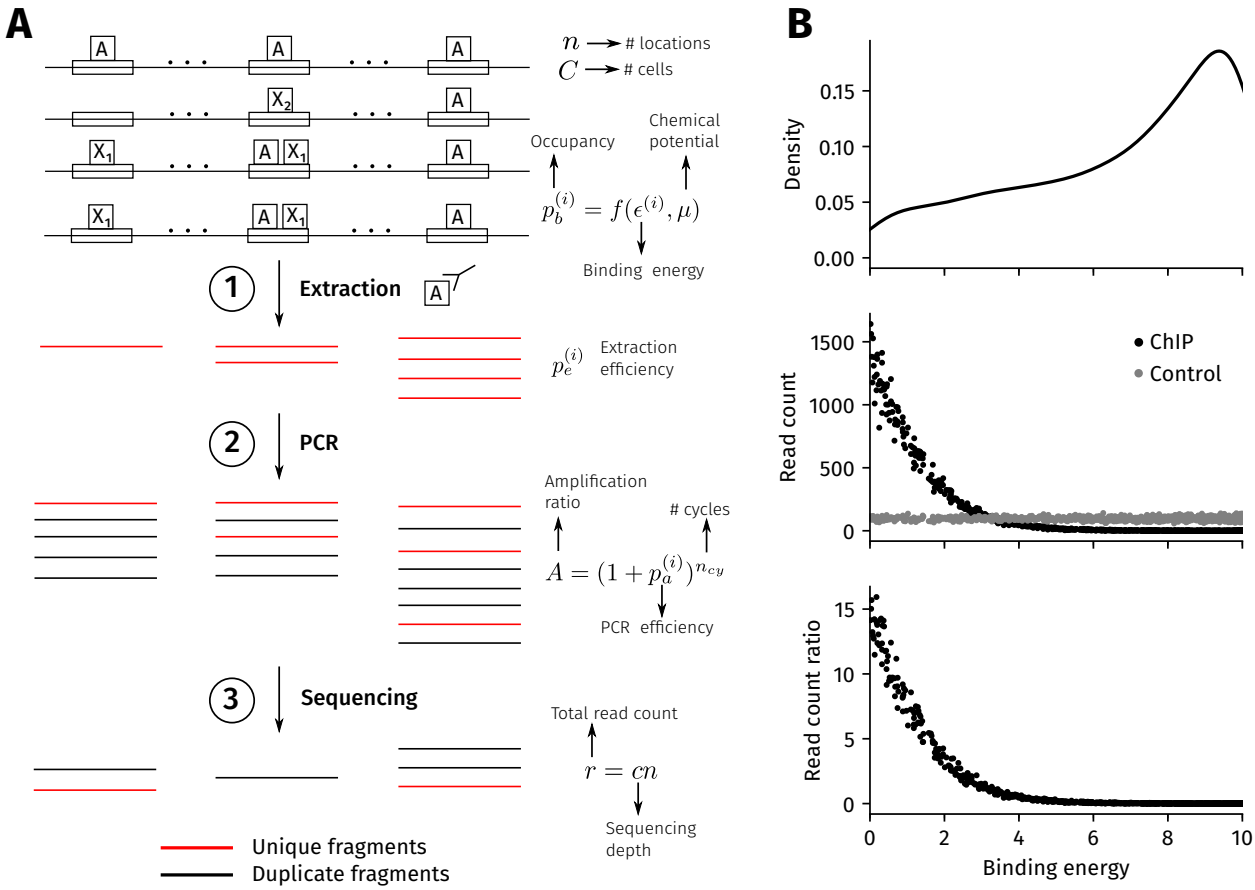

Figure 1: ChIP-seq simulation procedure. **(A) Procedure overview**. Each of n genomic regions contains a binding site with a binding energy $\epsilon^{(i)}$ (in units of k_BT), where a lower value represents a higher affinity binding site. In the ChIP sample, the binding energy at the i-th location, and the chemical potential, set the probability of occupancy ($p_b^{(i)}$) of the target TF. The probability of occupancy determines the number of cells (out of a total of C cells) where the i-th location is bound. In Step 1 of the simulation, a total of $F_e^{(i)}$ fragments are extracted at the i-th location from a binomial distribution with mean $Cp_e^{(i)}$. $p_e^{(i)}$ is the extraction efficiency that represents the probability of successfully extracting a bound fragment. All of these fragments are unique (in red) since each fragment originates from a different cell. In Step 2, the extracted fragments are amplified through n_{cy} cycles of PCR to give $F_a^{(i)}$ amplified fragments (duplicate fragments in black) with the probability of amplification of each fragment, or PCR efficiency, being $p_a^{(i)}$ in every cycle. The average number of amplified fragments A obtained from each extracted fragment is $A = (1 + p_a^{(i)})^{n_{cy}}$, which we refer to as the amplification ratio. In Step 3, r_i reads are obtained from the i-th location after sampling $r = cn$ fragments from all $\sum_{i=1}^{n} F_a^{(i)}$ amplified fragments, where r is the total read count and c is the sequencing depth. The number of unique reads at each location is the final output of the simulation. For the control input sample, read counts are simulated as for the foreground ChIP sample, except that the probability of occupancy is assumed to be constant at all n locations and the number of cells is assumed to be $0.1C$. **(B) Sample output from the simulation.** The top panel shows the distribution of binding energies employed in the simulation, which by default is a truncated power law between 0 and $10k_BT$ with an exponent of 0.5. The middle panel shows the number of unique reads from each genomic region in the ChIP (black) and input (gray) experiments. The bottom panel is the read count ratio between the ChIP and input samples at each location. The simulation parameters are $n = 1000$, $\mu = 3k_BT$, $\epsilon_{bg} = 1k_BT$, $C = 10^5$, $n_{cy} = 15$, and $A = 1000$. The extraction efficiencies in the ChIP and input samples follow a truncated normal distribution between 0 and 1, with a mean of 0.5 and standard deviation of 0.05 across the genome.

A schematic of our ChIP-seq simulation framework is shown in Figure 1A, with details in Methods. Broadly, the goal of our simulation is to take as an input the occupancy of a TF at multiple locations across a genome and output a set of read counts in both ChIP and input experiments at each location. A single TF's occupancy at a genomic location is determined by its site-specific binding energy, and its chemical potential, which depends on the logarithm of the concentration of the TF [27].

Each genomic location is associated with two experimental parameters — its extraction efficiency and PCR efficiency. The extraction efficiency, which determines how many bound fragments are successfully extracted in Step 1 in Figure 1A, can vary between different genomic locations. The extracted fragments are then subjected to a number of PCR amplification cycles in Step 2. The average number of amplified fragments obtained at each location depends on the PCR efficiency associated with the locus, which can also vary between different locations. Step 3 represents the sequencing step where the amplified fragments from all locations are sequenced. The number of sequenced fragments is the total read count of the experiment. Since unique and duplicate fragments (red and black in Fig. 1A) are tracked through the simulation, the exact number of unique reads at each location is known.

In addition to the ChIP experiment, we simulate the sequencing of a genomic control input sample. This simulation differs from the ChIP experiment in two ways. First, the background binding energy at each occupied genomic locus is set to a fixed value of $3k_BT$ for all locations (see Methods for justification). Second, we set the number of cells to be only 10% of the number of cells employed in the ChIP simulation. A larger number of cells are used in the ChIP simulation since, in practice, DNA is less efficiently extracted in the ChIP sample due to the use of an antibody whereas no antibody is used in the input sample. Thus, to extract a given amount of DNA, fewer cells are employed in the input sample than in the ChIP sample. In the event that the number of extracted fragments in the input is more than the number of extracted fragments in the ChIP sample, we down-sample the input fragments such that the total number of extracted fragments is identical in both ChIP and input samples before the PCR amplification step. Similarly, we down-sample the ChIP fragments if the number of extracted fragments in the ChIP sample is more than than that in the input sample. At each locus, the extraction and PCR efficiencies in the input are kept identical to the values used in the ChIP simulation.

After simulating both ChIP and input experiments, we compute the ratio of the read counts obtained from the ChIP sample to that of the input sample at each genomic locus. The read count ratio at a location is equivalent to peak intensity computed by commonly used peak callers from ChIP-seq reads [24]. The read counts obtained from our simulations are shown in Figure 1B with the values of the simulation parameters set to the default values described in Methods.

Key assumptions

When simulating genome-wide TF-DNA binding before Step 1 in Figure 1A, we assume that there is a single TF and a single binding site in each genomic region. We relax this assumption in later sections when simulating indirect and cooperative binding. We assume that binding in different genomic locations are independent of each other.

The extraction efficiency parameter in Step 1 aggregates the effect of the many steps of fragmen-

tation, cross-linking, pull down and size selection that occur in a ChIP-seq experimental protocol. We show in Methods that this parameter takes all these steps into account when modeling the extraction of DNA fragments from cells. In the PCR amplification step, we assume the amplification process to be in the exponential phase of amplification [28] where increasing the number of PCR cycles exponentially increases the number of amplified fragments. We assume that any potential PCR mutational errors do not change the amplification ratio at any genomic location. In the sequencing step 3, we assume that the sequencing and mapping are error-free, i.e., reads are uniquely and correctly mapped, and they can be perfectly de-duplicated at each location.

Motif inference is not affected by extraction and PCR efficiency whereas fidelity at locations with low read count ratios is affected by a low extraction efficiency

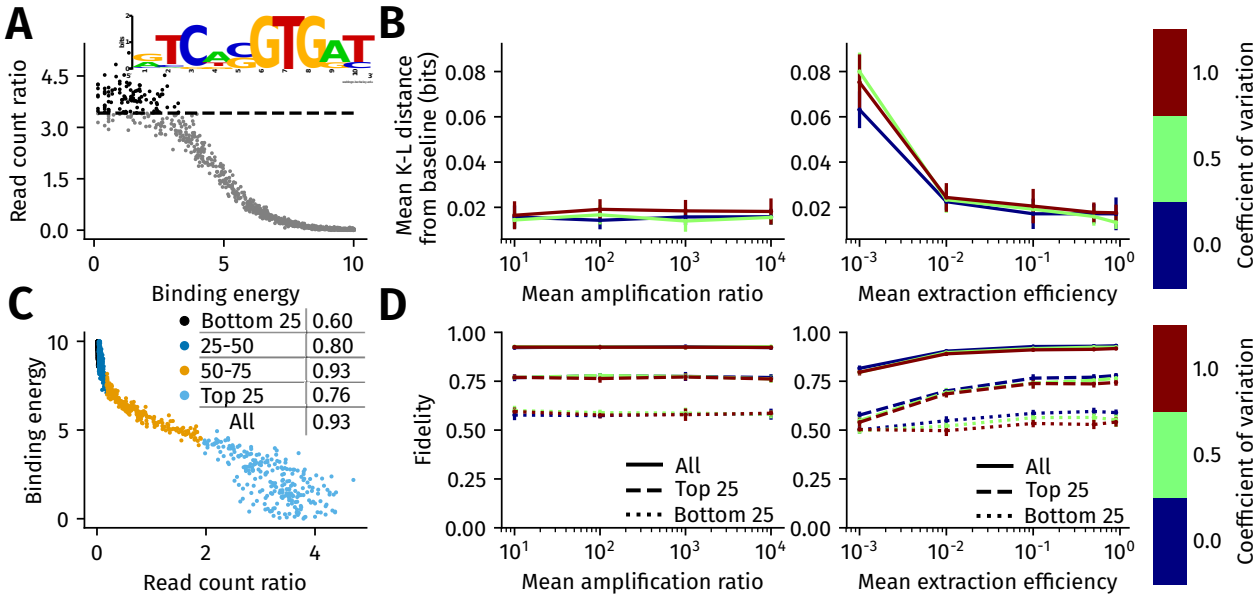

Figure 2: Impact of genome-wide heterogeneity in extraction and PCR efficiency on motif inference and ChIP-seq fidelity. (A) To simulate motif inference, 1000 binding energies were sampled from the default binding energy distribution. A binding site sequence of transcription factor Tye7 was assigned to each binding energy (Methods). After simulating ChIP-seq in the absence of extraction and amplification ratio heterogeneity, binding sites from locations with the 100 highest read count ratios were used to construct a baseline PWM of Tye7 (shown inset). **(B) The mean K-L distance between the baseline PWM and the motif inferred in the presence of heterogeneity in extraction efficiency (right) and amplification ratio (left).** The heterogeneity in extraction and amplification ratio is assumed to follow a truncated normal distribution, with the mean increasing from left to right on the x-axis in both panels. The coefficient of variation of the truncated normal varies from 0 (no variation, in blue) to 0.5 (green) and 1.0 (brown). The error bars are the standard deviation in the mean K-L distance computed after PWM was estimated in 10 replicates of ChIP-seq for each mean and coefficient of variation. **(C) ChIP-seq fidelity captures the monotonicity of the relationship between read count ratio and binding energy.** Fidelity is defined as the probability that if a location i has a read count ratio at least 10% higher than location j, then it implies that i has a lower binding energy than j. Fidelity is calculated by sampling 1000 pairs of locations, where each pair could be from anywhere in the genome or top 25th, 25-50th,50-75th, or bottom 25th percentiles of the read count ratio. Read count ratios falling in each percentile bin are marked in different colors in the scatter plot, where the y-axis is plotted on a logarithmic scale. The fidelity values in each of these bins is shown in the plot legend, along with the fidelity computed across all regions. **(D) Variation in ChIP-seq fidelity with heterogeneity in extraction efficiency (right) and PCR amplification ratio (left).** The x-axis and the three plot colors are defined identically to (B). The error bars are the standard deviation in the estimate of fidelity, which are computed from 10 replicates of simulation for a given mean and coefficient of variation.

We evaluated the impact of heterogeneity in extraction and PCR efficiencies on two outcomes of ChIP-seq. One is the recoverability of TF motif based on the top genomic locations in terms of their read count ratios. The second is fidelity, defined as the probability (frequency) with which a site X with a read count ratio that is at least 10% higher than site Y in fact contains a sequence with a higher affinity (or lower binding energy).

We simulated the process of inferring the motif of an arbitrarily chosen *S. cerevisiae* transcription factor (Tye7) from ChIP-seq (Figure 2A). We assigned binding energies for 1000 genomic regions from our default binding energy distribution, which is a power law distribution (with an exponent of 0.5) between 0 and $10k_BT$ (see Methods). By using the Tye7 binding energy matrix estimated by BEEML from protein binding microarray measurements [29, 13], we found the 10 bp sequence whose binding energy was closest to this assigned value, and virtually planted it at each location. We simulated a ChIP-seq experiment with a fixed extraction and PCR efficiency across the genome, following which we selected sites having the top 100 read count ratios and constructed the PWM of Tye7. This PWM, which we refer to as the *baseline PWM*, is used to compute the extent to which heterogeneity in extraction and PCR efficiency changes the derived PWM.

Ideally, one would use a PWM of Tye7 published in a database as the baseline motif against which to make these comparisons. However, these motifs can differ between databases due to differences not only in the experimental assay used to measure TF-DNA binding but also the underlying binding energy distribution of the bound sequences. This means that the baseline Tye7 PWM derived from the simulation may differ from that present in the ScerTF database [30]. However, our goal is to specifically assess the effect of extraction and PCR amplification heterogeneity on motif recoverability. For this purpose, it suffices to compare the motif derived in the presence of heterogeneity with the baseline motif derived in the absence of heterogeneity.

Figure 2B shows the difference between the baseline PWM and the Tye7 PWM that is derived in the presence of extraction or PCR amplification heterogeneity, respectively, when the efficiencies follow a truncated normal distribution. This difference is measured in terms of the mean Kullback-Leibler (K-L) distance, measured in bits (Methods), between the two PWMs. The impact of heterogeneity in these parameters on the inferred motif is small as the highest K-L distance from the baseline motif is about 0.045 bits. The effect of PCR amplification and extraction heterogeneity on motif inference is more pronounced when the heterogeneity follows a power law distribution (Supplementary Figure 1A) than when it follows a truncated normal distribution. The distortion of the PWM is still relatively low, however, with the highest K-L distance from the baseline motif being less than 0.1 bits per position. We also evaluated the impact of varying the chemical potential between $1k_BT$ and $6k_BT$ on the inferred motif when the baseline motif is derived at $\mu = 3k_BT$. Examples of the effect of a different chemical potential on the read count ratios obtained across the genome are shown in Supplementary Figure 2A. We found that the inferred motif deviates up to 0.08 bits per position from the baseline motif (Supplementary Figure 2B). This is within the same order of magnitude as the effect of a low mean extraction efficiency on the inferred motif.

We checked the extent to which the trends in Figure 2B depended on our choice of TF. We selected 21 TFs that possessed different DNA binding domains [31] (Supplementary Table 1), and for each TF, we simulated motif inference when there is heterogeneity in extraction efficiency (Supplementary Figure 3) or the PCR amplification ratio (Supplementary Figure 4) across the genome. The binding energy distribution of each TF, along with all other parameters of ChIPulate, was kept at their default values. We found that motif inference in each of these TFs was affected in

a manner similar to Tye7, where a low mean extraction efficiency distorts the inferred motif more than a low mean amplification efficiency. However, we found that the extent of motif distortion due to extraction heterogeneity was significantly correlated ($R^2 = -0.67, p < 10^{-3}$) with the motif information content of each TF (Supplementary Figure 3A), while there was no such correlation in the presence of amplification heterogeneity (Supplementary Figure 4A, $R^2 = -0.36, p = 0.1$). Finally, we found a weak dependence of the extent of motif distortion on the length of the motif, where a low mean extraction efficiency distorted a 6 bp motif upto nearly 0.1 bits per position with a lower magnitude of distortion for longer motifs (Supplementary Figure 5A). There was no such dependence of motif distortion on motif length when the amplification ratio varied across the genome (Supplementary Figure 5B).

Results shown in Figure 2C show that ChIP-seq fidelity is generally higher for sites with intermediate read count ratios (50th percentile to 75th percentile) and deteriorates at lower and higher read count ratios. This is partly due to the value of the chemical potential of $3k_BT$ employed in our simulation. When the chemical potential is varied between $1k_BT$ and $6k_BT$, the fidelity amongst the top 25th percentile of read count ratios noticeably decreases and reaches close to 0.5 at $\mu = 6k_BT$ (Supplementary Figure 2C). The fidelity amongst read count ratios of the bottom 25%, however, increases over the same range of the chemical potential. This is due to the fact that the occupancy of locations whose binding energies are much lower than the chemical potential are close to 1 while locations with binding energies above the chemical potential are close to zero. As the chemical potential increases, the increase in occupancy at locations with high binding energies is sufficiently large for their read count ratios to differ from one another. On the other hand, the occupancies of locations with low binding energies rapidly tend to one, which results in similar read count ratios between sites that possess different binding energies.

Figure 2D shows the impact of heterogeneity in extraction efficiency and PCR amplification ratio on fidelity. The fidelity is highest when computed across all binding site pairs, than amongst pairs in the top 25% of read count ratios. This is because the expected difference between the binding energies of sites in each pair is larger when the pairs are selected from the entire range. Changes in the mean and variance in the amplification ratio have little impact on fidelity, both overall as well as in each read count ratio bin. The fidelity across all regions is also not substantially affected by the mean and variance in extraction efficiency, but the fidelity is markedly lowered for sites with the lowest read count ratios. When PCR amplification heterogeneity is power-law distributed, its impact on fidelity is still quite low (Supplementary Figure 1B). In contrast, when the extraction efficiency is power-law distributed across the genome, its impact on fidelity is more drastic than when it is normally distributed.

A low mean extraction efficiency increases the probability of false positive peak calls but the PCR efficiency has no impact

We tested the effect of heterogeneity in extraction and PCR efficiency on the ability of ChIP-seq to distinguish peak calls that contain a binding site for the target TF (true positive) from those peak calls that do not contain a binding site for the target TF (false positive). Figure 3A shows a simulation where genomic locations that do not harbor a binding site for the target TF give rise to sequence reads. We refer to such locations as false positive peaks. This differs from indirectly bound peaks, where the target TF is part of the complex of TFs that binds DNA. In practice, a false positive peak can arise due to poor antibody quality, where the antibody binds epitopes on

TFs other than the target TF. We assign binding energies to these false positives locations from a truncated power law such that the mean occupancy ratio between true and false positive sites are fixed while the variance in binding energies amongst false positive sites and true positive sites are equal. Figure 3B shows the receiver operating characteristic (ROC) curve for the simulation shown in Figure 3A. The ROC curve shows the change in the number of true positives and false positives when the read count ratio threshold at which a location is declared as a peak is changed. The area under the ROC curve (auROC) provides an overall measure of the accuracy of using ChIP-seq read count ratios in distinguishing between locations with target TF binding sites from those without these sites.

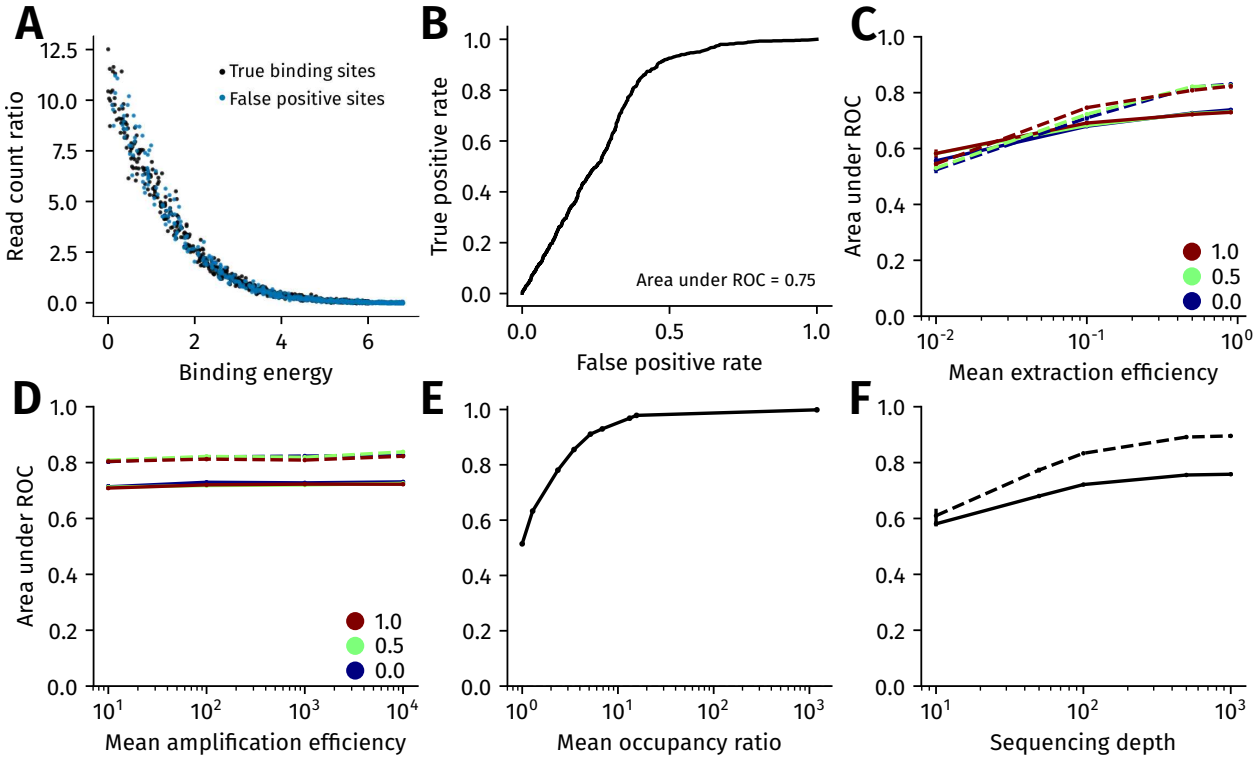

Figure 3: The impact of experimental and biological sources of variation on the sensitivity of ChIP-seq. (A) Simulating false positive binding sites. The binding energies of false positive genomic locations, which do not contain a target TF binding site, are distributed according to a truncated power law with exponent 0.76 in the range $[0, 6.78k_BT]$. Binding energies of true positive genomic locations, which contain a target TF binding site, are sampled from a truncated power law with exponent 0.5 in the range $[0, 6k_BT]$. **(B) Receiver operating characteristic (ROC) curve corresponding to the simulation shown in A. (C,D) Variation in auROC with the extraction efficiency and PCR amplification ratio.** The mean extraction efficiency in (C) and the mean amplification efficiency (after 15 cycles of PCR) in (D) increases along the x-axis. The efficiencies vary according to a truncated normal distribution with the blue, green and brown lines corresponding to a coefficient of variation of 0, 0.5 and 1.0, respectively. The solid and dashed lines are the auROC when the ratio of the mean occupancy of true positive binding sites to the mean occupancy of false positive binding sites is 2 (solid) and 10 (dashed) lines. **(E) Variation in auROC with the ratio of mean occupancy between true positive and false positive binding sites. (F) Variation in auROC with sequencing depth.** The ratio of the mean occupancy of true positive binding sites to the mean occupancy of false positive binding sites is set at 2 (solid line) and 10 (dashed line). In (C)-(F), the error bars are the standard deviation computed from ten replicates.

Figures 3C and 3D show the impact of heterogeneity in extraction efficiency and PCR efficiency, respectively, on the auROC of ChIP-seq read count ratios. We found that increases in the mean

PCR efficiency did not appreciably increase the auROC. Similarly, an increase in the heterogeneity of PCR efficiency, in terms of its coefficient of variation across the genome, did not decrease the auROC. In contrast, an increase in the mean extraction efficiency, and a decrease in its heterogeneity, led to a higher auROC. Thus, much like in the case of ChIP-seq fidelity, extraction efficiency has a greater impact on the auROC of ChIP-seq read count ratios than PCR efficiency.

Figure 3E shows that as expected, the auROC increases with the ratio of the mean occupancy of true and false positive binding sites. We illustrate in Methods that this mean occupancy ratio can be considered as a proxy for the specificity of the ChIP antibody used, where a higher mean occupancy ratio translates to a higher antibody specificity. A mean occupancy ratio of ~ 10 gives an auROC of around 0.9, which translates to a true positive rate of 0.8 at a false positive rate of 0.07. For a fixed mean occupancy ratio, an increase in the sequencing depth of both ChIP and input samples leads to an increase in the auROC (Figure 3F).

We further evaluated the impact of variation in chromatin accessibility across true and false positive binding sites. Our model of TF-DNA binding in the presence of chromatin assumes that TFs occupy a genomic location only when chromatin is accessible, i.e., the probability of occupancy depends on the energy of the binding site present and the probability that chromatin is accessible at that location (Equation (8)). We found that chromatin accessibility had a low impact on the auROC when the mean accessibility was at least 0.2 (Supplementary Figure 6). We note that equation (8) implicitly assumes that the probability of occupancy at a genomic location is zero in cells where chromatin is inaccessible. This does not hold for TFs such as pioneer TFs [32], which are capable of binding inaccessible chromatin and opening it up for further binding by other TFs. For these TFs, our model under-estimates the occupancy probability at a given genomic location, with the result that the auROC that we obtain in our simulation represents a lower bound.

Indirect binding affects motif inference and lowers sensitivity more than cooperative binding

The target TF (say, A) of a ChIP-seq experiment is said to be indirectly bound to DNA if it does not bind DNA directly but instead binds a second TF (say, B) that in turn directly binds DNA. Thus, an indirectly bound locus lacks a binding site for the target TF, and the number of reads mapped to this location do not depend on the binding energy of A. This can be seen in the top panel of Figure 4A, where the read count ratios of locations directly bound by A depend on the binding energy of A ($R^2 = -0.73, p < 10^{-115}$) whereas the read counts from indirectly bound regions do not ($R^2 = 0.004, p = 0.933$).

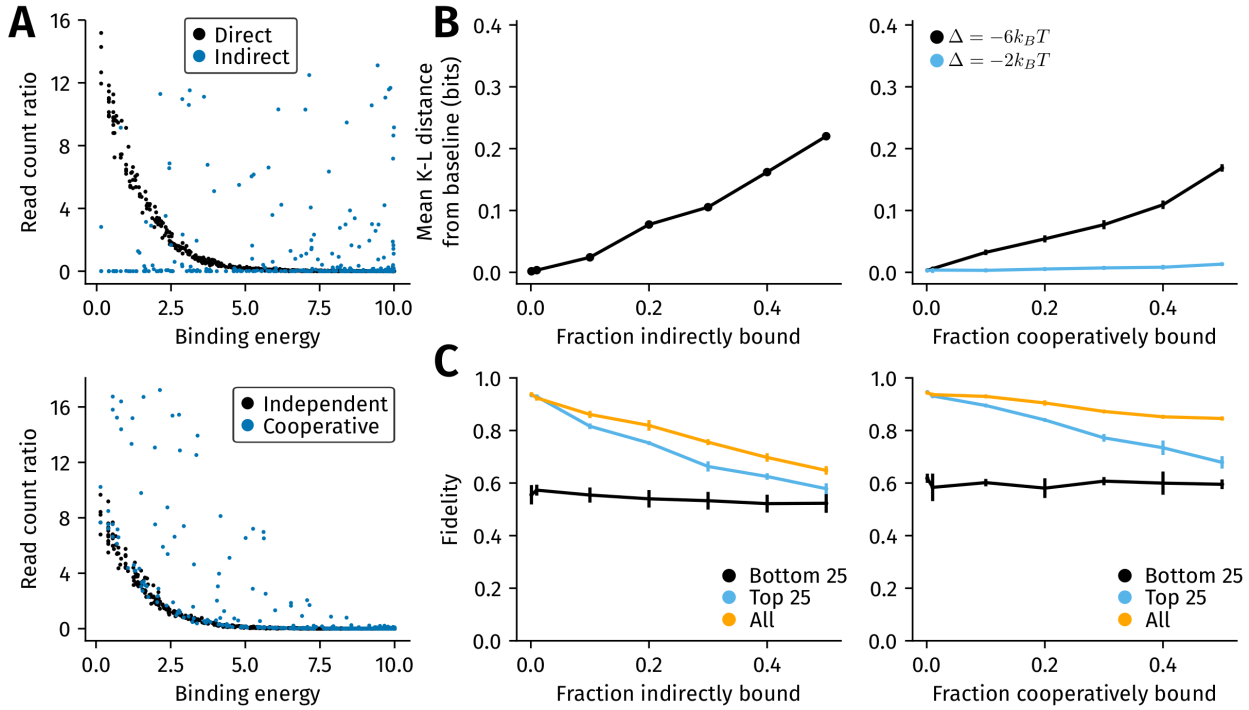

Figure 4: The relative impact of the inclusion of indirectly bound and cooperatively bound locations on motif inference and ChIP-seq fidelity. (A) Indirect binding and cooperative binding between TFs A and B are simulated as described in Methods. The scatter plots show the read counts from simulations where 30% of genomic locations are indirectly bound (top panel) or cooperatively bound (bottom panel, interaction energy (Δ) set to $-6k_BT$). In the top panel, the binding energy of A shown on the x-axis refers to the energy with which A would bind a location if it were in direct contact with DNA. The binding energies of A and B are sampled from a power law (with exponent 0.5) over a range of $[0, 10k_BT]$. **(B) Inclusion of indirectly bound locations [left] distorts the inferred motif more than the inclusion of cooperatively bound locations [right panel].** The change in the average K-L distance per base between the baseline motif and inferred motif (y-axis) is plotted against the fraction of indirectly bound regions [left] or cooperatively bound regions [right] on the x-axis. In the right panel, the cooperative interaction energy with the second TF is varied between $-6k_BT$ (black) and $-2k_BT$ (blue) (orange). The error bars are the standard deviation obtained from ten replicates of simulation. **(C) Indirect binding has a greater impact on ChIP-seq fidelity than cooperative binding.** The variation with fidelity and the fraction of indirectly bound [left] or cooperatively bound [right] locations is shown. The fidelity is calculated across all genomic locations (black), the top 25th percentile (blue) and the bottom 25th percentile (orange) of binding sites. The error bars are the standard deviation obtained from ten replicates of simulation.

In contrast to indirect binding, genomic locations that are cooperatively bound by the target TF and the second TF contain a binding site for each TF. The magnitude of the cooperative effect is represented by the interaction energy (Δ, in units of k_BT). Lower (negative) values of Δ represent a larger cooperative effect, with $\Delta = 0$ representing independent binding (see equation (4)). In the bottom panel of Figure 4A, read count ratios of some cooperatively bound locations ($\Delta = -6k_BT$) are higher than independently bound regions. Unlike in the case of indirect binding, however, read count ratios from cooperatively bound regions depend on the binding energy of the target TF ($R^2 = -0.66$, $p < 10^{-38}$), in addition to depending on the binding energy of the second TF and the interaction energy between both TFs.

In Figure 4B, we separately evaluated the impact of indirect and cooperative binding on the inferred PWM of the target TF. A detailed description of PWM inference in the presence of indirect and cooperative binding is given in Methods. The baseline PWM of the TF Tye7 is first computed

with direct binding alone. We then computed the mean K-L distance between this baseline PWM and the PWM inferred in the presence of either cooperative binding or indirect binding with a second TF. We found that the inclusion of a given fraction of indirectly bound sequences distorted the inferred PWM more than the inclusion of the same fraction of cooperatively bound sequences. The divergence of the inferred PWM from the baseline depended both on the binding energy range of the indirectly or cooperatively binding second TF, as well as the strength of the cooperative effect.

Similarly, we found that a given fraction of indirectly bound locations had a greater impact on ChIP-seq fidelity than the same fraction of cooperatively bound locations (Figure 4C). This was true even though the cooperativity is high ($\Delta = -6k_BT$). Indirect and cooperative binding particularly affected the fidelity of read count ratios from the strongest binding sites (the top 25th percentile of read count ratios).

More than two ChIP-seq replicates are needed to reliably estimate binding energies of low affinity sites from read counts

Next, we assessed the accuracy with which the binding energy at a locus can be estimated using ChIP-seq read count ratios, and the impact of replicates on this accuracy. If we denote the true binding energy of the i-th locus as $\epsilon^{(i)}$ and the estimate of the binding energy as $\overline{\epsilon}^{(i)}$, we expect $\overline{\epsilon}^{(i)}$ to move closer to $\epsilon^{(i)}$, on average, as more replicates of ChIP-seq are performed. The first step to quantifying the accuracy of $\overline{\epsilon}^{(i)}$ is to compute its posterior probability after having observed read count ratios $R^{(1)}, R^{(2)}, \ldots, R^{(n)}$ from n biological replicates, which is the conditional probability $P(\overline{\epsilon}^{(i)} = \epsilon | R^{(1)}, R^{(2)}, \ldots, R^{(n)})$. We define the absolute uncertainty or uncertainty of the binding energy estimate to be the 95% Bayesian credible interval of this posterior distribution [33], which we take to be the difference between its 97.5th and 2.5th quantiles. The posterior probability, and hence, the uncertainty of $\overline{\epsilon}^{(i)}$ after $n+1$ replicates are performed, can be computed from $P(\overline{\epsilon}^{(i)} = \epsilon | R^{(1)}, R^{(2)}, \ldots, R^{(n)})$ using Bayes' rule, as shown in Methods.

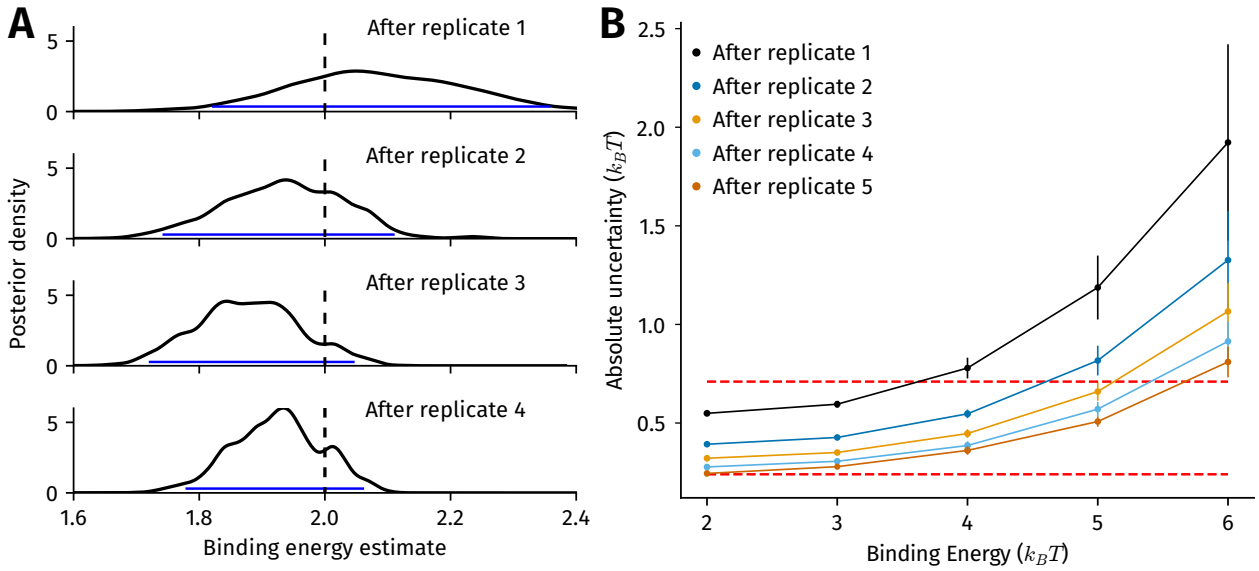

Figure 5: Impact of multiple biological replicates on errors in binding energy estimates. **(A) Posterior estimate of the binding energy at a single genomic locus after 1-4 replicates**. The true binding energy of the locus is $2k_BT$, indicated by the vertical dashed line. Four replicates of ChIP-seq were simulated, with the posterior density after each replicate, with the prior distribution set to be the default binding energy distribution employed in all our simulations (Methods). The blue lines indicate the absolute uncertainty of the binding energy estimate, which we define as the 95% Bayesian credible interval of the binding energy. This is calculated as the difference between the 2.5th and 97.5th quantiles of the posterior density. **(B) The reduction in absolute uncertainty of posterior binding energy estimates of five locations, each with a different binding energy, after three replicates of ChIP-seq.** The error bars are the standard deviation in estimates of the absolute uncertainty calculated after 100 trials; the filled circles represent the mean absolute uncertainty from these trials. The dashed red lines represent the values of absolute uncertainty required to distinguish between binding sites sequences that are a single mutation apart at least 50% of the time (upper,$0.71k_BT$) and at least 75% of the time (lower,$0.24k_BT$).

The expected decrease in the uncertainty of $\overline{\epsilon}^{(i)}$ as more replicates are performed can be seen in Figure 5A, where the area under posterior probability of $\overline{\epsilon}^{(i)}$ becomes more concentrated around $\epsilon^{(i)}$ with each additional replicate. We repeated this procedure to calculate the reduction in uncertainty at locations with binding energies between $2k_BT$ and $6k_BT$ in Figure 5B, and found that the uncertainty is high at sites with the lowest binding energies even after five replicates are performed. At the lowest binding energies, we also checked if the uncertainty after five replicates sufficed to distinguish between two binding sites that differ by one base pair from each other. From an analysis of binding energy matrices of 368 TFs in the BEEML database [13], we estimated that 75% of single mutations change the binding energy of a site by at least $0.24k_BT$, while 50% of single mutations change the energy by at least $0.71k_BT$. Thus, even three replicates of ChIP-seq would allow us to distinguish between two binding sites that are a single mutation apart with a probability between 50% and 75% at only the relatively high energy binding sites.

Simulating FASTQ reads using ChIPulate

In addition to simulating read counts from ChIP and control samples as described so far, ChIPulate can output single-end and paired-end reads for each genomic location if its chromosome coordinates are specified.

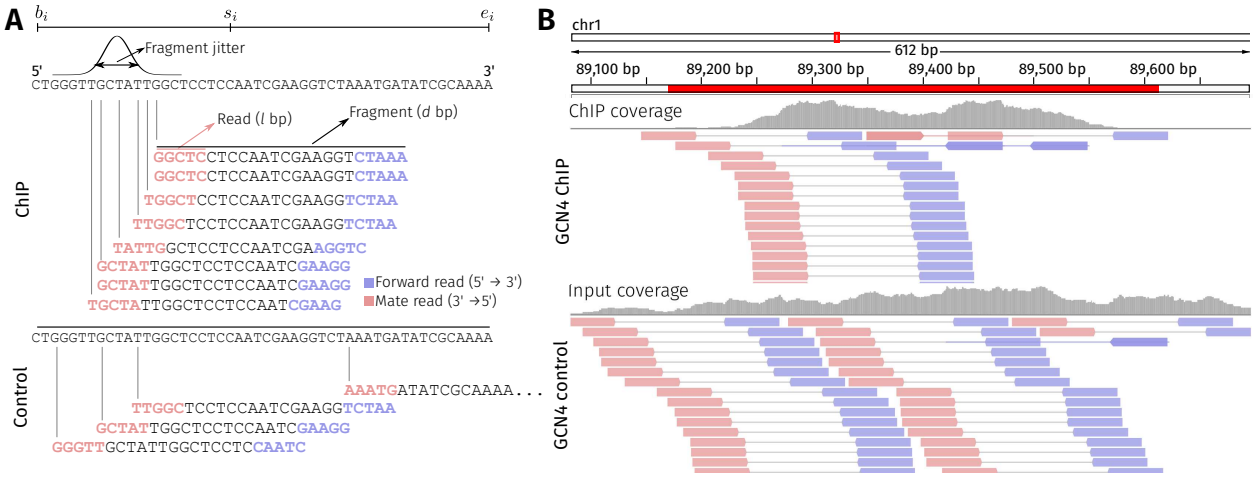

Figure 6: Example of simulating sequence reads using ChIPulate. (A) Illustration of ChIPulate's read generation method. The i-th genomic location is a genomic interval (b_i, e_i), with the "summit" of the interval located at s_i. The total read count (r_i) and unique read count (u_i) at the i-th interval in the ChIP and control samples are calculated as described in Figure 1. The starting positions of fragments in the ChIP sample are drawn from a Gaussian distribution with mean $s_i - d/2$ and standard deviation j, where d is the fragment length and j is the fragment jitter. In the control sample, fragment start positions are drawn from a Uniform$[b_i, e_i]$ distribution. The Phred-33 quality value of every base is set to be 75, which corresponds to a base-calling error probability of 10^{-42}. The read length is set to l bp, and both paired and single end reads can be simulated. **(B) Paired-end reads simulated from genomic intervals containing an experimentally determined GCN4 binding site.** The genomic intervals containing GCN4 binding sites were taken from an earlier publication [34]. Paired-end reads of 50 bp in length were simulated using ChIPulate, with the fragment jitter set at 50 bp and the fragment length set at 200 bp. The binding energy of the highest affinity GCN4 binding site in each interval was computed using the GCN4 binding energy matrix from the BEEML database. The remaining ChIPulate parameters were set at their default values.

Figure 6A illustrates the working of ChIPulate when it is run in FASTQ generation mode. To generate sequence reads, a set of n genomic intervals $(b_1, e_1), (b_2, e_2), \ldots, (b_n, e_n)$, along with a set of "summits" $s_1, s_2, \ldots, s_n$ is input to ChIPulate. These summits can be thought of as the physical location of TF-DNA binding in each interval. The total and unique read counts (r_i and u_i, respectively) at each interval are computed as described in Methods based on the binding energy (or energies) associated with each interval. Fragments that are PCR duplicates are tracked and the reads arising from duplicate fragments are named appropriately to reflect their status as a PCR duplicate. In the ChIP sample, the start positions of fragments in the i-th interval are sampled from a Gaussian distribution with mean $s_i - d/2$ and standard deviation j, where d is the fragment length and j is referred to as the fragment jitter. This captures the notion that a genomic position close to a TF-DNA binding event is more likely to give rise to a fragment in the ChIP sample than a genomic position that is far away. In the control sample, each position in an interval is assumed to be equally likely to give rise to a fragment. This corresponds to the Poisson background model assumed by peak callers such as MACS2 [24]. Thus, in the control sample, ChIPulate samples start positions in the i-th genomic interval from a Uniform(b_i, e_i) distribution. The l base pairs at the ends of each fragment in both ChIP and control samples represents the read length. Though Figure 6A illustrates the simulation of paired-end reads, ChIPulate can simulate both single-end reads and paired-end libraries, where the sequence of each read is assumed to be error-free.

Figure 6B is an IGV (Integrated Genomics Viewer) [35] snapshot of paired-end reads simulated from a genomic interval containing a GCN4 binding site in the *S. cerevisiae* genome. In this

example, we chose a set of 549 genomic regions that were found to be bound by GCN4 in an earlier publication [34], and computed the binding energy of GCN4 in each region using its binding energy matrix from the BEEML database. The starting position of the lowest energy binding site was set to be the summit of each region. We then generated the start positions of fragments (200 bp in length) in the ChIP sample by setting the fragment jitter to 50 bp. The read length for the paired-end reads were set at 50 bp at a sequencing depth of 100. The sequence of each read was assigned based on the sequence of the S288C R64-2-1 reference of the *S. cerevisiae* genome [36] from the *Saccharomyces* Genome Database [37]. The reads generated were aligned to this genome reference using the BWA aligner [38].

Discussion

ChIPulate substantially and qualitatively extends previous statistical models of read counts obtained in a ChIP-seq experiment. Early work by Zhang et. al assumed a distribution of binding affinities across the genome and computed occupancy using a similar biophysical model to ours [24], but did not simulate extraction and amplification. Bao et. al explicitly accounted for the effect of extraction efficiency on read counts [26], but did not model the underlying biophysical occupancy of genomic loci based on their binding energies. More recent work by Ruan et. al, accounted for the effect of chemical potential on TF-DNA occupancy and its impact on motif inference [14]. However, the authors evaluated the impact of experimental noise by adding Gaussian noise to the energies of binding sites rather than modeling the downstream processes of the ChIP-seq protocol.

Our pipeline combines a biophysical model of TF-DNA binding with a detailed simulation of fragment extraction, amplification and sequencing. This allows us to analyze the impact of heterogeneity in fragment extraction and PCR amplification on the motif inferred from bound genomic loci, and the fidelity of read count ratios in discriminating relatively close binding energies. We also evaluated how these factors, along with chromatin accessibility and sequencing depth, affect the false positive peak detection rate. The role of biological factors such as indirect binding and cooperative binding, which are frequent occurrences in ChIP-seq datasets, were also evaluated for their impact on motif inference and fidelity. Finally, we measured the accuracy of using read count ratios to infer the energy of a binding site and calculated the improvement in this accuracy after several biological replicates are performed.

In our simulations, fragment extraction encapsulates multiple steps in the ChIP-seq protocol, namely, cross-linking TF-DNA bound complex, antibody-mediated pull down of fragments bound by the TF, and the removal of cross-links. In addition to the mean efficiency, the nature of the extraction heterogeneity dictates the magnitude of its impact on motif inference and fidelity, with power law distributions having a greater impact than normally distributed heterogeneity. The effect of extraction heterogeneity on motif inference also depended on the structural class of the TF, where we found that a low mean extraction efficiency distorted motifs with a lower information content much more than those with high information content. However, for a given TF, it is unclear as to whether a normal or power law distribution better capture experimental variation in extraction efficiency across the genome. Previous reports suggest that variation in chromatin accessibility alone imposes a power-law distribution on extraction efficiency variation [18]. The development of alternate protocols that reduce the number of extraction steps [39, 40, 41] would improve the fidelity of read count ratios and reduce the number of false positive peak calls. In contrast to

extraction efficiency, variation in PCR amplification efficiency had a much lower impact on motif inference and the fidelity of read count ratios. However, in the presence of sequencing errors during amplification (which our model does not currently include) and imperfect mappability, a large mean amplification ratio in practice would increase false positive peak calls [42] and likely affects fidelity more than indicated by our simulations.

Though we explicitly modelled TF-DNA binding here, our conclusion on the impact of extraction and amplification efficiency on fidelity also likely holds for the ChIP-seq of histone marks. Though the occupancy of histones in nucleosomal complexes would have to be modelled in a very different manner to that of TFs [43, 44, 45, 46], this only affects the calculation of $p_b^{(i)}$. The remaining steps of extraction, PCR amplification and sequencing that are carried out in the ChIP-seq of a histone mark are identical to those that are simulated here. An issue of interest in a histone ChIP-seq dataset is the method employed to process the control sample [47, 48], where employing an antibody against the histone H3 in the control sample provides a better estimate of p_{ext} than an input sample [47]. Our simulations suggest that this would be true only if the mean extraction efficiency of fragments with these methods is not too much lower than using an input sample. This is because a low mean value of p_{ext} in the control sample would poorly sample any potential variation in extraction between genomic regions in the ChIP sample.

Extraction of DNA fragments due to indirect binding is believed to be frequent in ChIP-seq experiments [49, 10]. We find that the inclusion of indirectly bound regions has a more adverse effect on motif inference and fidelity than even strongly cooperatively bound regions. The extent of indirect binding in a dataset depends both on antibody quality [23, 11] and the cross-linking protocol employed [50] except when the ChIP-seq target protein is a co-factor which, by definition, is always indirectly bound. When antibody quality leads to indirect binding, we note that indirectly bound peaks and false positive peaks are hard to distinguish. Our work suggests that minimizing the extent of indirect binding and developing computational methods to detect them may provide a potent avenue to improving the fidelity of ChIP-seq experiments. For instance, protocols such as ChIP-exo [51, 11, 52] allow indirectly bound regions to be more easily detected through the peak shapes obtained after alignment and represent an important direction for further improvement of the ChIP-seq protocol.

Our results shows that several replicates of ChIP-seq are required if the energy of a binding site is to be inferred through read count ratios within a certain confidence interval. Since our simulation assumes a single site per peak, it should be noted that a single replicate of ChIP-seq can suffice to accurately infer a PWM for the TF [53], which in turn reliably estimates the energy of the binding site [54, 55, 56]. However, in the more general case where a peak may harbor multiple TF binding sites, the mapping between the binding energies of sites and the read count ratios in such a region is far more complex than in our simulation. This means that reliably inferring the occupancy at such a locus through read count ratios is likely to involve more replicates than we have estimated here. Since our simulations in this calculation assumed relatively low heterogeneity in extraction and PCR efficiency, further improvement in these steps of the protocol may not be the most effective avenue to reduce the number of replicates required to infer occupancy. Instead, reduction in sequencing cost to obtain multiple biological replicates of ChIP-seq will yield an accurate inference of *in vivo* occupancy at a genomic locus.

There is considerable scope for extending ChIPulate. We assumed that amplified fragments can be de-duplicated after sequencing in our simulation, which is possible with paired-end sequencing

and unique molecular identifier (UMI) based methods of PCR amplification [57, 58, 59]. Most ChIP-seq libraries are, however, subjected to single-end sequencing. It might be of interest to investigate the effect of imperfect de-duplication with single read sequencing. Furthermore, we did not take into account differences in mappability across the genome or sequencing errors, which can cause reads originating from a genomic locus to not map back to the same region [60]. These are relatively easy to account for in our simulation since it keeps track of individual fragments when they are extracted, amplified and sequenced. Finally, the FASTQ generation mode of ChIPulate can be used to comprehensively benchmark ChIP-seq peak callers and other computational pipelines that process ChIP-seq data.

Overall we have presented a detailed ChIP-seq simulation model and software pipeline that can be extended in various ways, including other bulk and single-cell sequencing protocols.

Acknowledgements

We thank Gautam Menon, Sandeep Krishna, Parul Singh and Aswin Sai Narain Seshasayee, for discussions. SH was funded in part by NSF award 1564785. VD was supported by the Simons Foundation. RS was supported by the PRISM 12th plan project at the Institute of Mathematical Sciences, funded by Department of Atomic Energy, Government of India.

Methods

ChIP-seq simulation framework

Genome-wide TF-DNA binding model

In the most basic version of our simulation, we consider the genome to be a set of n locations. Each of these locations contains a single binding site for the ChIP-seq target TF. We follow an approach similar to [27], where, at a given point in time, the i-th locus can be in one of two states — bound by the target TF or unbound. We assign a binding energy ϵ_u to the unbound state and a binding energy $\epsilon^{(i)}$ to the bound state. The Boltzmann weights of both these states are then $\exp(-\epsilon_u/k_BT)$ and $\exp(-(\epsilon^{(i)} - \mu)/k_BT)$, respectively. k_B is the Boltzmann constant T is the temperature at which the binding occurs, and μ is the chemical potential of the TF, which is proportional to the logarithm of the concentration of the TF [27]. The probability of finding the i-th locus in the bound state is then

$$p_b^{(i)} = \frac{\exp(-(\epsilon^{(i)} - \mu)/k_BT)}{Z}, \tag{1}$$

where $Z = \exp(-(\epsilon^{(i)} - \mu)/k_BT) + \exp(-\epsilon_u/k_BT)$. This can be re-written as

$$p_b^{(i)} = \frac{1}{1 + \exp((\epsilon^{(i)} - \epsilon_u - \mu)/k_BT)}, \tag{2}$$

Thus, the occupancy at a location is determined by the three quantities $\epsilon^{(i)}, \epsilon_u, \mu$. In the convention followed in this paper, the scale of $\epsilon^{(i)}$ is chosen such that the highest affinity binding site has a

binding energy of 0, with more positive values representing weaker binding sites. We set $\mu = 3k_BT$, which is within the range of values suggested by an earlier calculation in [27]. Finally, we set $\epsilon_u = 1.59k_BT$, which ensures that at the location with the highest affinity site i.e., when $\epsilon^{(i)} = 0$, the occupancy probability $p_b^{(i)}$ is 0.99.

$\epsilon^{(i)}$ can be thought of as a mismatch binding energy, where mutations that change the binding site sequence away from the highest affinity sequence increase its energy and thus lower the probability that it is occupied. This is in line with the convention followed in the binding energy matrices in the BEEML database.

To compute the probabilities of binding at all n locations, $p_b^{(1)}, p_b^{(2)}, \ldots, p_b^{(n)}$, using equation (1), we assign a binding energy to each locus. We sample binding energies from a power law that is truncated to a specified range. We define the probability density function f of this truncated power law, with parameters 1 ¿ $\alpha > 0, E_{max} > 0$, as

$$f(\epsilon) = \begin{cases} \frac{1-\alpha}{E_{max}^{1-\alpha}} \frac{1}{(E_{max}-\epsilon)^{\alpha}}, & 0 < \epsilon < E_{max} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Our choice of such a power law is motivated by the analysis of binding energies of sites that are occupied by the TFs CRP and FNR in the *E. coli* genome [61]. Unless stated otherwise, we set $\alpha = 0.5$ and $E_{max} = 10k_BT$. This corresponds to a situation where most of the binding sites in the genome have a low affinity (or high energy) for the target TF. $E_{max} = 10k_BT$ typically corresponds to a binding site that is 3-4 mutations away from the strongest binding site, since each mutation to the strongest sequence typically adds an energy $1 - 3k_BT$ to it for most TFs [27]. We set the chemical potential of the target TF, μ, to be $3k_BT$ by default in all our simulations in the main text.

In a population of C cells, when the i-th locus has a probability $p_b^{(i)}$ of being bound, then the number of bound fragments extracted from the i-th locus, denoted $F_b^{(i)}$, follows a Binomial$(C, p_b^{(i)})$ distribution, with the mean number of bound fragments is $Cp_b^{(i)}$.

Input sample: In most ChIP-seq experiments, a small number of cells are used to carry out a control, or input, experiment where fragments from cells are extracted without the use of antibody specific to the target TF. The read counts obtained in the input experiment are used to normalize, in a region or locus-specific manner, the read counts obtained from the ChIP sample. The input sample helps take into account heterogeneity that may arise in the extraction, amplification and sequencing of fragments from different genomic locations. The difference between the input and ChIP samples is in the number of cells employed and the probability of occupancy employed for each location. We assume that 10% of the number of cells used in the ChIP experiment are used in the input sample. The occupancy of genomic regions in the background sample consists of the non-specific binding of all the TFs in the cell, along with different sets of TFs that may be specifically bound at each genomic location. We phenomenologically model this scenario by assigning a fixed background binding energy $\bar{\epsilon}_{bg}$ to each of the n genomic locations, from which we use equation (1) to compute the probability of occupancy $\bar{p_b}$. We set $\mu = 0$ in equation (1) for the input sample and keep $\epsilon_u = 1.59k_BT$.

If the number of cells used in the input sample is C', then the number of bound fragments at each location follows a Binomial$(C', \bar{p_b})$ distribution. Unless stated otherwise, we set $\bar{\epsilon}_{bg}$ to $1k_BT$ in simulations where we do not model false positive binding sites in the ChIP sample. When

we simulate false positive binding sites, we set $\overline{\epsilon}_{bg}$ to a value where the ratio between the mean probability of occupancy in the ChIP sample to the mean probability of occupancy in the input sample is 10.

Aside from these differences in the TF-DNA binding model in the input and ChIP samples, the simulation procedures used in the description of the extraction, amplification and sequencing steps are common to both samples.

Cooperative binding

Consider TFs A and B, where A is the target of the ChIP-seq experiment, however, its binding depends on nearby binding of TF B. To simulate such cooperative binding between A and B, we have each of the n genomic locations in our simulation to contain a single binding site each for A and B, with binding energies $\epsilon_A^{(i)}$ and $\epsilon_B^{(i)}$ and an interaction energy Δ between them.

A given genomic locus can be in one of four states — unbound, bound by A, bound by B, or bound by both A and B. The probabilities of the locus being in each of these four states is [62]

$$\begin{aligned}
\text{P(unbound)} &= \exp(-\epsilon_u/k_BT)/Z, \\
\text{P(bound by A only)} &= \exp((-\epsilon_A^{(i)} + \mu_A)/k_BT)/Z, \\
\text{P(bound by B only)} &= \exp((-\epsilon_B^{(i)} + \mu_B)/k_BT)/Z, \\
\text{P(bound by A and B)} &= \exp((-\epsilon_A^{(i)} - \epsilon_B^{(i)} + \mu_A + \mu_B - \Delta)/k_BT)/Z,
\end{aligned}$$

where $Z = \exp(-\epsilon_u/k_BT) + \exp((-\epsilon_A^{(i)} + \mu_A)/k_BT) + \exp((-\epsilon_B^{(i)} + \mu_B)/k_BT) + \exp((-\epsilon_A^{(i)} - \epsilon_B^{(i)} - \Delta + \mu_A + \mu_B)/k_BT)$. $\Delta = 0$ corresponds to independent binding between A and B, $\Delta < 0$ corresponds to cooperative binding between A and B and $\Delta > 0$ represents competitive binding between A and B.

Since the ChIP experiment involves extracting fragments that are bound by A, the probability of a location being bound by A is the sum of the probability of the location being bound only by A and the probability of being bound by A and B. Thus, the occupancy probability of the i-th location, $p_b^{(i)}$, as seen in the ChIP-seq of A is :

$$p_b^{(i)} = \exp((-\epsilon_A^{(i)} + \mu_A)/k_BT) + \exp((-\epsilon_A^{(i)} - \epsilon_B^{(i)} - \Delta + \mu_A + \mu_B)/k_BT)/Z \tag{4}$$

We assume here that the binding site of A is not occupied by B, and vice versa. Further, we assume that there is no cooperativity between two distinct genomic locations i and j. We do not take the chemical potential of individual TFs into account, and thus implicitly assume that the concentrations of both TFs are similar.

Indirect binding

Consider TFs A and B, where A is the target of the ChIP-seq experiment. We say that A indirectly binds DNA if A binds to B which in turn binds DNA. To simulate indirect binding, we assume that

the probability of occupancy $p_b^{(i)}$ is proportional to the binding energy of B at the i-th location, $\epsilon_B^{(i)}$ as

$$p_b^{(i)} = \frac{1}{1 + \exp\left((\epsilon_B^{(i)} - \epsilon_u - \mu_B)/k_B T\right)}. \tag{5}$$

On the other hand, at locations directly bound by A, $p_b^{(i)}$ is dependent on $\epsilon_A^{(i)}$ according to equation (1). We set the chemical potential of A and B to zero and assume that B cannot occupy binding sites of A, and *vice versa*.

False positive binding sites

We consider a detected binding site to be false positive if it is bound by a TF (or TFs) other than the target TF of a ChIP-seq, or not bound by any TF at all, but fragments from which are nonetheless extracted in Step 2 of simulation. This differs from indirect binding, where the target TF is part of the complex of TFs that is bound to DNA. False positives reflect antibody quality or specificity.

In our simulations of false positive binding sites in Figure 3, we set the number of false positive binding sites to be equal to the number of true positive binding sites. We note that the area under the ROC curve is independent of the number of false positive binding sites simulated.

We set the binding energy distribution of true positive binding sites to be a power law between 0 and $6k_BT$ with $\alpha = 0.5$. This differs from the default power law distribution employed for true positive binding energies in our other simulations. This was because we found that binding sites with energies in the range $7 - 10k_BT$ had read count ratios that were very close to each other. Thus, when we computed the ROC curve with binding sites in this energy range, we saw that the true positive rate showed an abrupt jump as the threshold on the read count ratio was lowered below a point. Such jumps in the true positive rate make it difficult to compare ROC curves. By restricting the energy range of true positive binding sites to between 0 and $6k_BT$, we ensured that the true positive rate did not display this jump-like behaviour, which in turn allowed us to use the auROC as an accurate measure of detection performance.

We set the binding energies of the false positive sites to follow a power law distribution where the minimum binding site energy was $0k_BT$. The maximum binding energy (E_{max}^{fp}) and α_{fp} was set such that the ratio of the means of true and false positive binding energies could be fixed at a constant value while the variances were kept equal. Suppose the true positive binding energy distribution follows a power law over the energy range $[0, E_{max}^{tp}]$ and exponent α_{tp} such that its mean is m and variance is v. Then, setting α_{fp} and E_{max}^{fp} as

$$\begin{aligned} \alpha^{fp} &= 3 + 2/(r-2), \\ E_{max}^{fp} &= fm(2 - \alpha^{tp}), \end{aligned} \tag{6}$$

where $r = v/(g^2m^2) + 1$, sets the mean false positive binding energy to be fm while the variance remains at v.

Once the true and false positive binding energies are sampled, we compute the ratio between the mean occupancy of the true positive sites to that of the false positive binding sites. This is

calculated as

$$m_{occ} = \frac{\sum_{i=1}^{N_{tp}} p_{b,tp}^{(i)}/N_{tp}}{\sum_{j=1}^{N_{fp}} p_{b,fp}^{(j)}/N_{fp}} \tag{7}$$

where N_{fp} and N_{tp} are the number of false and true positive binding sites, respectively, and $p_{b,tp}^{(i)}$ and $p_{b,fp}^{(i)}$ are occupancies at true and false binding sites respectively.

In Figure 3, we set $\alpha_{fp} = 0.76$ and $E_{max}^{fp} = 6.78 k_B T$, which corresponds to a mean occupancy ratio of ~ 2. This is in line with the antibody recommendations of the ENCODE consortium [23], whose recommendation for antibody quality is that the primary reactive band should be at least 50% of the signal on the blot, which we interpret to represent a mean occupancy ratio between true positive and false positive sites in the ChIP sample to be at least 2.

Finally, we set the background binding energy, which determines the occupancy of fragments in the input sample, to a value of $4.96 k_B T$. This sets the ratio of the mean occupancy of true positive sites in the ChIP sample to the mean occupancy of the same sites in the input sample to 10. This is in line with the 5− to 13− fold-enrichment observed in read counts of ChIP-seq peaks between the ChIP and input samples in the ENCODE project [23].

Chromatin accessibility

We modeled chromatin accessibility using a sigmoidal prior based on DNAse-seq hypersensitivity data following [63, 3]. Briefly, the occupancy probability calculated based on the binding energies in equation (1) is multiplied by the sigmoidal function—

$$p_{acc} = \frac{1}{1 + \exp(-\beta d)}, \tag{8}$$

where d is the DNAse-seq read count density, which is average per-base DNASe-seq read coverage within a 150 bp window.

Thus, for a genomic locus whose occupancy probability is p_b, the probability that it is bound in Step 1 of our simulation is the product $p_b p_{acc}$. Thus, the occupancy probability essentially provides a scaling factor for binding probability.

Instead of simulating a profile of DNAse-seq read counts across the genome, we used read counts from DNAse-seq data of the *M. musculus* DBA/2 cell line (accession number ENCFF871YIT, lab of John Stamatoyannopoulos), available from the ENCODE Consortium [64, 65]. We used the filtered binary alignment map (BAM) files, which were aligned to the mm10 genome assembly using the ENCODE data processing pipeline. We chose the DNAseq-seq profile of this cell line as a reference DNAse-seq profile since it cleared all quality audit error categories laid out by the ENCODE Consortium for DNAse-seq data.

In order to consider only regions with accessible chromatin, we counted the number of reads falling into non-overlapping 100 kb windows of the genome and retained the top 10% of windows that had the highest read counts. We finally counted the reads falling into non-overlapping 150 bp windows and stored these read counts. In each ChIP-seq simulation, we uniformly sampled read

counts of n of these regions, where n is the number of binding locations being simulated, and plugged them into equation (8) (after dividing the counts by 150) to compute chromatin accessibility.

The extraction process

Given the number of bound fragments $F_b^{(1)}, \ldots, F_b^{(n)}$, each fragment needs to be extracted from the cell population. This extraction process involves several steps that include the lysing of cells to extract DNA, cross-linking of bound proteins to DNA, the size selection of sheared fragments, etc. After each of these steps, the number of bound fragments either stays the same or reduces. Our model of fragment loss during these steps tracks the reduction in fragment numbers prior to PCR amplification step.

We assume that the number of extracted fragments from the i-th genomic locus, denoted $F_e^{(i)}$, follows a Binomial$(F_b^{(i)}, p_e^{(i)})$ distribution. $p_e^{(i)}$ is the probability of a bound fragment from the i-th region being successfully extracted and present in the pre-PCR fragment pool. p_e can vary across the genome [66, 67] and is also dependent on the extraction procedure employed [68]. Unless stated otherwise, we assume that p_e follows a normal distribution truncated to lie between 0 and 1, with a mean of 0.5 and a standard deviation of 0.05.

Since we assume that $F_e^{(i)}$ is a binomial sample of $F_b^{(i)}$ fragments, the addition of more extraction steps into the simulation only changes the value of $p_e^{(i)}$ without necessarily violating the binomial assumption. Suppose we add an additional extraction step with efficiency e_i into the simulation such that $F_e^{(i)} \sim$ Binomial(X_i, e_i), where $X_i \sim$ Binomial$(F_b^{(i)}, p_e^{(i)})$. Then, by the laws of total variance and total expectation, the mean of $F_e^{(i)}$, for a fixed value of $F_b^{(i)}$, is $F_b^{(i)} p_e^{(i)} e_i$ and its variance is $F_b^{(i)} p_e^{(i)} e_i (1 - p_e^{(i)} e_i)$. By setting $p_i^* = p_e^{(i)} e_i$, we see that $F_e^{(i)}$ (conditioned on $F_b^{(i)}$) follows a Binomial$(F_b^{(i)}, p_i^*)$ distribution up to the first two moments. Thus, the parameter $p_e^{(i)}$ can be thought to implicitly take into account a variety of steps employed in the extraction process.

As stated in the previous section, we employ the same values of $p_e^{(i)}$ in both ChIP and input samples. In a single run of the simulation, it can happen that the number of fragments extracted in the ChIP sample exceeds that of the input sample, or vice versa. In practice, since fragments are typically more efficiently extracted in the input sample than in the ChIP sample, DNA extracted from the input library is diluted such that the number of fragments in both ChIP and input samples are equalized before amplification.

In our simulation, we compute a down-sampling factor $D = \sum_{i=1}^{n} F_e^{(i,\text{ChIP})} / \sum_{i=1}^{n} F_e^{(i,input)}$, which is the ratio of the total number of extracted fragments in the ChIP sample to that in the input sample. If $D < 1$, we down-sample extracted fragments in the input sample by replacing $F_e^{(i,input)}$ with the sample $X_i \sim$ Binomial$(F_e^{(i,input)}, D)$. If $D > 1$, we similarly down-sample the fragments from the ChIP sample instead of the input sample.

Mean occupancy ratio as a measure of antibody specificity

In equation (7), we defined the mean occupancy ratio as the ratio between the mean occupancy of true positive binding sites to that of false positive binding sites. This ratio can also be interpreted as a measure of antibody specificity in the following way.

Consider a genome with n true positive binding sites and m false positive binding sites. Suppose the extraction efficiency is $t_e^{(1)}, \ldots, t_e^{(n)}$ at true positive binding sites and $f_e^{(1)}, \ldots, f_e^{(m)}$ at false positive binding sites. Then, the average number of fragments extracted from the i-th true positive binding site in the ChIP sample is $Cp_b^{(i)} t_e^{(i)}$, and the average number extracted from the j-th false positive binding site is $Cp_b^{(j)} f_e^{(j)}$, where $\{p_b^{(i)}\}$ are occupancies at each location. The ratio (r_e) between the average number of fragments extracted from true positive binding sites to that from false positive binding sites is then

$$r_e = \frac{Cm \sum_{i=1}^{n} p_b^{(i)} t_e^{(i)}}{Cn \sum_{j=1}^{m} p_b^{(j)} f_e^{(j)}}.$$

This is the ratio that determines the accuracy of ChIP-seq in distinguishing between true and false positive binding sites. If r_e is large, true positive binding sites will give rise to more fragments than false positive binding sites, leading to true positive sites having larger read count ratios. A high value of r_e is achieved if, for $t_e^{(i)} = f_e^{(j)} = p_e$ across the genome, true positive sites have a larger mean occupancy than at false positive sites. Setting $t_e^{(i)} = f_e^{(j)} = p_e$ in the above equation reduces r_e to the mean occupancy ratio defined in equation (7).

Conversely, if $p_b^{(1)}, p_b^{(2)}, \ldots, p_b^{(n+m)} = p_b$ across the genome, then r_e reduces to the ratio

$$r_e = \frac{m \sum_{i=1}^{n} t_e^{(i)}}{n \sum_{j=1}^{m} f_e^{(j)}},$$

which is merely the ratio between the mean extraction efficiency at true positive binding sites to that at false positive binding sites. This constitutes a measure of the specificity of the antibody employed against the target TF of the ChIP-seq. Thus, ensuring that the average extraction efficiency at true positive sites is higher than at false positive sites ensures that the former give rise to higher read count ratios.

PCR amplification

We simulate PCR amplification using the model in [69] which we briefly explain below. In this model, each DNA fragment has a probability p_a of being amplified in a cycle of PCR and gives rise to two fragments. The fragment fails to undergo amplification and remains as a single fragment with probability $1 - p_a$.

Suppose $S_{n_{cy}}$ is a random variable that represents the number of amplified fragments after n_{cy} cycles of PCR. $S_{n_{cy}}$ can assume values between 1 and $2^{n_{cy}}$.

$$S_{n_{cy}} = S_{n_{cy}-1} + X, \tag{9}$$

where $X \sim \text{Binomial}(S_{n_{cy}-1}, p_a)$. If we represent $S_{n_{cy}-1}$ as a vector of length $2^{n_{cy}-1}$, the probability distribution of $S_{n_{cy}}$ can then be calculated from $S_{n_{cy}-1}$ in terms of an updating matrix $M^{(n_{cy})}$ (of dimension $2^{n_{cy}} \times 2^{n_{cy}-1}$) by setting $S_{n_{cy}} = M^{(n_{cy})} S_{n_{cy}-1}$ where

$$M_{i,j}^{(n_{cy})} = \binom{i}{j} p_a^j (1-p_a)^{i-j}. \tag{10}$$

Thus, for a given value of p_a, the distribution of $S_{n_{cy}}$ can be computed recursively from $S_{n_{cy}-1}, S_{n_{cy}-2}, \ldots, S_1$ when we start from a single DNA fragment. For computational efficiency, for values of n_{cy} between 10 and 15, we precomputed and stored the distributions of $S_{n_{cy}}$ for values of p_a between 0.01 and 0.99 in steps of 0.01. In order to simulate n_{cy} cycles of PCR amplification for $F_e^{(i)}$ fragments, we draw $F \equiv F_e^{(i)}$ samples $S_1^{(i)}, S_2^{(i)}, \ldots, S_F^{(i)}$ from the distribution $S_{n_{cy}}$ corresponding to the value of $p_a^{(i)}$ at that locus.

The mean and variance in the number of amplified fragments, $F_a^{(i)}$, starting with a single fragment at the first cycle and an amplification efficiency of p_a, can be calculated from branching process theory [70]. After a single cycle of PCR, the mean (m_1) and variance (v_1) are

$$\begin{aligned} m_1 &= (2p_a) + (1-p_a) = 1 + p_a \\ v_1 &= 1 + 3p_a - m_1^2 = p_a(1-p_a). \end{aligned}$$

After n_{cy} cycles of amplification, $m_{n_{cy}}$ and $v_{n_{cy}}$ are given by —

$$\begin{aligned} m_{n_{cy}} &= m_1^{n_{cy}} \\ v_{n_{cy}} &= \frac{v_{n_{cy}} m_1^{n_{cy}} (1 - m_1^{n_{cy}})}{m_1(1-m_1)} \quad (m_1 \neq 1) \end{aligned}$$

Substituting the expressions for m_1 and v_1 into the above equation, we have

$$\begin{aligned} m_{n_{cy}} &= (1+p_a)^{n_{cy}} \\ v_{n_{cy}} &= (1-p_a)(1+p_a)^{n_{cy}-1}\left((1+p_a)^{n_{cy}} - 1\right) \end{aligned} \tag{11}$$

We refer to $A \equiv m_{n_{cy}} = (1+p_a)^{n_{cy}}$ as amplification ratio of PCR at that location.

We assume that the PCR efficiencies at a locus remains constant across all cycles of amplification. We also assume that the number of amplified fragments obtained from each starting DNA fragment is statistically independent of amplifications at other fragments. This implies that starting with $F_e^{(i)}$ fragments, the mean number of amplified fragments after n_{cy} cycles of PCR is $AF_e^{(i)}$. The expressions in equation (11) are valid only when PCR has not yet reached a saturation of amplification where the concentration of PCR primers, polymerase, etc. relative to the concentration of the amplified fragments is enough to ensure that p_a does not decrease as n_{cy} is increased.

By default, we set $A = 1000$ and $n_{cy} = 15$ across the genome in all our simulations unless stated otherwise. This corresponds to a value of p_a that is ≈ 0.58.

Sequencing

At the end of the PCR amplification step, there are $\sum_{i=1}^{n} F_a^{(i)}$ amplified fragments from all n genomic locations. In this set of amplified fragments, there are $\sum_{i=1}^{n} F_e^{(i)}$ unique fragments, each

of which come from different cells, and the remaining $\sum_{i=1}^{n}\left(F_a^{(i)} - F_e^{(i)}\right)$ fragments are duplicates obtained through PCR. During the sequencing step, a total of $r = cn$ fragments are sampled from this amplified pool and sequenced. r denotes the total read count of the experiment and c is the sequencing depth.

Since each fragment has an equal probability of being sampled, the read count sample $r_1, r_2, \ldots, r_n$ from amplified fragment pools $F_a^{(1)}, F_a^{(2)}, \ldots, F_a^{(n)}$, where $r = \sum_{i=1}^{n} r_i$, follows a multivariate hyper-geometric distribution. To draw a sample from this distribution, we implemented the following procedure [71]—

Given two pools of fragments of pool sizes A and B from which r reads are to be sampled, the probability that k reads come from the first pool is

$$P(k) = \binom{A}{k}\binom{B}{r-k}\Big/\binom{A+B}{r}.$$

We extend this distribution to sampling from n pools of fragments, each of size $F_a^{(1)}, F_a^{(2)}, \ldots, F_a^{(n)}$ by first drawing r reads from two pools of size $A_1 = F_a^{(1)}$ and $B_1 = \sum_{i=2}^{n} F_a^{(i)}$. The probability of observing r_1 reads from the first locus (i.e. pool A_1) is

$$P(r_1) = \binom{A_1}{r_1}\binom{B_1}{r-r_1}\Big/\binom{A_1+B_1}{r}. \tag{12}$$

To draw reads for the second locus , we draw a sample of size $r - r_1$ from two pools of size $A_2 = F_a^{(2)}$ and $B_2 = \sum_{i=3}^{n} F_a^{(i)}$.

$$P(r_2) = \binom{A_2}{r_2}\binom{B_2}{r-r_1-r_2}\Big/\binom{A_2+B_2}{r-r_1}. \tag{13}$$

We continue this sampling process till all r reads have been sampled.

The read count at the i-th location, r_i, consists of a mixture of u_i unique reads and d_i duplicate reads such that $r_i = u_i + d_i$. We keep track of the fragments that were obtained via amplification versus those obtained in extraction step, and therefore, by simulation design, the u_i reads can be perfectly separated out from the d_i reads, which is detailed below.

As stated in the earlier section, $F \equiv F_e^{(i)}$ fragments extracted from the i-th location undergo PCR amplification to give $F_a^{(i)} = \sum_{j=1}^{F} S_j^{(i)}$ fragments, where $S_1^{(i)}, S_2^{(i)}, \ldots, S_F^{(i)}$ are the number of amplified fragments obtained from each of the $F_e^{(i)}$ fragments. To sample u_i we first draw samples $r_1^{(i)}, r_2^{(i)}, \ldots, r_F^{(i)}$ from the fragments pools $S_1^{(i)}, S_2^{(i)}, \ldots, S_F^{(i)}$ such that $r_i = \sum_{j=1}^{F} r_j^{(i)}$. We can write $r_j^{(i)} = u_j^{(i)} + d_j^{(i)}$, where $u_j^{(i)}$ and $d_j^{(i)}$ are the numbers of unique and duplicate fragments sampled from $S_j^{(i)}$. Since we assume that each of the $\sum_{j=1}^{F} S_j^{(i)}$ fragments has an equal probability of being sampled, $r_1^{(i)}, r_2^{(i)}, \ldots, r_F^{(i)}$ is a multivariate hyper-geometric sample of $S_1^{(i)}, S_2^{(i)}, \ldots, S_j^{(i)}$, which we generate using the same procedure outlined for sampling $r_1, r_2, \ldots, r_n$ reads from $F_a^{(1)}, F_a^{(2)}, \ldots, F_a^{(n)}$ fragments.

We then compute u_i for each location as follows. If $r_j^{(i)} = 1$, then $u_j^{(i)} = 1$ and $d_j^{(i)} = 0$. Since only one fragment in $S_j^{(i)}$ is unique, $u_j^{(i)}$ cannot exceed 1, which means that if $r_j^{(i)} > 1$, then, $u_j^{(i)} = 1$ and $d_j^{(i)} = r_j^{(i)} - 1$. We can then obtain u_i by setting $u_i = \sum_{i=1}^{F} u_j^{(i)}$.

In the process of simulating read counts, We do not account for mutations than can cause alignment errors in actual ChIP-seq reads, which in turn can change read counts at a genomic locus.

FASTQ read generation

To generate sequence reads, a set of genomic n intervals $(b_1, e_1), (b_2, e_2), \ldots, (b_n, e_n)$, along with a set of "summits" $s_1, s_2, \ldots, s_n$ need to be input to ChIPulate. These summits can be thought of as the physical location of TF-DNA binding in each interval. The total and unique read counts (r_i and u_i, respectively) at the i-th interval are computed as described above, based on the binding energy (or energies) associated with each interval.

In the ChIP sample, after r_i and u_i are computed, u_i fragment start positions $\{a_1^{(i)}, a_2^{(i)}, \ldots, a_{u_i}^{(i)}\}$ are sampled from a Gaussian distribution with mean $s_i - d/2$ and standard deviation j, where d is the fragment length and j is referred to as the fragment jitter. Suppose the copy number of each of the u_i fragments is $c_1, c_2, \ldots, c_{u_i}$, such that $c_1 + c_2 + \cdots + c_{u_i} = r_i$, where $c_i \geq 1$. Then, a set of r_i fragment start positions can be obtained by repeating the j-th unique fragment start position c_j times to give the set $\{a_1, \ldots, a_{c_1}, a_2, \ldots a_{c_2}, \ldots, a_{u_i}, \ldots, a_{c_i}\}$. The fragment jitter parameter j controls the dispersion of the fragment start positions from the position s_i.

The fragment start positions are simulated differently in the control sample. Here, given a unique read count u_i and total read count r_i at the i-th genomic interval, unique fragment start positions $\{a_1, a_2, \ldots, a_{u_i}\}$ are sampled from a Uniform(b_i, e_i) distribution. Given this set of unique fragment positions, the set of r_i fragment positions is constructed in the same way as in the ChIP sample, with the j-th unique fragment's starting position a_j being repeated c_j times.

Suppose the read length being simulated is l base pairs. At the end of the previous step, we have a total of $r = cn$ fragment start positions each in the ChIP and control samples, where c is the sequencing depth and r is the total read count. Suppose the the start position of the i-th read is a_i and the genome is a N letter sequence $g_1 g_2 \ldots g_N$, with the reverse complement sequence denoted $g'_1, g'_2 \ldots g'_N$. Then, if single-end sequencing is being simulated, we assign the sequence read to begin from the positive strand or the negative strand with equal probability. If the strand assigned to the read is positive, we assign the sequence $g_{a_i} g_{a_i+1} \ldots g_{a_i+l}$ to the i-th read. If the strand assigned to the read is negative, we assign the sequence $g'_{a_i+d-l+1} g'_{a_i+d-l} \ldots g'_{a_i+d}$. If paired-end sequencing is being simulated, we generate r read pairs, i.e., $2r$ reads in total, where each fragment gives rise to one read from the positive strand with the sequence $g_{a_i} g_{a_i+1} \ldots g_{a_i+l}$ and one read from the negative strand with the sequence $g'_{a_i+d-l+1} g'_{a_i+d-l} \ldots g' a_i + d$. ChIPulate uses the `getfasta` subcommand of `pybedtools` from the input genome FASTA file to extract these sequences [72, 73].

The quality of each base call in every sequence read is set to be K, the maximum possible value on the Illumina Phred-33 quality scale. This corresponds to a base call error probability of 10^{-42} [74].

Default values

Unless stated otherwise, we use the following values by default—
Genome-wide TF-DNA binding model: $n = 1000$, $\epsilon_1, \ldots, \epsilon_n$ are sampled from a power law (with $\alpha = 0.5$) between 0 and $10k_BT$, $\bar{\epsilon}_{bg} = 1k_BT$, $C = 10^5$(ChIP sample), $\mu = 3k_BT$, $C = 10^4$(input sample)
Extraction process: $p_e^{(1)}, \ldots, p_e^{(n)}$ are sampled from a $N(0.5, 0.025)$ distribution that is truncated to lie in $[0, 1]$. Further, at each location, the extraction efficiency is set to be the same in both ChIP and input samples.
Amplification process: $A = 1000$, $n_{cy} = 15$
Sequencing: $c = 100$

Motif estimation

To simulate motif estimation, we first sample binding energies $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ from a binding energy distribution. We associate each binding energy $\epsilon^{(i)}$ with a binding site sequence of the target TF whose energy ϵ'_i is closest to this assigned value. The sampled energy $\epsilon^{(i)}$ is then replaced with ϵ'_i, and the set of energies $\epsilon'_1, \epsilon'_2, \ldots, \epsilon'_n$ are used for all occupancy calculations in the TF-DNA binding simulation. In the simulations shown here, the energies ϵ'_i for each binding site sequence were taken from BEEML binding energy matrices (BEMs) that were fit to *in vitro* protein-binding microarray affinity measurements of different TFs [13].

Once ChIP-seq is simulated based on these binding energies, the top 10% of locations based on their read count ratios are chosen and the binding sites in each of them are used to construct the PWM of the target TF of the ChIP-seq. Since the BEMs we employed were for 10 bp long sequences, the PWMs we estimated were also 10 base pairs in length. Each term of the $\mathbf{W}$ was computed using the formula [75]

$$\mathbf{W}_{ij} = \log_2 \left(\frac{N_{ij} + b_i}{b_i(N+1)} \right), \tag{14}$$

where N_{ij} is the number of sequences with the i-th base at the j-th position, b_i is the background frequency of the i-th sequence and N is the number of sequences used to estimated $\mathbf{W}$. The background frequencies are added as pseudo-counts to deal with positions where N_{ij} is 0.

Motif estimation in the presence of indirect binding and cooperative binding

As stated in Methods, in indirect binding, a fraction of genomic locations are bound by the ChIP target TF A while the remaining locations are bound by a second TF B, which is in turn bound to A. The binding site sequences of A are assigned to each genomic locus based on the BEM of A but we do not utilize any BEM of B to assign a binding site sequence for it. Instead, we associate with each binding energy of B a 10 bp sequence generated from a dinucleotide model of the *S. cerevisiae* genome. We chose a dinucleotide model to generate indirectly bound sequences since this has been shown to closely approximate background genomic DNA that is not bound by the target TF [61].

After ChIP-seq is simulated in the presence of indirect binding, the top 10% of genomic locations

are chosen according to their read count ratio to estimate the PWM of A. The sequences used for PWM estimation are the binding sites of A from directly bound locations that fall in the top 10% and the randomly generated binding site sequences of B from the remaining indirectly bound locations.

In the case of cooperative binding, a fraction of locations is bound cooperatively by A and B while the remaining locations are independently bound by both TFs. At each genomic locus, binding sites of A and B are associated with the binding energies of each location based on their respective BEMs. When the top 10% of locations are chosen for PWM estimation, the binding sites of A from both cooperatively and independently bound locations are chosen.

Posterior density estimation

We drew samples from the posterior binding energy distribution $P(\bar{\epsilon}_1 = \epsilon | r^{(1)}, r^{(2)}, \ldots, r^{(k)})$ using a combination of kernel density estimation and the Metropolis-Hastings algorithm [76]. Here, $r^{(1)}, r^{(2)}, \ldots, r^{(k)}$ are read count ratios observed in k sequential replicates from a genomic locus whose true binding energy is ϵ_T.

We first compute the posterior distribution when $k = 1$. From Bayes' rule, we have

$$P(\bar{\epsilon}_1 = \epsilon | R^{(1)} = r^{(1)}) = P(R^{(1)} = r^{(1)} | \bar{\epsilon}_1 = \epsilon)\pi(\bar{\epsilon}_1 = \epsilon)/C_1, \tag{15}$$

where C_1 is the normalization constant $C_1 = \int P(R^{(1)} = r^{(1)} | \bar{\epsilon}_1 = \epsilon)\pi(\bar{\epsilon}_1 = \epsilon)d\epsilon$ and π is a prior distribution, which we set to be the default binding energy distribution.

$P(R^{(1)} = r^{(1)} | \bar{\epsilon}_1 = \epsilon) \equiv p(\epsilon_1)$ represents the probability density of read count ratios from a locus containing a single binding site with energy ϵ. To compute this distribution, we simulated 10^3 replicates of ChIP-seq across $n = 1000$ locations whose binding energies $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ were sampled from the default binding energy distribution. The remaining simulation parameters of the ChIP-seq were set to their default values. This gave us 10^3 replicates of read count ratios for each of the binding energies $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$. We used these read count ratios to compute Gaussian kernel density estimators (KDE) of $p(\epsilon_1), p(\epsilon_2), \ldots, p(\epsilon_n)$, which we denote as $\widehat{p}(\epsilon_1), \widehat{p}(\epsilon_2), \ldots, \widehat{p}(\epsilon_n)$. The KDE estimators were computed using the `gaussian_kde` method in the Python scipy library (v0.19.1) [77]. Given this set of KDE estimators, we can compute $\widehat{p}$ at any arbitrary ϵ in the range $[\min \epsilon^{(i)}, \max \epsilon^{(i)}]$ using a linear interpolation between values ϵ_j and ϵ_{j+1} as

$$\widehat{p}(\epsilon) = \lambda \widehat{p}(\epsilon_j) + (1 - \lambda)\widehat{p}(\epsilon_{j+1}), \tag{16}$$

where $\epsilon_j < \epsilon < \epsilon_{j+1}$ and $\lambda = (\epsilon - \epsilon_j)/(\epsilon_{j+1} - \epsilon_j)$. Substituting this into equation (15), we get

$$P(\bar{\epsilon}_1 = \epsilon | R^{(1)} = r^{(1)}) \approx \widehat{p}(\epsilon_1)\pi(\bar{\epsilon}_1 = \epsilon)/C_1, \tag{17}$$

where $C_1 = \int \widehat{p}(\epsilon_1)(r^{(1)})\pi(\bar{\epsilon}_1 = \epsilon)d\epsilon$. To sample from $P(\bar{\epsilon}_1 = \epsilon | R^{(1)} = r^{(1)})$, we implemented the Metropolis-Hastings algorithm [76], which is an algorithm that draws samples from a Markov chain whose stationary distribution is the posterior density from which we wish to sample.

Using the same ChIP-seq simulation parameters as used in generating the 10^3 replicates, but with the binding energy of the first locus ϵ_1 set to ϵ_T, we simulated k replicates of ChIP-seq and store the sequence of read count ratios $r^{(1)}, r^{(2)}, \ldots r^{(k)}$ from the first locus. These are the read count ratios that are used to sequentially update the posterior binding energy estimate in equation (15).

Metropolis-Hastings algorithm

The output of the Metropolis-Hastings algorithm is a set of samples $\epsilon_k^{(0)}, \epsilon_k^{(1)}, \ldots, \epsilon_k^{(m)}$ that follows a target distribution, which in this case is the posterior distribution $P(\bar{\epsilon}_1 = \epsilon | R^{(1)} = r^{(1)}, R^{(2)} = r^{(2)}, \ldots, R^{(k)} = r^{(k)})$. We explain the algorithm for the case when $k = 1$. .

We first sample ϵ from a Uniform$[0, 10]$ distribution and set $\epsilon_1^{(0)} = \epsilon$, and compute $f(\epsilon_0) \equiv P(\epsilon_0 | R^{(1)} = r^{(1)}) = \widehat{p}_\epsilon \pi(\epsilon^{(i)})$. $\epsilon_1^{(0)}$ is the initial state of the Markov chain. We then sample ϵ' from a proposal distribution, which we denote as $g(\epsilon' | \epsilon_1^{(0)})$, and then compute $f(\epsilon') = \widehat{p}(\epsilon')\pi(\bar{\epsilon}')$ using equation (16). If $f(\epsilon') > f(\epsilon_1^{(0)})$, we update the current state of the sampler to $\epsilon_1^{(1)} = \epsilon'$. If $f(\epsilon') < f(\epsilon_1^{(0)})$, we compute a ratio

$$r = \frac{f(\epsilon')g(\epsilon' | \epsilon_1^{(0)})}{f(\epsilon_1^{(0)})g(\epsilon_1^{(0)} | \epsilon')}, \tag{18}$$

and sample a random number u from a Uniform$[0, 1]$ distribution. If $u < r$, we set the current state of the sampler to $\epsilon_1^{(1)}$ or leave it unchanged at $\epsilon_1^{(0)}$ if $u > r$. If $\epsilon_1^{(0)}$ takes on values close to the values 0 or $10k_BT$, then the probability density $g(\epsilon' | \epsilon_1^{(0)})$ must be 0 when $\epsilon' < 0$ and $\epsilon' > 10k_BT$. The choice of g which we use for the proposal distribution is a truncated normal distribution. The forward proposal distribution $g(\epsilon' | \epsilon_1^{(i)})$, where $\epsilon_1^{(i)}$ is the i-th sample drawn, is a normal distribution with mean $\epsilon_1^{(i)}$ and standard deviation $0.02\epsilon_1^{(i)}$ that is truncated to lie within $[0, 10k_BT]$. The backward proposal distribution, $g(\epsilon_1^{(i)} | \epsilon')$ is a normal distribution with mean ϵ' and standard deviation $0.2\epsilon'$ that is truncated to lie within $[0, 10k_BT]$.

We repeat this process and store the states of the sampler $\epsilon_1^{(0)}, \epsilon_1^{(1)}, \ldots, \epsilon_1^{(b+tm)}$, where t is a thinning factor that we fix at 100, b is the burn-in period that is set to 10000 and m is the number of samples to be retained, which we set at 10000. We then retain only every t-th sample beginning from the b-th sample i.e. $\epsilon_b^{(1)}, \epsilon_{b+t}^{(1)}, \ldots, \epsilon_{b+tm}$, and use it to construct a Gaussian kernel density estimator of $P(\bar{\epsilon}_1 = \epsilon | R^{(1)} = r^{(1)})$, which we denote as $\widehat{P}(\bar{\epsilon}_1 = \epsilon | R^{(1)} = r^{(1)})$. Note that in this process, the normalization constant C_1 in equation (15) need not be computed.

When the read count ratio $r^{(2)}$ is observed after the second replicate of ChIP-seq, we substitute $\widehat{P}(\bar{\epsilon}_1 = \epsilon | R^{(1)} = r^{(1)})$ in place of the prior π in equation (15) and update the posterior $P(\bar{\epsilon}_1 | R^{(1)} = r^{(1)}, R^{(2)} = r^{(2)})$—

$$\begin{aligned}
P(\bar{\epsilon}_1 = \epsilon | R^{(1)} = r^{(1)}, R^{(2)} = r^{(2)}) &= P(R^{(1)} = r^{(1)}, R^{(2)} = r^{(2)} | \bar{\epsilon}_1 = \epsilon)\widehat{P}(\bar{\epsilon}_1 = \epsilon | R^{(1)} = r^{(1)})/C_2 \\
&= P(R^{(2)} = r^{(2)} | \bar{\epsilon}_1 = \epsilon)P(R^{(1)} = r^{(1)} | \bar{\epsilon}_1 = \epsilon)\widehat{P}(\bar{\epsilon}_1 = \epsilon | R^{(1)} = r^{(1)})/C_2 \\
&\approx \widehat{P}(R^{(2)} = r^{(2)} | \bar{\epsilon}_1 = \epsilon)\widehat{P}(R^{(1)} = r^{(1)} | \bar{\epsilon}_1 = \epsilon)\widehat{P}(\bar{\epsilon}_1 = \epsilon | R^{(1)} = r^{(1)})/C_2
\end{aligned}$$

where C_2 is a normalization constant. This process is repeated k times in order to sample from the posterior binding energy distributions $P(\bar{\epsilon}_1 = \epsilon | R^{(1)} = r^{(1)}), \ldots, P(\bar{\epsilon}_1 = \epsilon | R^{(1)} = r^{(1)}, \ldots, R^{(k)} = r^{(k)})$

Change in binding energy due to a single nucleotide change in a binding site

We obtained BEM models of 368 TFs from the BEEML database (`http://stormo.wustl.edu/beeml/`) [13]. These matrices follow the convention where the highest affinity sequence is assigned an energy of 0, with all other sequences being assigned a positive binding energy. The matrices in this database were calculated by fits of binding energy models to protein binding microarray data [13] which measured the affinity of TFs for different 10 bp long sequences. For each TF, there were two BEMs (each of dimension 4×10) available since the matrices were fit individually to replicates of microarray measurements, with the quality of the fit being determined by the R^2 value of predictions from the matrix with measurements. For each TF, we chose the BEM that had the higher R^2 value for further analysis.

For each TF, we picked the sequence whose energy was closest in value to a baseline energy that were in the range $2 - 6k_BT$ in steps of $1k_BT$. For a given baseline energy, we then computed the binding energy of each of the 30 sequences that were a single mutation away from this sequence and stored the absolute values of the energy differences. We repeated this process for every TF in the database and computed the 50-th and 25-th quantiles of binding energy differences. The 50-th quantiles for the baseline energies from $2 - 6k_BT$ were (in units of k_BT) $0.24, 0.24, 0.23, 0.24, 0.24$ and the 25-th quantiles were $0.76, 0.72, 0.7, 0.7, 0.71$.

References

[1] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

[2] Xiao-Yong Li, Sean Thomas, Peter J Sabo, Michael B Eisen, John A Stamatoyannopoulos, and Mark D Biggin. The role of chromatin accessibility in directing the widespread, overlapping patterns of drosophila transcription factor binding. *Genome biology*, 12(4):R34, 2011.

[3] Tommy Kaplan, Xiao-Yong Li, Peter J Sabo, Sean Thomas, John A Stamatoyannopoulos, Mark D Biggin, and Michael B Eisen. Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early drosophila development. *PLoS genetics*, 7(2):e1001290, 2011.

[4] Sam John, Peter J Sabo, Robert E Thurman, Myong-Hee Sung, Simon C Biddie, Thomas A Johnson, Gordon L Hager, and John A Stamatoyannopoulos. Chromatin accessibility predetermines glucocorticoid receptor binding patterns. *Nature genetics*, 43(3):264–268, 2011.

[5] Michael J Guertin, André L Martins, Adam Siepel, and John T Lis. Accurate prediction of inducible transcription factor binding intensities in vivo. *PLoS genetics*, 8(3):e1002610, 2012.

[6] Helge G Roider, Aditi Kanhere, Thomas Manke, and Martin Vingron. Predicting transcription factor affinities to dna from a biophysical model. *Bioinformatics*, 23(2):134–141, 2006.

[7] Qiong Cheng, Majid Kazemian, Hannah Pham, Charles Blatti, Susan E Celniker, Scot A Wolfe, Michael H Brodsky, and Saurabh Sinha. Computational identification of diverse mechanisms underlying transcription factor-dna occupancy. *PLoS genetics*, 9(8):e1003571, 2013.

[8] Xin He, Chieh-Chun Chen, Feng Hong, Fang Fang, Saurabh Sinha, Huck-Hui Ng, and Sheng Zhong. A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PloS One*, 4(12):e8155, 2009.

[9] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. Meme suite: tools for motif discovery and searching. *Nucleic acids research*, 37(suppl_2):W202–W208, 2009.

[10] Jie Wang, Jiali Zhuang, Sowmya Iyer, XinYing Lin, Troy W Whitfield, Melissa C Greven, Brian G Pierce, Xianjun Dong, Anshul Kundaje, Yong Cheng, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research*, 22(9):1798–1812, 2012.

[11] Stephan R Starick, Jonas Ibn-Salem, Marcel Jurk, Céline Hernandez, Michael I Love, Ho-Ryun Chung, Martin Vingron, Morgane Thomas-Chollier, and Sebastiaan H Meijsing. Chip-exo signal associated with dna-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome Research*, 25(6):825–835, 2015.

[12] Gary D. Stormo. Dna binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[13] Yue Zhao and Gary D Stormo. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature biotechnology*, 29(6):480, 2011.

[14] Shuxiang Ruan and Gary D Stormo. Inherent limitations of probabilistic models for protein-dna binding specificity. *PLoS computational biology*, 13(7):e1005638, 2017.

[15] Peter V Kharchenko, Michael Y Tolstorukov, and Peter J Park. Design and analysis of chip-seq experiments for dna-binding proteins. *Nature biotechnology*, 26(12):1351, 2008.

[16] Peter J Park. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

[17] Daechan Park, Yaelim Lee, Gurvani Bhupindersingh, and Vishwanath R Iyer. Widespread misinterpretable chip-seq bias in yeast. *PLoS One*, 8(12):e83506, 2013.

[18] Majid Kazemian, Hannah Pham, Scot A Wolfe, Michael H Brodsky, and Saurabh Sinha. Widespread evidence of cooperative dna binding by transcription factors in drosophila development. *Nucleic Acids Research*, 41(17):8237–8252, 2013.

[19] Dhawal Jain, Sandro Baldi, Angelika Zabel, Tobias Straub, and Peter B Becker. Active promoters give rise to false positive 'phantom peaks' in chip-seq experiments. *Nucleic acids research*, 43(14):6959–6968, 2015.

[20] Yair Benita, Ronald S. Oosting, Martin C. Lok, Michael J. Wise, and Ian Humphery-Smith. Regionalized gc content of template dna as a predictor of pcr success. *Nucleic Acids Research*, 31(16):e99, 2003.

[21] Daniel Aird, Michael G Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B Jaffe, Chad Nusbaum, and Andreas Gnirke. Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. *Genome biology*, 12(2):R18, 2011.

[22] Katharine Best, Theres Oakes, James M Heather, John Shawe-Taylor, and Benny Chain. Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Scientific Reports*, 5, 2015.

[23] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, et al. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome Research*, 22(9):1813–1831, 2012.

[24] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):R137, 2008.

[25] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nature biotechnology*, 27(1):66, 2009.

[26] Yanchun Bao, Veronica Vinciotti, Ernst Wit, and Peter AC 't Hoen. Joint modeling of chip-seq data via a markov random field model. *Biostatistics*, 15(2):296–310, 2013.

[27] Ulrich Gerland, J David Moroz, and Terence Hwa. Physical constraints and functional characteristics of transcription factor–dna interaction. *Proceedings of the National Academy of Sciences*, 99(19):12015–12020, 2002.

[28] Nadia Lalam, Christine Jacob, and P Jagers. Modelling the pcr amplification process by a size-dependent branching process and estimation of the efficiency. *Advances in Applied probability*, 36(2):602–615, 2004.

[29] Yue Zhao, David Granas, and Gary D Stormo. Inferring binding energies from selected binding sites. *PLoS computational biology*, 5(12):e1000590, 2009.

[30] Aaron T Spivak and Gary D Stormo. Scertf: a comprehensive database of benchmarked position weight matrices for saccharomyces species. *Nucleic acids research*, 40(D1):D162–D168, 2011.

[31] Volker Matys, Olga V Kel-Margoulis, Ellen Fricke, Ines Liebich, Sigrid Land, A Barre-Dirrie, Ingmar Reuter, D Chekmenev, Mathias Krull, Klaus Hornischer, et al. Transfac® and its module transcompel®: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(suppl_1):D108–D110, 2006.

[32] Kenneth S Zaret and Jason S Carroll. Pioneer transcription factors: establishing competence for gene expression. *Genes & development*, 25(21):2227–2241, 2011.

[33] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[34] Yashpal Rawal, Răzvan V Chereji, Vishalini Valabhoju, Hongfang Qiu, Josefina Ocampo, David J Clark, and Alan G Hinnebusch. Gcn4 binding in coding regions can activate internal and canonical 5' promoters in yeast. *Molecular cell*, 70(2):297–311, 2018.

[35] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature biotechnology*, 29(1):24, 2011.

[36] Stacia R Engel, Fred S Dietrich, Dianna G Fisk, Gail Binkley, Rama Balakrishnan, Maria C Costanzo, Selina S Dwight, Benjamin C Hitz, Kalpana Karra, Robert S Nash, et al. The reference genome sequence of saccharomyces cerevisiae: then and now. *G3: Genes, Genomes, Genetics*, 4(3):389–398, 2014.

[37] J Michael Cherry, Eurie L Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T Chan, Karen R Christie, Maria C Costanzo, Selina S Dwight, Stacia R Engel, et al. Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Research*, page gkr1029, 2011.

[38] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[39] Sivakanthan Kasinathan, Guillermo A Orsi, Gabriel E Zentner, Kami Ahmad, and Steven Henikoff. High-resolution mapping of transcription factor binding sites on native chromatin. *Nature methods*, 11(2):203, 2014.

[40] Gabriel E Zentner, Sivakanthan Kasinathan, Beibei Xin, Remo Rohs, and Steven Henikoff. Chec-seq kinetics discriminates transcription factor binding sites by dna sequence and shape in vivo. *Nature communications*, 6:8733, 2015.

[41] Peter J Skene and Steven Henikoff. An efficient targeted nuclease strategy for high-resolution mapping of dna binding sites. *Elife*, 6:e21856, 2017.

[42] Arvind YM Sundaram, Timothy Hughes, Shea Biondi, Nathalie Bolduc, Sarah K Bowman, Andrew Camilli, Yap C Chew, Catherine Couture, Andrew Farmer, John P Jerome, et al. A comparative study of chip-seq sequencing library preparation methods. *BMC genomics*, 17(1):816, 2016.

[43] Ranjith Padinhateeri and John F Marko. Nucleosome positioning in a model of active chromatin remodeling enzymes. *Proceedings of the National Academy of Sciences*, 2011.

[44] Jyotsana J Parmar, John F Marko, and Ranjith Padinhateeri. Nucleosome positioning and kinetics near transcription-start-site barriers are controlled by interplay between active remodeling and dna sequence. *Nucleic acids research*, 42(1):128–136, 2013.

[45] Brendan Osberg, Johannes Nuebler, and Ulrich Gerland. Adsorption-desorption kinetics of soft particles. *Physical review letters*, 115(8):088301, 2015.

[46] Răzvan V Chereji and David J Clark. Major determinants of nucleosome positioning. *Biophysical journal*, 2018.

[47] Christoffer Flensburg, Sarah A Kinkel, Andrew Keniry, Marnie E Blewitt, and Alicia Oshlack. A comparison of control samples for chip-seq of histone modifications. *Frontiers in genetics*, 5:329, 2014.

[48] Ryuichiro Nakato and Katsuhiko Shirahige. Recent advances in chip-seq analysis: from quality management to whole-genome annotation. *Briefings in bioinformatics*, 18(2):279–290, 2016.

[49] Timothy L Bailey and Philip Machanick. Inferring direct dna binding from chip-seq. *Nucleic acids research*, page gks433, 2012.

[50] Ping-Yao Zeng, Christopher R Vakoc, Zhu-Chu Chen, Gerd A Blobel, and Shelley L Berger. In vivo dual cross-linking for identification of indirect dna-associated proteins by chromatin immunoprecipitation. *Biotechniques*, 41(6):694–698, 2006.

[51] Ho Sung Rhee and B Franklin Pugh. Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419, 2011.

[52] Naomi Yamada, William KM Lai, Nina Farrell, B Franklin Pugh, and Shaun Mahony. Characterizing protein-dna binding event subtypes in chip-exo data. *bioRxiv*, page 266536, 2018.

[53] Justin B Kinney, Gašper Tkačik, and Curtis G Callan. Precise physical models of protein–dna interaction from high-throughput data. *Proceedings of the National Academy of Sciences*, 104(2):501–506, 2007.

[54] Otto G Berg and Peter H von Hippel. Selection of dna binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *Journal of molecular biology*, 193(4):723–743, 1987.

[55] Sebastian J Maerkl and Stephen R Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–237, 2007.

[56] Xiaoyan Ma, Daphne Ezer, Carmen Navarro, and Boris Adryan. Reliable scaling of position weight matrices for binding strength comparisons between transcription factors. *BMC bioinformatics*, 16(1):265, 2015.

[57] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72–74, 2012.

[58] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 11(2):163–166, 2014.

[59] Qiye He, Jeff Johnston, and Julia Zeitlinger. Chip-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology*, 33(4):395–401, 2015.

[60] Pei Fen Kuan, Dongjun Chung, Guangjin Pan, James A Thomson, Ron Stewart, and Sündüz Keleş. A statistical framework for the analysis of chip-seq data. *Journal of the American Statistical Association*, 106(495):891–903, 2011.

[61] Marko Djordjevic, Anirvan M Sengupta, and Boris I Shraiman. A biophysical approach to transcription factor binding site discovery. *Genome research*, 13(11):2381–2390, 2003.

[62] Madeline A Shea and Gary K Ackers. The or control system of bacteriophage lambda: A physical-chemical model for gene regulation. *Journal of molecular biology*, 181(2):211–230, 1985.

[63] Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome research*, pages gr–112623, 2010.

[64] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[65] Cricket A Sloan, Esther T Chan, Jean M Davidson, Venkat S Malladi, J Seth Strattan, Benjamin C Hitz, Idan Gabdank, Aditi K Narayanan, Marcus Ho, Brian T Lee, et al. Encode data at the encode portal. *Nucleic acids research*, 44(D1):D726–D732, 2015.

[66] Leonid Teytelman, Bilge Özaydın, Oliver Zill, Philippe Lefrançois, Michael Snyder, Jasper Rine, and Michael B Eisen. Impact of chromatin structures on dna processing for genomic analyses. *PloS one*, 4(8):e6700, 2009.

[67] Leonid Teytelman, Deborah M Thurtle, Jasper Rine, and Alexander van Oudenaarden. Highly expressed loci are vulnerable to misleading chip localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences*, 110(46):18602–18607, 2013.

[68] Torsten Waldminghaus and Kirsten Skarstad. Chip on chip: surprising results are often artifacts. *BMC genomics*, 11(1):414, 2010.

[69] Justus M Kebschull and Anthony M Zador. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research*, 43(21):e143–e143, 2015.

[70] Theodore E Harris. *The theory of branching processes*. Courier Corporation, 2002.

[71] James E Gentle. *Random number generation and Monte Carlo methods*. Springer Science & Business Media, 2006.

[72] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[73] Ryan K Dale, Brent S Pedersen, and Aaron R Quinlan. Pybedtools: a flexible python library for manipulating genomic datasets and annotations. *Bioinformatics*, 27(24):3423–3424, 2011.

[74] Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771, 2009.

[75] Ivan Kulakovskiy, Victor Levitsky, Dmitry Oshchepkov, Leonid Bryzgalov, Ilya Vorontsov, and Vsevolod Makeev. From binding motifs in chip-seq data to improved models of transcription factor binding sites. *Journal of bioinformatics and computational biology*, 11(01):1340004, 2013.

[76] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[77] Eric Jones, Travis Oliphant, and Pearu Peterson. Scipy: open source scientific tools for python. 2014.