Contents

Illustrations

Tables

Contributors

Jonathan Cairns *Department of Oncology, University of Cambridge*
Andy G Lynch *Department of Oncology, University of Cambridge*
Simon Tavaré *Department of Oncology, University of Cambridge*

1

Statistical Aspects of ChIP-seq Analysis

1.1 Introduction – The purpose of the ChIP-seq experiment

Every cell is host to a diverse ecosystem of proteins, each protein having its own functional properties. Cell behaviours and processes, such as growth, are dependent on the levels of these various proteins.

Through microarrays, and subsequently high-throughput sequencing, we have become adept at quantifying the expression levels of genes, in various cell types under different conditions. These expression levels provide us with a proxy, albeit an imperfect one, for the levels of protein production in a cell. (Estimates of the correlation between mRNA levels and protein levels are typically very variable (Gry *et al.*, 2009).)

We can infer links between cause and effect by perturbing a cell in some way, and measuring which genes' mRNA levels change in response. However, this information is, by itself, unsatisfying – specifically, we seek to understand the regulatory mechanisms underlying these links. This has clinical significance – ultimately, we may be able to manipulate these mechanisms ourselves, potentially leading to novel therapies.

Consider the following example: In breast cancer, malignant tumours typically exhibit an invasive, proliferative behaviour characterized by abnormal growth (Weinberg, 2007). A critical factor in establishing this behaviour is the estrogen receptor (ER). In particular, ER is known to encourage mitotic cell division. This function is, in itself, not dangerous – cells in healthy tissue commonly divide in order to replace lost cells. However, ER is known to be a key factor in establishing cancer, and around 70% of breast tumours are labelled as "ER-positive" – that is, their cells have ER content above a certain threshold (Tannock and Hill, 1998; Mohibi *et al.*, 2011).

ER is an example of a transcription factor (TF) – a protein that binds to a DNA region, called a transcription factor binding site (TFBS), and alters the way in which that region is interpreted by transcription machinery. This effect can either encourage or discourage transcription. In our example, ER usually encourages the transcription of various genes that facilitate cell division, such as c-Myc and cyclin D1 (Tannock and Hill, 1998; Dubik and Shiu, 1992; Eeckhoute *et al.*, 2006).

Since ER expression is critical for tumour formation and growth, there have been many attempts to disable its function. Notably, the drug Tamoxifen inhibits ER, and can prolong disease-free survival in ER-positive tumours (Fisher *et al.*, 1989; Smith and Dowsett, 2003).

However, some ER-positive tumour cells are resistant to Tamoxifen, or evolve resistance over time (Shou *et al.*, 2004) – as a result, tumours can regrow, even after Tamoxifen treatment. Either the tumour's dependency on ER has been circumvented, or ER is still functional and could be inhibited in some other way – for example, by inhibiting a protein that is required for ER binding, such as FOXA1 (Hurtado *et al.*, 2011).

To better understand Tamoxifen resistance, we seek to clarify ER's action. For this, it is important to determine the locations to which ER binds, the strength of the binding, and the TFs that bind nearby, affecting ER's function. This problem has a broad scope – many TFs have unknown function, or have transcription programs that are not fully elucidated, and so methods developed in this area can be applied in a wide range of settings across molecular biology.

ChIP (Chromatin ImmunoPrecipitation) locates and quantifies interactions between DNA and proteins (such as those involving ER in the example above). We do this by creating a ChIP library: a pool of DNA, each molecule of which should contain at least one TFBS. By investigating the sequences in the library, we can clarify the motifs to which the TF binds. Moreover, if we align these sequences back to a reference genome, we can locate the original binding events, and quantify the "strength" of each event based on the number of fragments obtained from that location. (Typically, "strength" or "affinity" within a cell population is defined as the proportion of cells in which a TFBS is occupied.)

Until recently, microarray technology was the usual method for determining the contents of ChIP libraries. This technique, known as ChIP-chip, is useful, but suffers from the drawbacks of microarrays: in particular, we must choose which sequences to test for, in advance. Without prior knowledge of TFBS locations, a tiling microarray is most appropriate. The human genome has size 3×10^9 base pairs, and so a genome-wide

array with 5×10^6 features can only provide a resolution of 600 base pairs. However, if we have prior knowledge of the locations to which the TF binds, then we can improve on this resolution by choosing a more specific array design. For example, if we believe that a TF binds to promoter regions, then we can use probe sequences that tile promoter regions, rather than the entire genome (Buck and Lieb, 2004).

Increasingly, researchers are turning to high-throughput sequencing when investigating ChIP libraries. The combined use of ChIP and sequencing is referred to as ChIP-seq. In theory, this should be much more powerful than ChIP-chip, with potential advantages including single-nucleotide fragment resolution, lower amount of DNA required, and better coverage (Park, 2009). However, in practice, ChIP-seq analysis is still in its infancy, and available methods do not take full advantage of the data. In particular, as we shall see, the single-nucleotide resolution is often not exploited, as read counts are often collected over large bins.

There is, therefore, great demand for appropriate statistical procedures that can navigate the vast data sets produced through high-throughput sequencing experiments, and extract the key biological information.

1.2 Aims

Typically, we want to answer questions such as:

- Where does the TF bind?
- Which sites exhibit stronger/weaker binding?
- Are there binding differences between two conditions or phenotypes?
- Do other regulatory proteins interact with the TF, and are any of these coregulators required for its action?
- Which genes have expression levels altered by the TF?

1.3 Experimental overview

A ChIP-seq experiment consists of making a ChIP library from a large cell population, followed by sequencing. Typically, the cell population size is of the order of 10^7 cells (Schmidt *et al.*, 2009; Kidder *et al.*, 2011). The essential steps are as follows:

Cross-linking: Formaldehyde is used to bind the TF covalently to its associated DNA. Note that this step's effect is not specific to

interactions between the TF of interest and its TFBSs – all other binding events occurring in the cell are affected, as well.

Fragmentation (Sonication): A high-frequency sonic pulse is used to randomly break the DNA into small fragments.

ImmunoPrecipitation (IP): We use antibodies, specific to the TF of interest, to pull out any instances of the TF along with the DNA fragments to which they are bound. Any remaining fragments are washed away. (The sensitivity and specificity of the antibody must be identified in order to interpret the data correctly. Here, we assume that the sensitivity and specificity are high.)

Reverse cross-linking and purification: The protein is removed from the fragments of interest, usually by subjecting the library to high-temperatures that reverse the cross-linking step. The DNA is then purified, so that we are left with a solution containing only the DNA fragments of interest.

Sequencing is used to determine the contents of this pool of DNA. We simplify down only to the ChIP-seq-pertinent steps:

Polymerase Chain Reaction (PCR) amplification: In order to increase the material present for sequencing, the library is "amplified" using PCR – that is, the DNA molecules in the library are repeatedly duplicated. This procedure is known to be biased, duplicating certain sequences at faster rates (Aird *et al.*, 2011).

Size selection: Gel electrophoresis is used to collect only DNA fragments whose lengths fall within a defined interval. The interval choice can vary, but is typically around 200-300bp (Schmidt *et al.*, 2009). This step is required by most sequencing protocols.

Sequence reads: We sequence a subset of the DNA fragments in the library as follows. For each fragment, one of its two strands is chosen at random, and we determine the first k-mer (that is, the first k base pairs) of that strand's sequence. Thus, we obtain a set of k-mers, each of which could have come from either end of its associated fragment. Alternatively, both ends of each fragment can be sequenced simultaneously – so-called "paired-end sequencing" is discussed in Section (1.3.2). For most sequencing platforms, k is 36 or greater, although the value of k can often be increased, at extra cost, by extending run time.

Alignment: The k-mers are aligned back to a reference genome; that is, each read is mapped to its genomic location of origin, either on the + strand or on the − strand (the − strand's sequence

being the reverse complement of the + strand's). For simplicity, in this chapter we will assume that each sequence maps to a unique genomic position – typically, this assumption is not true in real sequencing data. We will consider only the $5'$ ends of each read, for the majority of this chapter.

Thus, any region surrounding a TF-DNA interaction accumulates a large number of reads.

1.3.1 Control experiments

Of course, none of these steps can be performed perfectly. Each will introduce some level of noise, or bias towards certain properties. Therefore, it is desirable to perform a control experiment alongside any ChIP-seq experiment, in an attempt to quantify noise, and also to compensate for any cell-specific biases. Different methods can be used to produce control data (Kidder *et al.*, 2011) – we will not discuss them in detail but, briefly, the most common control types are:

TF-null: Knockdown or knockout the TF.

IgG: Use Immunoglobulin G (IgG) as the antibody in the IP step.

Input: Omit the IP step in its entirety.

There is no consensus on which of these is "best". The knockdown strategy, while theoretically ideal, can be impractical since it yields very little DNA; if increased PCR amplification is used to compensate for the low yield, our data can become noisy or unusable. The other strategies are useful practical compromises.

1.3.2 Paired-end sequencing

When a ChIP library is investigated with paired-end sequencing, the procedure is known as ChIP-PET ("Chromatin Immunoprecipitation with Paired-End Tags" – Wei *et al.* (2006)). ChIP-PET is more expensive than single-end ChIP-seq, and thus tends to be much less common.

However, there are advantages to using paired-end sequencing in the context of ChIP – notably, we obtain information about fragment length (Section 1.3.8), and PCR duplication errors (Section 1.3.4). We could also detect structural instabilities, if they are of interest.

1.3.3 The data

Consider a simplified example, where a TF has exactly one TFBS, which consists of a single base pair. After cross-linking and fragmenting, the TF is bound to a DNA fragment of random length. We pull out the TF-DNA complex during the IP step, and sequence one of the DNA fragment's strands, at random. As we perform ChIP simultaneously on a large number of cells, we have many replicates of this procedure. Thus, when we align back to the genome, reads map to two regions near the TFBS: one on the positive strand, and the other on the negative strand. Each region is located $5'$ of the TFBS, relative to its own strand. This procedure results in a characteristic "peak" shape (see Figure 1.1).

1.3.4 Potential sources of error and bias

In an ideal situation, only these peak shapes would be present.

However, in practice, the IP step is not perfect, and so our DNA pool also contains "noise" – that is, DNA fragments that were not bound to the TF. These fragments contribute reads that are randomly distributed across the genome. Perhaps counter-intuitively, this positional distribution is not uniform, and there is clear bias towards certain sequences. To optimize ChIP-seq analysis, we need to take account of these biases.

The PCR amplification step can introduce so-called "PCR artefacts" – excessive amplification of certain sequences. This can be problematic, as a PCR artefact may be mistaken as representing a true binding event. It is also conceivable that a true event could be masked, if a PCR artefact distorts a peak's shape beyond recognition, or merges two nearby peaks.

As a result, some sources advocate deleting duplicates, the aim being to remove any observed duplication events that could have been introduced by PCR. A common procedure is to delete any reads that map to a given location, beyond the first instance (Visel *et al.*, 2009). This strategy can be improved by looking at SNPs, since two reads in the same location but with different sequence cannot be PCR duplicates.

Deleting duplicates accounts for situations where multiple copies of a read have been observed due to PCR over-amplification. However, the effect of PCR amplification is not limited to the appearance of duplicates – PCR over-amplication can increase the presence of a particular sequence in the library, thus increasing the probability of observing that sequence once. We cannot account for this situation by removing duplicates.

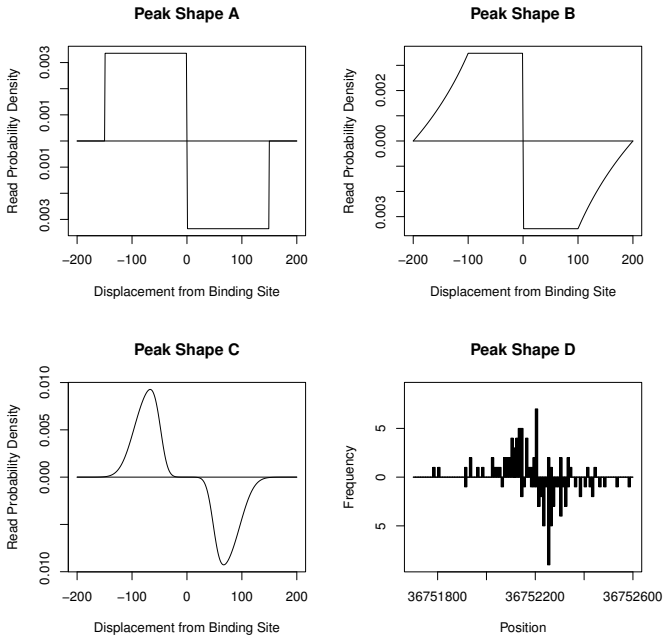

Figure 1.1 Histograms representing the characteristic ChIP-seq "peak" shape. Shapes A-C represent simulated data, under various arbitrary mechanistic assumptions. The y-axis represents the probability of a read being located in a particular location, where values below the x-axis represent the probability density of reads occurring on the negative strand. For peak shape A, we assume that fragment length is fixed, and that a TF's location on a fragment is uniform. When moving to peak shape B, the assumption that fragment length is fixed is relaxed, and now assumed to be uniform. For peak shape C, the TF's location on a fragment is assumed to have truncated normal distribution, with standard deviation proportional to fragment size. Shape D was plotted using actual data from a p53 ChIP-seq experiment.

Moreover, the duplicate-removing strategy removes "true" duplicates as well, restricting our dynamic range in regions with signal. As sequencing depth increases, this will seriously limit our data, as we will reach the saturation point (that is, the point where the majority of base pairs in the genome have an associated read). Therefore, deleting duplicates cannot be an effective long term strategy, and we will need techniques to specifically target these PCR duplicates – either from knowledge of the sequence, or by allowing for anomalous counts in our statistical models.

It has been suggested that an appropriate compromise is to remove duplicates when peak-calling (Section 1.4), but to reinstate them when performing any in-depth analysis on peak regions (Chen *et al.*, 2012).

If we have paired-end data, then we can make a stronger inference. Suppose that a ChIP fragment contains a sequence that is vulnerable to PCR overamplification. This fragment spawns many PCR duplicates, all of which begin and end at the same genomic locations. In contrast, "real" fragments, obtained from pulling down a TFBS, do not have to start or end in the same place. Therefore, we improve our duplicate-removal strategy by deleting a pair of reads if and only if its ends identically match another pair's ends (Chen *et al.*, 2012).

Other "artefacts" may exist in the data. There are genomic regions that always appear to be significantly enriched for aligned reads, regardless of the experiment. The ENCODE consortium (Myers *et al.*, 2011) have compiled a list of these regions – the list is available in the mappability section of their data repository.

These artefacts have the potential to confuse our fragment length estimation algorithms, and so it is common to delete reads that fall in these regions. This is a reasonable short-term strategy, but it prevents us from detecting any changes in read density in those regions. Additionally, uncommon artefacts may occur outside of the blacklist regions – for example, if there is significant contamination in our DNA library.

In RNA-seq, it is well known that the count of a given sequence is biased by that sequence's GC content – that is, the percentage of bases that are G or C (Zheng *et al.*, 2011; Benjamini and Speed, 2012). However, it is extremely difficult to correct for GC bias in ChIP-seq data, because any observed GC bias could be a property of the biology being studied. For example, if a TF commonly binds to GC-rich regions, we might expect to see a "bias" towards high GC content.

In ChIP-seq, it is possible that the sonication step, or the sequencing step, is biased towards particular sequence content, although we lack a consensus on errors of this type. Hansen *et al.* (2010) did not find bias in the sonication step. In contrast, Cheung *et al.* (2011) reported sequencing biases, and suggested a read-weighting normalization method based on GC, mappability, and other local effects.

Of course, there may also be biases that depend on the sequencing platform used. Each new platform will have its own biases, so algorithms should evolve to accommodate appropriate corrections and keep up with changes in technology.

1.3.5 Histones/nucleosomes

The ChIP-seq experiment is not limited to investigating TF binding. In theory, we can detect any genomic binding event, provided that we can use antibodies to pull out the molecules involved in that event.

DNA is wrapped around histones, in structures called nucleosomes. Histones have "tails" that can be modified, acting as signalling flags for other molecular machinery. There are various potential histone modifications, each specified using a code. For example, "H3K4me3" specifies that the 4th amino acid ("K4") on histone subunit 3 ("H3") has been trimethylated ("me3").

Though the biological function of these marks is not fully understood, some relationships between marks and expression levels are known. Certain histone marks associate with active genes – for example, H3K36me3 (Sims and Reinberg, 2009) and H3K4me3 (Okitsu *et al.*, 2010). Other marks associate with inactive genes – for example, H3K9me3 and H3K27 (Barski *et al.*, 2007). Combinations of modifications can enable additional signal types (Bannister and Kouzarides, 2011). We can perform ChIP-seq using antibodies that are specific to these modified histone tails, allowing us to identify modification events.

Usually, data obtained from histone mark ChIP-seq have very different qualities to data from transcription factor ChIP-seq. In particular, there are often many more events of interest in the histone mark data. We are typically interested in locating large regions, each of which contains a high number of histone modifications. This means that we must use different methodologies when analysing histone mark data.

Not all histone marks have these properties. H3K4me3 (an activatory histone mark) is usually observed only in gene promoter regions, with peak shapes that resemble those from TF ChIP-seq. H3K4me3 data are typically analysed using TF peak-calling methods.

Similarly, we can use antibodies specific to nucleosomes themselves, in order to determine their locations. This procedure, known as "nucleosome positioning", is discussed in greater detail in Chapter **??**.

1.3.6 ReChIP

If we are interested in a genomic event where two or more transcription factors bind in the same locus, then we can use a technique called reChIP. For example, Ross-Innes *et al.* (2010) were interested in the interaction between ER and RARα. When crosslinking, all binding events

are reinforced by covalent bonds, and so any interactions between ER, RARα and DNA will result in a covalently-bound complex. Thus, we can perform two successive immunoprecipitation steps – one enriching for ER, and another for RARα. The DNA sequences pulled down in this way should correspond to the locations where both ER and RARα bind.

The peaks in these data are very similar to normal ChIP-seq peaks and, therefore, the same analytical techniques are typically used. However, we obtain less DNA from two ChIPs than one. This may mean that we have to perform more amplification, magnifying PCR effects.

1.3.7 Other experiments

Other genomic events can be detected using an antibody, and thus can be investigated through experiments similar to ChIP-seq. Data obtained from these experiments typically have very different properties to ChIP-seq data, and therefore we will not discuss them in any great detail.

BrdU-seq identifies regions of the chromosome that are undergoing replication (Morstyn *et al.*, 1983). MeDIP-seq is used to find methylated regions of DNA, as opposed to methylated histones (Li *et al.*, 2010b).

ChIA-PET (Chromatin Interaction Analysis – Paired-End Tags) is used to discover long-range interactions: a TF-DNA complex is ChIPped out as before, and DNA fragments bound to the same TF are encouraged to bind to each other. By sequencing these joins, we determine when a TF is bound to multiple DNA loci that are far apart in terms of genomic coordinates, but close in 3-dimensional space (Li *et al.*, 2010a).

Similar experiments investigate the 3-dimensional structure of DNA, but without reference to any particular TF. In order of increasing scope, these are: 3C, 4C, 5C and HiC (De Wit and De Laat, 2012).

1.3.8 Estimating fragment length

Define "fragment length" as the number of base pairs contained in a given DNA fragment. For nearly all ChIP-seq analysis, knowledge of the fragment length distribution is required; usually, only the mean fragment length is taken into account. We can directly infer this quantity, if we performed a paired-end experiment. If we did not, then this quantity can be estimated via other means:

1. We can perform experiments that directly measure the lengths of fragments in the library (Lee *et al.*, 2006).

2. We can estimate average fragment length from the data; for example, using MACS' mode-based technique (Zhang *et al.*, 2008a) or Tag Autocorrelation from HOMER (Heinz *et al.*, 2010).

In the authors' experience, these estimates often substantially disagree. Fragment length estimation is frequently built into analysis programs but, since fragment length estimation algorithms are imperfect, programs should allow this estimation procedure to be overridden.

1.4 Peak-calling in TF data

Our first aim is to "peak-call" – that is, find the locations where a transcription factor binds. Formally, we can think of a peak-caller as a function that maps a set of reads to a set of genomic intervals.

There are two main strategies used to call peaks:

Count-based: Define regions. Count the number of reads falling into each region. When a region contains a statistically significant number of reads, call that region as a peak.

Shape-based: Consider individual candidate binding sites. Model the spatial distribution of reads in surrounding regions, and call a peak when the read distribution conforms to the expected distribution near a binding site.

The implementations of these strategies can vary wildly. It is common to see new peak-callers published on an experiment-by-experiment basis, and thus great variation exists in peak-callers.

We will take a look at each strategy in turn, but first we examine a number of preliminary steps that are common to both strategies.

1.4.1 Strategy-independent issues

A simple mathematical model of ChIP-seq is as follows: Our ChIP-seq library contains a mixture of fragments. During the sequencing step, we take Q different fragments from this library, at random, sequence one end of each (with the choice of end again being random) and align the read sequences to a reference genome. Let X_i^+ be the number of reads that align to the + strand, and have $5'$ end located at genomic position i. Similarly, let X_i^- be the number of reads on the − strand that have $5'$ end at position i. Some of the subsequent results in this chapter apply

equally to both strands – where this is the case, we replace the strand with a star: $X_i^\star$.

Provided that Q is large enough, but not so large as to approach the number of fragments in the library, we can consider $X_i^\star$ to be a Poisson count (Marioni *et al.*, 2008) with mean $\mu_i^\star = Q\rho_i^\star$, where $\rho_i^\star$ is the fraction of sequence $s_i^\star$ present in the library.

$$X_i^\star \sim Pois(\mu_i^\star)$$

We assume that the variables $X_i^\star$ are independent – this is true if our sequencing step, where we randomly select DNA fragments from the ChIP-seq library, occurs without bias. Now, if we take a bin that starts at position $i = a$ and ends at $i = b$, then the read count in that bin (considering only reads on one strand) has conditional distribution

$$X_{[a,b]}^\star = \sum_{i=a}^{b} X_i^\star \sim Pois\left(\sum_{i=a}^{b} \mu_i^\star\right) \stackrel{def}{=} Pois\left(\mu_{[a,b]}^\star\right)$$

Note that these results do not hold if duplicates were removed since, in this case, $X_i^\star$ are Bernoulli variables with differing parameters.

Normalization

Ideally, when comparing multiple replicates, one should normalize the data. Techniques exist for bin count data (Robinson and Oshlack, 2010), but there is no well-established whole-genome inter-sample normalization technique. As such, it is common to perform a simple normalization step, such as scaling any parameters representing mean counts by factors proportional to the number of reads in each sample. (In general, enriched and unenriched regions should have distinct scaling factors.)

Even in the case where we have reduced to bin count data, normalization should be performed with care in ChIP-seq data (Section 1.8.1).

Overdispersion

So far, we have assumed that the means $\mu_i^\star$ are non-random. This is true for a single library at a single site. However, we must be wary in the following situations:

1. modelling bin counts at one site across multiple libraries.
2. modelling bin counts from multiple sites in one library.

Consider situation 1. Suppose that we have sequenced multiple libraries, each derived from a distinct cell population (see Section 1.6.5).

Define a genomic bin $[a, b]$ and, considering reads from both strands, let X be the read count in this bin. Thus, after sequencing multiple libraries, we have multiple samples from the variable X. If we use the model $X \sim Pois(\mu)$ with constant μ, then we enforce the equality $Var(X) = \mathbb{E}(X)$. This equality is typically not observed in actual data – in fact, the variance is often much greater than the mean. As such, the Poisson distribution is not an appropriate fit.

As an illustration of this phenomenon in practice, we verified it in a set of three input libraries derived from so-called "IMR-90" cells, taking read counts from bins of width 100 that tiled the region $10^7 - 10^8$ of chromosome 1 (this region selected to avoid the centromere and telomeres). We applied the Z-test for overdispersion (Hilbe, 2011), rejecting the null hypothesis that the count data had mean equal to variance, at a significance level of $p < 2 \times 10^{-16}$.

The property that, in our data, the variance greatly exceeds the mean is known as "overdispersion". One explanation for the presence of overdispersion in ChIP-seq data is that μ is not constant across libraries, and should instead be considered an emission from a random variable, M. We now consider the entire ChIP-seq experiment as a two-step process: firstly, we create N ChIP libraries, equivalent to taking N independent samples $\mu_1, \ldots, \mu_N$ from $M^\star$. Secondly, since these values are hidden, we must sequence each library, obtaining one sample from each of the distributions $Pois(\mu_1), \ldots, Pois(\mu_N)$. Now, by the law of total variance,

$$\begin{aligned} Var(X) &= \mathbb{E}(Var(X|M)) + Var(\mathbb{E}(X|M)) \\ &= \mathbb{E}(M) + Var(M) \\ &= \mathbb{E}\left(\mathbb{E}(X|M)\right) + Var(M) \\ &= \mathbb{E}(X) + Var(M) \end{aligned}$$

Thus, the overall variance is a sum of two terms: the first representing Poisson variation, and the second representing an additional contribution from the underlying mixture distribution.

A common statistical solution to situation 1 is to use the Negative Binomial (NB) distribution, a distribution of which the Poisson distribution is a special case. Setting X to have NB distribution is equivalent to choosing M to be gamma-distributed. The NB distribution allows arbitrary, independent choices of mean and variance, enabling us to model the overdispersion in the data.

Now, consider the second situation. Suppose that we have sequenced one library. Define multiple bins $[a_i, b_i]$ across the genome, choosing the bins such that they have constant width. For simplicity, let us assume that each bin is "enriched" or "unenriched", with bin i having an associated bin count X_i. We might assume that the enriched bins' counts have identical Poisson distribution, and similarly for the unenriched bins. As before, this assumption requires equivalence of mean and variance, a property that is contradicted in actual data (Spyrou *et al.*, 2009).

To account for this effect, we can again introduce a mixing distribution M that describes the biological variability between sites. For example, we can set the X_i to have common NB distribution as before. A popular alternative is to allow site-specific distributions – for example, set $X_i \sim Pois(\mu_i)$, allowing site-specific means μ_i. If using a Bayesian approach, we can reconcile this with the previous approach by setting a Gamma prior distribution for Θ_i.

Ideally, we should ensure that our choice of mixing distribution can handle both of these situations simultaneously. Some peak-callers do not allow for overdispersion, which may lead to overestimation of means and underestimation of variances. The latter effect is especially problematic, as it typically inflates the significance of hypothesis tests.

Modelling Fragment Length

There is a dependency between the bin counts $X^+_{[a,b]}$ and $X^-_{[a',b']}$, and we do well to model it when peak-calling.

Suppose that TFBS i is located at location s. This TFBS contributes fragments to the library – typically, we assume the following:

1. Each TFBS's contributions are independent of other TFBSs.
2. We have rotational symmetry – that is, reads contributed by the TFBS align to location $s - Y$ (+ strand) or $s + Y'$ (− strand) with equal probability, where Y, Y' are independent, identically distributed discrete random variables. The size selection procedure enforces that Y falls in some interval $[y_{min}, y_{max}]$.
3. Y has symmetric distribution about some integer λ – thus, $\mathbb{E}(Y) = \lambda$.
4. λ is constant (otherwise, the computation is typically intractable).

The average fragment length is $\mathbb{E}\left((s + Y') - (s - Y)\right) = 2\mathbb{E}(Y) = 2\lambda$. It can be shown, based on the above assumptions, that

$$X^+_a \stackrel{d}{=} X^-_{a+2\lambda}$$

That is, after a shift of 2λ, the positive and negative read counts have

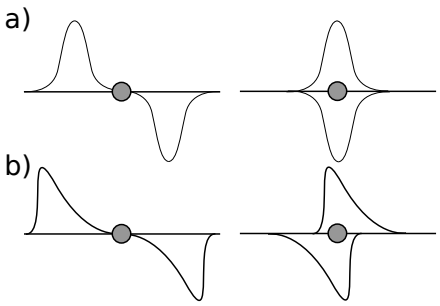

Figure 1.2 Cartoon illustration of the symmetry assumption. The grey circle represents a TFBS, with hypothetical distributions of reads plotted for each strand. If Y has a symmetric distribution, as in a), then shifting the reads on each strand by a half-fragment length results in the two strands having equal distribution. This is not the case when Y has an asymmetric distribution, as in b).

equal distribution (see Figure 1.2). By summing over multiple genomic locations, we obtain the corollary that bin counts $X^+_{[a-\lambda,b-\lambda]} \stackrel{d}{=} X^-_{[a+\lambda,b+\lambda]}$.

Thus, a common modelling strategy is to obtain $\hat{\lambda}$, some estimate of the half-fragment length λ, and link the distributions of $X^+_{[a-\hat{\lambda},b-\hat{\lambda}]}$ and $X^-_{[a+\hat{\lambda},b+\hat{\lambda}]}$. From a computational point of view, we can either model these offset bins *in situ*, or we can shift the reads by adding $\hat{\lambda}$ to the positions of reads on the positive strand, and subtracting $\hat{\lambda}$ from the positions of reads on the negative strand. The latter allows us to forget about the offset between strands when modelling, and focus on a single set of genomic co-ordinates.

Pile-up

"Pile-up" is an alternative fragment-modelling strategy. So far, we have used only the $5'$ end of each read. Pile-up counts are obtained by treating each base pair in a read as a separate observation, and binning these observations as before. These counts can be useful for visualization.

The early peak-caller XSET (Robertson *et al.*, 2007) uses this strategy, as does CSAR (Kaufmann *et al.*, 2009). Each read is extended to the average fragment length estimate. Subsequently, pile-up is calculated, and each pile-up count is modelled as a Poisson random variable.

Using pile-up in this fashion is equivalent to a two-step process: shift read locations by $\hat{\lambda}$, as before, and then smooth using a symmetric uniform kernel whose width is equal to the average fragment length.

When pile-up is looked at from this perspective, it becomes clear that this kernel choice, chosen for its computational ease, is completely arbitrary. Boyle *et al.* (2008) Valouev *et al.* (2008) and Beck *et al.* (2012) smooth with a Gaussian kernel, and Wu *et al.* (2010) use a wavelet-based smoothing technique. There is yet to be a thorough investigation of kernel choice in the literature.

1.4.2 Count-based strategies

Count-based strategies search for regions where there are significantly more reads in the treated data than in the control data (or background regions, when there are no control data).

Usually, we wish to find the smallest such significant regions – intuitively, these best describe the binding event of interest, and reduce the computation required for later steps such as motif analysis.

Historically, ChIP-seq experiments have been performed without replicates, and most current peak-callers assume that only 1 ChIP-treated sample and 1 control sample exist (or that each sample's replicates are sufficiently similar that their reads have been combined) – indeed, this assumption is made by all of the peak-callers in this section. In this setting, we have no estimate of the variation between samples. As such, we need to make some additional assumptions.

In this section, we assume that fragment length has been corrected for. T^+ and T^- represent the ChIP-treated sample's read count in a given bin of size w, using only reads on the $+$ or $-$ strand as appropriate. C^+ and C^- represent the corresponding bin counts in the control sample. The superscript is dropped when reads from both strands are counted.

MACS (Zhang *et al.*, 2008a) is, arguably, the most commonly used peak-caller. MACS makes the key assumption that, under the null hypothesis, the ChIP sample's bin count has distribution $T \sim Pois(\hat{\lambda})$, where $\hat{\lambda}$ is estimated from the control sample: $\hat{\lambda} = C$ (with an adjustment when C is unusually high compared to estimates from larger bins).

To apply this method to the entire genome, MACS slides a window of size w across each chromosome and tests for significance at each location. Whenever a window passes the significance threshold, all of the base pairs in that window are marked as enriched. The output peak-calls consist of the union of all enriched base pairs. There is a trade-off when selecting w: if w is large, this approach can inflate the size of peak-calls, but if w is small, we risk calling too many false positives.

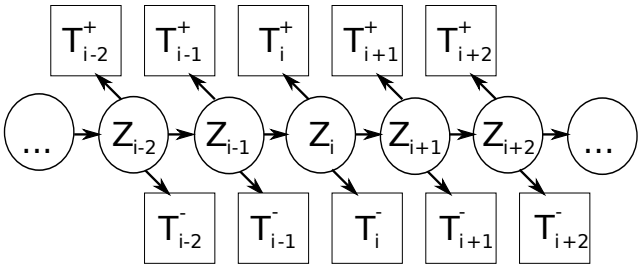

Figure 1.3 Use of a Hidden Markov Model – A graphical model showing the dependencies of bin counts $T_i^\star$ on the hidden states Z_i.

Many peak-callers similarly use a Poisson-based bin count threshold (Johnson *et al.*, 2007; Robertson *et al.*, 2007; Kaufmann *et al.*, 2009).

Other peak-callers use different enrichment tests. F-seq (Boyle *et al.*, 2008) calculates a T threshold by assuming that data are normally distributed. ChIP-PaM (Wu *et al.*, 2010) assumes a negative binomial model. QuEST (Valouev *et al.*, 2008), having used Gaussian smoothing, applies a number of arbitrary thresholds – requiring, for example, that 600 tags are present in a given 300bp window, and that the local maximum in a bin has height at least 20-fold above control.

HMM-based approach

One downside of the count-based methods described so far is that they do not model the dependence of nearby bins – namely, if we tile the genome with bins of width w, where w is small compared to peak width, then we expect peaks to consist of multiple adjacent enriched bins.

We can model this spatial dependency with a Hidden Markov Model (HMM), a structure that has been used successfully in many biological settings (see Durbin *et al.* (1998)). Here, we have an underlying Markov chain, Z_i, the states of which are not observed. Each state Z_i takes a value of 0 or 1, representing absence or presence of TF binding, respectively. Z_i emits observations T_i^+ and T_i^-, where $T_i^+\big|_{Z_i=1}$ takes larger values than $T_i^+\big|_{Z_i=0}$. The probabilities of transitioning between Z_i states are represented in a transition matrix P (see Figure 1.3).

BayesPeak (Spyrou *et al.*, 2009; Cairns *et al.*, 2011) uses an HMM in this way. Gibbs sampling is used to sample from the posterior distribution of (P, Z). In contrast, the algorithm HPeak (Qin *et al.*, 2010) uses a frequentist approach to fit hidden states.

Other peak-callers based on HMMs are specifically designed for histone modification data; these are described in Section 1.5.

We can model the dependency between bin counts in other ways. For example, iSeq (Mo and Liang, 2010) uses a Hidden Ising Model in place of an HMM, although it ignores fragment length and strand information.

1.4.3 Shape-based strategies

Shape-based strategies consist of modelling the positions of reads relative to a candidate TFBS, rather than modelling the counts of reads at a given position. This allows us to model the shape of a peak, explicitly.

Intuitively, shape-based modelling strategies may seem more appropriate than their count-based counterparts, as they directly model the underlying physical process at work. However, shape-based peak-calling is computationally difficult in comparison to count-based techniques, and so the implementations can be processor-intensive. The size of the human genome is approximately 3×10^9 base pairs and, if we consider each base pair as a potential TFBS, then we need to fit a read distribution near each point. Fitting individual spatial distributions via Monte Carlo approaches, such as MCMC or Gibbs Sampling, is infeasible, as we must perform this procedure approximately 3×10^9 times.

Therefore, it is typically necessary to perform a preliminary step that reduces the scope of our analysis down from the full genome to a number of sites of interest. We must choose a very permissive preliminary step, since any regions excluded at this point cannot be recovered.

Another potential difficulty (though by no means insurmountable) is that a successful method must distinguish binding sites that are very close to one another. In particular, if there is a binding element that contains two adjacent sites, then the local distribution will be a mixture of two peaks – our peak-calling algorithm ought not to be confused by this. In extreme cases, there could even be extensive binding elements that a TF can bind to at any point – a shape-based algorithm should explicitly model this situation, otherwise such a region could be rejected.

Perhaps as a result of these complexities, there are fewer shape-based peak-callers than count-based ones. PICS (Zhang *et al.*, 2011) assumes that the reads about a TFBS have locations that follow Student's t-distribution. This algorithm is described in full in Chapter **??**.

ChIP-PaM (Wu *et al.*, 2010) invokes a shape-assessment step based on wavelet-smoothing and the assumption that, near TFBSs, we obtain sinusoidal data after subtracting − strand read counts from + strand

read counts. This strategy could have difficulty distinguishing adjacent TFBSs, as reads from one event can cancel out reads from the other.

We can model peak shape without referring to binding events. T-PIC (Hower *et al.*, 2010) constructs a "tree" based on the pile-up track's branching shape in candidate regions, and performs a hypothesis test based on the tree's topological structure.

1.5 Peak-calling in Histone Mark Data

Recall from Section 1.3.5 that, in histone mark data, we typically seek large regions of enrichment rather than individual binding events.

In theory, peak-calling in histone mark ChIP-seq data should be easier than in TF ChIP-seq data, since the data contain a stronger signal.

A pragmatic approach, used by SICER (Zang *et al.*, 2009), is "peak-merging": use a TF peak-caller, then merge nearby peak-calls to form estimates of modified regions. The downside of peak-merging is that it does not model large-scale properties of modified histone regions – individual histones are analysed in isolation, and we do not use information from surrounding regions. Moreover, peak-merging entails a number of arbitrary decisions, and it is not clear how these impact our results.

Other methods divide the genome into large bins and model the counts within. CCAT (Xu *et al.*, 2010) defines bins of width 1kb, categorizing them according to a signal-noise model. RSEG (Song and Smith, 2011) models bin count differences $T_i - C_i$ as emissions from a two-state Hidden Markov Model. Afterwards, boundaries between signal and noise bins are adjusted. These methods are useful, but discard the single base pair resolution that is often cited as an advantage of ChIP-seq.

1.6 Validation

Our understanding of binding events is still in its infancy and, since no gold standard exists, it is difficult to verify that a particular method outputs an "accurate" set of peak calls. Nevertheless, there are several methods that one can use to evaluate the accuracy of a given peak-caller.

There are two levels of validation:

Functional binding: TF binds to DNA in a given location, affecting nearby transcription (or some other process of interest).

Binding: TF binds to DNA in a given location, but need not have an effect on nearby transcription.

1.6.1 Functional binding site validation

If only ChIP-seq data are available, then there is no known way to show that a binding event has a functional effect, since ChIP-seq is designed to test for TF presence only. If, for example, we wish to show that a candidate binding event affects the transcription of a given gene, it is preferred to perform an independent experimental validation such as a luciferase reporter assay (Corbo *et al.*, 2010). However, this technique is extremely laborious for more than a handful of TFBS/gene pairs.

Ideally, we would like to find functional binding events on a genome-wide scale, through integration with differential expression data. Methods for this purpose are still in their infancy – see Section 1.9.1.

1.6.2 Binding site validation

To validate a (possibly non-functional) binding event, a common strategy is to use "ChIP-qPCR" – in this experiment, we investigate the pool of ChIPped DNA by using qPCR instead of sequencing. For example, Mortazavi *et al.* (2006) assessed 113 potential binding sites of the neural repressor NRSF/REST with ChIP-qPCR. The advantage of ChIP-qPCR is that focusing on a small set of loci gives us high statistical power. The main downside is that we must design and manufacture individual primers for each locus of interest – thus, time and cost issues usually restrict us to testing only a small number of loci.

A custom ChIP-chip experiment, using probes designed for the locations of interest, may also serve as useful validation for a ChIP-seq experiment. Microarray platforms are currently cheaper than sequencers, and are associated with better-developed analyses.

An important caveat is that ChIP-qPCR and ChIP-chip experiments are not independent of ChIP-seq, since all three rely on performing a ChIP experiment. Thus, we cannot control for ChIP biases through these validation techniques.

Usually, we do not want to go to the trouble of experimental validation until we are reasonably confident in the reliability of our ChIP-seq data and analysis. As such, we might also like to assess the quality of the ChIP-seq data without using independent experimental sources.

The spatial positions of our peaks can often be used to assess validity.

Many (but not all) TFs bind to known promoter or enhancer regions. Similarly, some histone marks localize to particular regions – for example, H3K4me3 localizes to promoter regions (Okitsu *et al.*, 2010). In these cases, we can map each peak-call to its nearest annotation feature (for example, using ChIPpeakAnno from Zhu *et al.* (2010)), and check that these mapping distances are sufficiently small.

1.6.3 Motif analysis

Peak-call validity can sometimes be assessed with motif analysis (for example, see Das and Dai (2007)). If the target TF has a well-characterized motif, then it should appear in a larger proportion of peak-call regions than it does in appropriately chosen "background" regions.

Alternatively, if one is extremely confident in a TF's motif, it can be used to filter peak-calls. ChIP-PaM (Wu *et al.*, 2010) calls peaks, and filters the resultant peak list with motif analysis. This strategy removes many "noisy" peaks but, as it can only find peaks with the known motif, it is clearly not appropriate for *de novo* regulatory element discovery.

1.6.4 Replication

A simple measure of accuracy is to replicate the experiment, and find the overlap between the peak sets obtained for each of our N replicates. (The manner in which we should replicate is discussed in Section 1.6.5.)

We will assume here that two peak-calls overlap if they share at least one base pair in common. This need not be the definition used – we can instead require that multiple base pairs must be held in common, or allow peak-calls to overlap if they lie within k base pairs of each other.

- For the two-way comparison, $N = 2$, we count the peaks in one set that overlap with at least one in the other set, and interpret this count – typically, as a percentage of the total number of peaks in one of the two samples, or as a Venn diagram.
- When $N > 2$, a naive approach is to consider all possible pairwise comparisons as before. However, these results are difficult to visualize. It is more common to take the union of all of the peak regions to form a "master" peak list, then find the overlap that each peak set has with the master list. This second strategy is usually easier to interpret, such as with a Venn diagram (see Figure 1.4).

  Asymptotically, taking the union in this manner is not a logical

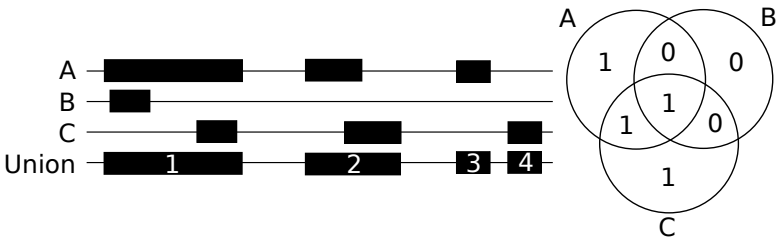

Figure 1.4 An example of assessing the overlap between 3 data sets: A, B and C. The black rectangles represent peak-calls. We have taken the union across data sets, obtaining 4 overall regions. We then compare each data set to this master list, as a Venn diagram. Note that this method is sensitive to peak width – wider peaks will result in fewer union peaks, leading to inflated overlap (observe the difference between peak 2, and peaks 3 and 4).

strategy. In the limit as $N \to \infty$, our consensus regions would tend to the entire genome, which is not ideal. As such, it is arguably better to include a base pair in our consensus regions if and only if that base pair lies within a peak for at least m of our N samples, for appropriate m. Note that this method assigns equal weight to each sample, which may not be appropriate if, for example, the samples are correlated.

When describing "overlap percentages", it is vital to explain exactly what method has been used – there are a number of arbitrary choices to be made, and different researchers will select different choices. Without these details, one cannot correctly interpret or replicate an analysis.

Note that overlap percentage is only an indirect method of assessing peak accuracy, and hence has serious limitations:

- It cannot detect any systematic biases in peak-caller output. For example, if our peak-caller always returns a particular artefact region as a peak-call, this will incorrectly inflate our overlap percentage.
- It is biased towards peak-callers with inflated peak-call widths, since these peaks are more likely to overlap. Thus, blindly attempting to maximise overlap percentage may conflict with our aim to find narrow, precise estimates that represent the actual peaks in the data.
- Its outcome heavily depends on the number of true binding events in the data. Suppose that there are T true binding sites, and our combined ChIP-seq experiment and peak-caller is 100% sensitive; that is, it always correctly identifies true binding sites. However, suppose that we always call F false peaks, located in uniformly-distributed random

genomic positions (thus, these false peaks are extremely unlikely to overlap across replicates). If we were to perform two such experiments, then the overlap percentage between them would be approximately

$$\frac{T}{T+F} = \left(1 + \frac{F}{T}\right)^{-1}$$

If T is much larger than F, then this overlap percentage will be close to 100%. However, if T is much smaller than F (that is, if our TF is specific only to a small number of sites) then the overlap percentage will be close to 0%. This is an example of the False Positive Paradox (Schoenfeld, 2007) – despite having 100% sensitivity, the TF is hard to detect when T is small, as we call many false positives. Therefore, a low overlap percentage does not imply a failed experiment.

- Our power to detect events is limited by the weakest sample. Ideally, we should be able to "pool" knowledge across samples, increasing our power. But if we demand that a binding event must be visible in both a high power and a low power analysis, then visibility in the low power analysis is the limiting step, wasting the high power analysis.
- It will usually underestimate overlap, because peak-calling applies an arbitrary threshold. Suppose that we sequence the same ChIP-seq library twice, and consider the overlap between these two data sets. There are many "peaks" that will just be on the threshold of being called, and many of these will be called in one library but not the other, by random chance. These will not contribute to the overlap percentage. (Bardet *et al.* (2012) describe this phenomenon, drawing a parallel with the "winner's curse" phenomenon that occurs in GWAS analysis.) We could potentially circumvent this problem by not thresholding – for example, if we were to collect posterior probabilities of enrichment throughout the entire genome, then we might be able to test for these values being similar between two replicates.

Alternatives exist to taking overlap percentages. Irreproducible Discovery Rate (IDR), described in full in Li *et al.* (2011), returns a set of peak-calls that are "common" to two lists, based on estimating the probability of observing a result in both lists.

Methods also exist to determine the overall similarity between two datasets; allowing us, for example, to investigate the hypothesis that two TFs bind to similar regions (Chikina and Troyanskaya, 2012).

As analysis of ChIP-seq improves, it should become possible to call

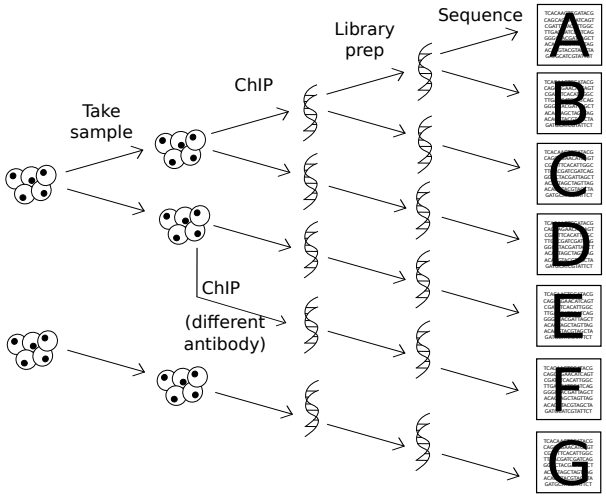

Figure 1.5 Different methods of replication. Data sets A-F are technical replicates of each other, since all of them have been obtained from the same population of cells. Data set G is a biological replicate of any of A-F, since it is derived from a different population of cells.

peaks as a function of multiple replicates, bypassing the need to look at overlap between peak-calls from individual samples.

1.6.5 Technical and biological replication

A ChIP-seq experiment can be repeated in multiple different ways, each with distinct statistical properties (see Figure 1.5).

Two broad classes of replication exist: "biological" and "technical".

For ChIP-seq, "biological replication" refers to taking multiple cell populations and performing ChIP-seq on each, independently. (In Figure 1.5, data set G is a biological replicate of data set A.) We usually find fewer consistent calls among biological replicates than among technical replicates, because of increased variability. However, our findings are "externally valid": their scope is not limited to a single population of cells. For this reason, biological replication is often most appropriate.

"Technical replication" refers to repeating one or more steps of a ChIP-seq experiment, but without using independent populations of cells. (In Figure 1.5, data sets A-F are technical replicates of each other.) Of course, repeating more steps increases the experiment's cost. Technical replicates allow us to find effects that are "internally valid": we

cannot assume that any results apply to other cell populations. Many different forms of technical replication are insufficient by themselves – for example, sequencing the same ChIPped library multiple times, or repeatedly making libraries from the same ChIPped sample. Together, these give us information about only one ChIP experiment, and so any errors introduced during that experiment cannot be corrected.

A preferable method of technical replication is to replace a step with a different but equivalent step. For example, have different personnel perform the ChIP, or use an alternative antibody. It may seem counter-intuitive that introducing additional variance in this manner can improve our analysis but, by doing so, we increase our confidence that the obtained results have scope that is not limited to one protocol.

A common example of technical replication is to use "polyclonal antibodies" – that is, a mixture of different antibodies that bind to different locations on the target TF (Lipman *et al.*, 2005).

The difference between external and internal validity can also be illustrated using Section 1.4.1's mixture framework. To investigate the distribution of M at a given site, we must collect hidden samples $\mu_1, \ldots, \mu_N$ from M, then sample individual observations x_{ij}, for $1 \leq j \leq J$, from the Poisson random variables $X_i \sim Pois(\mu_i)$. Suppose that, instead, we take a single sample μ_1 from M, then take multiple Poisson observations x_j, for $1 \leq j \leq J$, from $X \sim Pois(\mu_1)$. (This procedure corresponds to sequencing the same library multiple times.) Here, we learn a lot about μ_1, but we learn very little about the distribution of M, which is the variable of interest.

An additional problem with using biological replicates is that one must source multiple independent populations of cells. In particular, collecting clinical samples of tumour cells in sufficient quantities for ChIP is difficult, not only due to ethical/regulatory concerns, but also because samples are commonly preserved in a way that precludes standard ChIP protocols (Fanelli *et al.*, 2011). These issues can be circumvented with cell lines. A cell line is a population of cells that has been collected from primary tissue, and then artificially grown *ex vivo*. We may be able to infer biological information from studying these cell lines but we should be cautious as, through the growth procedure, cells are selected for their ability to grow, potentially introducing bias. Additionally, using multiple samples from one cell line is a form of technical replication, since the cells are originally from the same cell population. If possible, we should use samples from multiple cell lines, which constitutes biological replication.

Of course, there are situations in which biological replication is inap-

propriate: for example, if investigating heterogeneity within a given cell population, technical replication is required.

1.7 Assessing the reliability of peak-callers

Various methods can be used to show that a given peak-caller is reliable.

Spike-in data sets have been of great benefit in ChIP-chip analysis. A spike-in data set is created through taking an input sample, and artificially simulating IP enrichment by inserting a known quantity of DNA (Johnson *et al.*, 2008). We then assess whether or not an analysis can detect these changes in the data. Unfortunately, spike-in data are of little use for ChIP-seq, since it is difficult to generate DNA strands that resemble the "triangular" shapes expected from a successfully ChIPped binding event. Amplification of the spiked-in DNA fragments typically generates "rectangular" peaks. Due to this qualitative difference, it is not clear that spike-in data are appropriate for training peak-callers.

Simulated data sets are often used to validate statistical methods. However, there is again the question of whether simulated ChIP-seq data bear any resemblance to real ChIP-seq data, since we do not yet have a full understanding of the biases present within experimental data sets. An example ChIP-seq data simulator is Zhang *et al.* (2008b).

In the absence of a gold standard for ChIP-seq analysis validation, it is common to benchmark peak-callers by applying them to data sets in which peaks have been validated (see Section 1.6.2). For example, Laajala *et al.* (2009) benchmarked peak-callers using three quality measures: overlap percentage, motif presence, and ChIP-qPCR.

The scope of peak-caller assessments should always be considered carefully, especially if the number of data sets used is small. Rankings derived in one data set may not apply to others; in particular, authors may have presented only the data sets that best represent their own peak-caller.

1.8 Differential count-based strategies

For some time, peak-calling has been the predominant method of analysis for ChIP-seq data. However, a set of "peak-calls" is only a summary of a given data set, and may not retain information of interest. From a mathematical perspective, peak-callers are typically designed to assign a 0/1 binary value to each base pair, labelling it as either "bound" or

"not bound". (At best, we have a score representing how confident we are that a given location is a peak, and not noise.) This is arguably an oversimplification of the system under investigation. In reality, the library obtained from a ChIP-seq experiment is an "average" over all of the cells present in the sample at the time of cross-linking.

Let us illustrate this point by considering a simple hypothetical example, removing all sources of noise other than intercellular variation. Suppose that we use N cells in a ChIP-seq experiment, and that they are sufficiently homogeneous that they share the same set of potential TFBSs, $\{s_1, \ldots, s_K\}$. At the time of cross-linking, each site s_i is TF-occupied in any given cell with some probability p_i. Assuming a perfect antibody, each TF-occupied site will contribute a fragment to the DNA pool obtained after the IP step.

Thus, if we could observe the hidden variable Z_i, the number of pre-amplification fragments that contain site s_i, we would expect it to have distribution $Z_i \sim Bin(N, p_i)$, approximated as $Z_i \sim Pois(Np_i)$, since N is large. Samples from Z_i can give us information about the probabilities p_i. A higher p_i does not prove that the site is more likely to be "functional" (since binding need not have an effect), but we might expect p_i to be correlated with functionality.

In summary, the immunoprecipitated DNA sample reflects not only the locations of TF-DNA interactions that are occurring in the data, but also the frequency with which those interactions occur. These frequencies tell us about the activity of a binding site. Since most peak-callers do not consider this complexity, we lose information when peak-calling.

Peak-calling strategies also introduce complications when we compare a number of samples, or combine information across replicates (since it is difficult to combine peak-calls and meaningfully summarize their scores). Therefore, we can usually get a lot more information about a given binding site by returning to the raw aligned reads.

To this end, there is an increasing tendency towards using differential count analysis, normally used in the context of assessing differential expression in RNA-seq data. In RNA-seq, we expect reads to fall in pre-defined genomic (or transcriptomic) regions. Equivalent annotational resources are being developed for ChIP-seq (for example, **?**) – however, these resources can only be as reliable as the tools that we use to analyse our ChIP-seq data, and we have limited methodology available at present.

1.8.1 Analysis protocol

A differential count-based ChIP-seq analysis typically consists of 5 steps:

1. Define "bins" (locations of interest).
2. Shift reads by half the average fragment length.
3. Count the number of reads that fall into the bins, for each sample.
4. Estimate appropriate normalization factors.
5. Run a differential count algorithm on the count data.

Step 2 was discussed in Section 1.3.8. Step 3 is trivial, and step 5 was addressed in Chapter **??** – appropriate algorithms include BaySeq (Hardcastle and Kelly, 2010), EdgeR (Robinson *et al.*, 2010) and DESeq (Anders and Huber, 2010). We therefore address steps 1 and 4.

Consider step 1. An obvious strategy is to define bins based on peak-calls: if we have N ChIP-seq samples, each with an associated control, we simply call peaks in each data set. Then, define consensus regions, as we did in Section 1.6.4. This strategy suffers from two main disadvantages: firstly, we inherit all of the noise associated with peak-calling multiple times. Secondly, since peak-calls have been generated by selecting regions of high enrichment, the counts in these bins will be biased towards larger values. This could have implications when normalizing, or when selecting prior distributions for differential count algorithms.

If we know which genomic features our TF binds to, we can simply define bins about these features. Alternatively, we can select bins that tile the genome, provided that analysis remains computationally tractable.

Now consider step 4 – normalization. Count normalization algorithms that are used in RNA-seq analysis typically assume that a large proportion of sites do not show differential read counts between conditions (Robinson and Oshlack, 2010). This assumption may be invalid for many ChIP-seq experiments – for example, if we define bins based on peak-calls, as described above. Even if we define bins based on annotation features, the assumption may still be invalid if there are many TFBSs in the vicinity of these features, resulting in a large proportion of bins containing ChIP signal. This effect is especially problematic in many histone mark data sets, where we expect to see widespread enrichment across the genome. We should be careful to use a normalization method that is appropriate for the data.

Two Bioconductor packages that facilitate ChIP-seq differential count analyses are *DiffBind* (Ross-Innes *et al.*, 2012) and *Repitools* (Statham *et al.*, 2010).

1.9 The future of ChIP-seq

The world of sequencing is changing rapidly, and our algorithmic strategies should adapt to accommodate these changes.

In the short term, sequencing technology is becoming faster and cheaper at a dramatic pace. As such, there will likely be a corresponding increase in depth available to us. This will afford us greater power to detect "weaker" binding events, compounding the shift of focus towards quantifying binding events, rather than simply calling regions.

Of course, there may be significant changes to the ChIP protocol itself.

As an example, our fragment length modelling methods (from Section 1.3.8) are highly dependent on the fragment length size selection procedure. So far, gel electrophoresis has been the usual method of size selection. However, Illumina have recently changed their sequencing library preparation protocol, now using the Nextera kit. This replaces the gel-based size selection step with a bead-based method. Libraries prepared with Nextera have substantially different fragment length distributions to libraries prepared using previous technologies (Adey *et al.*, 2010), and so we may need to re-evaluate our assumptions about fragment length when analysing these libraries, especially if comparing with earlier non-Nextera experiments.

In the long term, a new generation of sequencing machines are looming on the horizon – in particular, Oxford Nanopore (Eisenstein, 2012) claim that their machine can fully sequence large DNA fragments. (In ChIP-seq, since the DNA fragments are relatively small, it is already possible to sequence them in their entirety with current paired-end sequencing protocols. However, to do this, reads must be extended and this comes at a cost, since sequencers must be run for additional cycles; typically, the minor improvement in data quality is not considered to be worth the additional cost.) If whole-fragment sequencing becomes widespread, we will no longer have to extrapolate fragment locations from read positions. This will likely improve low-level analysis, such as alignment, but is unlikely to alter downstream ChIP-seq algorithms; aligned fragments can be treated as arbitrarily long paired-end reads.

Error introduced through the need to model fragment length may be reduced through the use of the modified protocol "ChIP-exo" (Rhee and Pugh, 2011). In this experiment, DNA fragments are trimmed to their smallest possible size while still crosslinked to the TF of interest. Thus, in theory, we have greater precision when finding binding sites.

In RNA-seq, we are beginning to see the establishment of single-cell se-

quencing (Tang *et al.*, 2009). Current ChIP-seq protocols cannot do this, because the DNA present in one cell is insufficient for ChIP. However, in the future, we may succeed in collecting single-cell binding information. When performing ChIP-seq on a population of cells, we obtain an average measure of affinity. If we were able to pull out TFs on a single-cell level, this averaging step would be removed and we would obtain an indicator of binding in each cell – usually ternary data (no binding, binding on one chromosome, binding on both chromosomes), except in cells with abnormal copy number. One benefit of such an experiment is that it would provide information on binding events that occur simultaneously. Though we may not be able to achieve the single-cell limit in the near future, we may be able to approach it. The amount of starting material required for ChIP is decreasing rapidly – 10^7 cells are required for most current ChIP protocols, but techniques are emerging that are claimed to require of the order of 10^5 cells – for example, Chromatrap (Bryant, 2012). In particular, reducing the number of required cells permits more precise assessments of heterogeneity. For example, we could collect 5 different samples from a tumour, and find out if the binding activity of a given TF varies between samples.

Correspondingly, as ChIP-seq protocols become more efficient, our ability to perform multiple ChIP-seq experiments simultaneously ("high-throughput ChIP") is also improving. An important consequence of this is the capability to screen many TFs, in the case where we do not know which ones are responsible for an observed effect.

1.9.1 Integrating ChIP-seq with expression data

One of our aims, as defined at the start of this chapter, was to combine ChIP-seq data with expression data, identifying any transcriptional effects of binding events. Since ChIP-seq analysis is still evolving, there is no consensus on how to do this. A common technique is to peak-call in the ChIP-seq data, map each peak to the nearest gene, and then look at which DE genes have associated peaks. This approach typically suffers from an inability to quantify the ChIP-seq signal – a gene's ChIP-seq status is binary ("mapped peak" or "no mapped peaks") and thus it is difficult to rank targets or assess significance. We might try counting the number of mapped peaks, or summing peak-call widths, but it is difficult to justify the biological significance of these properties.

The recent Bioconductor package *Rcade* circumvents these issues by adopting a Bayesian approach, using *baySeq* (Hardcastle and Kelly,

2010) to perform count-based ChIP-seq analysis, as per Section 1.8.1, and integrating the results with a differential expression analysis. Thus, we can rank genes by their posterior probability of being enriched for both ChIP-seq and DE.

High-throughput sequencing is providing the opportunity to tackle key unanswered questions in transcription regulation and, as such, we expect this exciting field to be a fertile ground for novel research in the upcoming years.

References

Adey, Andrew, et al. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology*, **11**, R119.

Aird, Daniel, et al. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, **12**, R18.

Anders, Simon, and Huber, Wolfgang. 2010. Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106.

Bannister, Andrew J, and Kouzarides, Tony. 2011. Regulation of chromatin by histone modifications. *Cell Research*, **21**, 381–395.

Bardet, Anaïs F, et al. 2012. A computational pipeline for comparative ChIP-seq analyses. *Nature Protocols*, **7**, 45–61.

Barski, Artem, et al. 2007. High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–37.

Beck, Dominik, et al. 2012. Signal analysis for genome wide maps of histone modifications measured by ChIP-seq. *Bioinformatics (Oxford, England)*, 1–8.

Benjamini, Yuval, and Speed, Terence P. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 1–14.

Boyle, Alan P, et al. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics (Oxford, England)*, **24**, 2537–8.

Bryant, Robert Rusty. 2012. Microplates, assay reagents, screening consumables, and kits. *Journal of Biomolecular Screening*, **17**, 550–2.

Buck, Michael J, and Lieb, Jason D. 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349–360.

Cairns, Jonathan, et al. 2011. BayesPeak - An R package for analysing ChIP-seq data. *Bioinformatics (Oxford, England)*, 1–2.

Chen, Yiwen, et al. 2012. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods*, **9**, 1–9.

Cheung, Ming-Sin, et al. 2011. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Research*, **39**, e103.

Chikina, Maria, and Troyanskaya, Olga. 2012. An effective statistical evaluation of ChIPseq dataset similarity. *Bioinformatics*, **28**, 607–13.

Corbo, Joseph C, et al. 2010. CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Research*, **20**, 1512–1525.

Das, Modan K, and Dai, Ho-Kwok. 2007. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, **8 Suppl 7**, S21.

De Wit, E, and De Laat, W. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes & Development*, **26**, 11–24.

Dubik, D, and Shiu, R P. 1992. Mechanism of estrogen activation of c-myc oncogene expression. *Oncogene*, **7**, 1587–1594.

Durbin, R, et al. 1998. *Biological Sequence Analysis*. Cambridge University Press.

Eeckhoute, Jérôme, et al. 2006. A cell-type-specific transcriptional network required for estrogen regulation of cyclin D1 and cell cycle progression in breast cancer. *Genes & Development*, **20**, 2513–2526.

Eisenstein, Michael. 2012. Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology*, **30**, 295–296.

Fanelli, Mirco, et al. 2011. Chromatin immunoprecipitation and high-throughput sequencing from paraffin-embedded pathology tissue. *Nature Protocols*, **6**, 1905–19.

Fisher, B, et al. 1989. *A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors.* Tech. rept. National Surgical Adjuvant Breast and Bowel Project (NSABP) Headquarters, Pittsburgh, PA 15261.

Gry, Marcus, et al. 2009. Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics*, **10**, 365.

Hansen, Kasper D, et al. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, **38**, e131.

Hardcastle, Thomas J, and Kelly, Krystyna A. 2010. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.

Heinz, Sven, et al. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, **38**, 576–589.

Hilbe, Joseph M. 2011. *Negative Binomial Regression*. Vol. 70. Cambridge University Press.

Hower, Valerie, et al. 2010. Shape-based peak identification for ChIP-Seq. *BMC Bioinformatics*, **12**, 12.

Hurtado, Antoni, et al. 2011. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nature Genetics*, **43**, 27–33.

Johnson, David S, et al. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, **316**, 1497–502.

Johnson, David S, et al. 2008. Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Research*, **18**, 393–403.

Kaufmann, Kerstin, et al. 2009. Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower. *PLoS Biology*, **7**, e1000090.

Kidder, Benjamin L, et al. 2011. ChIP-Seq: technical considerations for obtaining high-quality data. *Nature Immunology*, **12**, 918–922.

Laajala, Teemu D, et al. 2009. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics*, **10**, 618.

Lee, Tong Ihn, et al. 2006. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nature Protocols*, **1**, 729–748.

Li, Guoliang, et al. 2010a. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biology*, **11**, R22.

Li, Ning, et al. 2010b. Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods San Diego Calif*, **52**, 203–212.

Li, Qunhua, et al. 2011. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, **5**, 1752–1779.

Lipman, Neil S, et al. 2005. Monoclonal versus polyclonal antibodies: distinguishing characteristics, applications, and information resources. *ILAR journal / National Research Council, Institute of Laboratory Animal Resources*, **46**, 258–68.

Marioni, John C, et al. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, **18**, 1509–17.

Mo, Qianxing, and Liang, Faming. 2010. A hidden Ising model for ChIP-chip data analysis. *Bioinformatics*, **26**, 777–783.

Mohibi, Shakur, et al. 2011. Mouse models of estrogen receptor-positive breast cancer. *Journal of Carcinogenesis*, **10**, 35.

Morstyn, G, et al. 1983. Bromodeoxyuridine in tumors and chromosomes detected with a monoclonal antibody. *The Journal of Clinical Investigation*, **72**, 1844–50.

Mortazavi, Ali, et al. 2006. Comparative genomics modeling of the NRSF/REST repressor network: from single conserved sites to genome-wide repertoire. *Genome Research*, **16**, 1208–1221.

Myers, Richard M, et al. 2011. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biology*, **9**, 21.

Okitsu, Cindy Yen, et al. 2010. Transcriptional activity affects the H3K4me3 level and distribution in the coding region. *Molecular and Cellular Biology*, **30**, 2933–2946.

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

Qin, Zhaohui S, et al. 2010. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, **11**, 369.

Rhee, Ho Sung, and Pugh, B Franklin. 2011. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell*, **147**, 1408–1419.

Robertson, Gordon, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, **4**, 651–7.

Robinson, Mark D, and Oshlack, Alicia. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**, R25.

Robinson, Mark D, et al. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Ross-Innes, Caryn S, et al. 2010. Cooperative interaction between retinoic acid receptor-alpha and estrogen receptor in breast cancer. *Genes & Development*, **24**, 171–82.

Ross-Innes, Caryn S, et al. 2012. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–93.

Schmidt, Dominic, et al. 2009. ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods (San Diego, Calif.)*, **48**, 240–8.

Schoenfeld, Alan H. 2007. *Assessing Mathematical Proficiency MSRI Publications Volume 53, 2007*. Vol. 53. Cambridge University Press.

Shou, Jiang, et al. 2004. Mechanisms of tamoxifen resistance: increased estrogen receptor-HER2/neu cross-talk in ER/HER2-positive breast cancer. *Journal Of The National Cancer Institute*, **96**, 926–935.

Sims, Robert J, and Reinberg, Danny. 2009. Processing the H3K36me3 signature. *Nature genetics*, **41**, 270–1.

Smith, Ian E, and Dowsett, Mitch. 2003. Aromatase inhibitors in breast cancer. *The New England Journal of Medicine*, **348**, 2431–42.

Song, Qiang, and Smith, Andrew D. 2011. Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics*, **27**, 870–871.

Spyrou, Christiana, et al. 2009. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC Bioinformatics*, **10**, 299.

Statham, Aaron L, et al. 2010. Repitools: an R package for the analysis of enrichment-based epigenomic data. *Bioinformatics*, **26**, 1662–1663.

Tang, Fuchou, et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, **6**.

Tannock, Ian F, and Hill, Richard P. 1998. *The Basic Science of Oncology*. Third edit edn. McGraw-Hill.

Valouev, Anton, et al. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, **5**, 829–834.

Visel, Axel, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.

Wei, Chia-Lin, et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell*, **124**, 207–219.

Weinberg, Robert A. 2007. *The Biology of Cancer*.

Wu, Song, et al. 2010. ChIP-PaM: an algorithm to identify protein-DNA interaction using ChIP-Seq data. *Genes & Development*, **7**, 18.

Xu, Han, et al. 2010. A Signal-Noise Model for Significance Analysis of ChIP-seq with Negative Control. *Bioinformatics (Oxford, England)*, **26**.

Zang, Chongzhi, et al. 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics (Oxford, England)*, **25**, 1952–8.

Zhang, Xuekui, et al. 2011. PICS: probabilistic inference for ChIP-seq. *Biometrics*, **67**, 151–163.

Zhang, Yong, et al. 2008a. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, **9**, R137.

Zhang, Zhengdong D, et al. 2008b. Modeling ChIP Sequencing In Silico with Applications. *PLoS Computational Biology*, **4**, 10.

Zheng, Wei, et al. 2011. Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics*, **12**, 290.

Zhu, Lihua J, et al. 2010. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, **11**, 237.