Title
Statistical Aspects of ChIP-Seq Data Analysis

Author
Mayba, Oleg Sergeyevich

Publication Date
2011

Peer reviewed|Thesis/dissertation

Statistical Aspects of ChIP-Seq Data Analysis

by

Oleg Sergeyevich Mayba

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

and the Designated Emphasis
in

Computational and Genomic Biology

in the
Graduate Division
of the

University of California, Berkeley

Committee in charge:

Professor Terence P. Speed, Chair
Professor Sandrine Dudoit
Professor John Ngai

Spring 2011

Abstract

Statistical Aspects of ChIP-Seq Data Analysis

by

Oleg Sergeyevich Mayba

Doctor of Philosophy in Statistics
and the Designated Emphasis
in
Computational and Genomic Biology

University of California, Berkeley

Professor Terence P. Speed, Chair

ChIP-Seq experiments combine the recently developed next-generation sequencing technology with the established chromatin immunoprecipitation assays to study the interactions between various classes of proteins and DNA in the cell nucleus. The experiments consist of isolating the protein-DNA complexes from the nucleus, enriching the pool of DNA fragments for those bound to the protein of interest, and sequencing the resulting pool of fragments, producing millions of short reads that can be aligned to the genome. Despite the fact that the ChIP-Seq technology has been developed very recently, a great number of studies have been carried out on the DNA binding of a variety of transcription factors in different species and tissue types. ChIP-Seq approaches have also been used to study cellular epigenomic states such as histone modifications.

As with any nascent technology, a number of methodological issues need to be addressed before a proper data analysis pipeline for ChIP-Seq can be established. Some of the issues that need to be addressed are image processing and analysis, alignment of the reads to a genome or a subset of it, and identifying the signal sites along the genome. This work focuses on the issue of signal identification, the problem known as peak-finding in the literature.

We describe the data-generating process for ChIP-Seq experiments and review properties of the data and various sources of biases in Chapter 1. We then review various approaches to peak-finding in Chapter 2. We provide a detailed overview of some common strategies, their relative advantages and disadvantages, and describe the statistical models used by some popular peak-finding tools. We formalize the conceptual framework of peak-finding by introducing the notions of enrichment measures and enrichment statistics and categorize various peak-finders in terms of this framework. We discuss in some detail the different kinds

of control samples used in ChIP-Seq experiments, and how they are incorporated into the peak-finding procedure. We also address the important issue of validation in the context of ChIP-Seq experiments and the shortcomings of the currently available validation approaches.

In Chapter 3 we propose a novel peak-finding strategy for experiments involving trancription factor binding that lack appropriate control samples (so-called one-sample experiments). Our approach accounts for genomic sequence biases in the data, namely the GC and mappability effects, and utilizes the knowledge of the shape of the read density profile in the vicinity of the true binding sites. We use deduced sets of true positive and true negative enriched regions to demonstrate that our approach is better at removing non-specifically enriched regions from the set of identified binding sites than other one-sample approaches and provides a superior spatial resolution to most examined peak-finders.

Finally, in Chapter 4 we discuss the important issue of combining data from replicate samples. We discuss different kinds of replicates common in the ChIP-Seq literature and the standard approaches used to integrate data across replicates. We develop several diagnostic plots for assessing whether the standard assumption of Poisson variance holds and observe that the assumption can break down even for technical replicates due to flow cell-specific sequence composition effects.

Acknowledgments

I would like to thank:

My advisor Terry Speed for his support, encouragement and invaluable advice. I feel very lucky to have shared in your wisdom and would like to think that I am a better statistician and a better person for it.

All the wonderful students, postdocs and faculty in the department who have been extremely helpful in making sense of genomics in general and sequencing in particular: Sandrine, Elizabeth, Yuval, Kasper, Margaret, Jim, Su Yeon, Pierre, Henrik, Stephen, Xiaoyue, Leath and many others. Thank you for all the insights and all the stimulating conversations over the years. I have learnt a lot from you.

Dale Leitman and Xiaoyue Zhao, for getting me started with ChIP-Seq.

My biologist collaborators, Wally Wong and Yu Zhang, for their knowledge, their data and for being so very patient.

My fellow stat newbies: Harry, Charlotte, Irma, Moorea, Daisy, Jing, and Allan. Thank you for all the good times and for all the group therapy sessions we've had. I do not think I would have gotten through my first years without you.

The great support stuff of the department. Angie - thank you for helping me figure out all the intricacies of the system and for being a calming influence in times of panic. Phil and Ryan - thank you for all the help you've given me over the years, I'm a better programmer because of you.

My dissertation committee: Terry Speed, Sandrine Dudoit, and John Ngai. Thank you for all your comments and advice - you've helped to make this dissertation readable.

All the teachers I've worked with for their teaching advice, and all of my students for helping me overcome my fears, laughing at my jokes (intended or otherwise) and letting me help you.

My fellow bears: Alex, Harry,and Jeff for all the shared excitement and heartbreak. There's always the next year! GO BEARS!

My dear freinds, especially Neil and Joel, for reminding me that there is life outside of Evans.

My family for their support, their food, and their love. In answer to your question: yes, it IS done.

My awesome and wonderful Effie for being π-tastic and for putting up with more than is right.

To Effie

Contents

List of Figures

List of Tables

Chapter 1

Introduction to ChIP-Seq

1.1 High-Throughput Sequencing and its Applications

Over the last few years a variety of DNA-sequencing platforms have emerged that are commonly referred to as high-throughput sequencing (HTS) or next-generation sequencing (NGS) technologies. The primary difference between these platforms and the older well-established Sanger capillary sequencing is the drastic increase in the number of DNA molecules sequenced simultaneously at the expense of the length of the high-quality portion of the sequenced molecule. The overall effect of the larger number of sequenced reads of shorter length is the many-fold increase in the amount of sequence output by the dedicated machines both in total quantity and in terms of output per machine run and per time required for the run to complete. This increase is achieved through a variety of proprietary chemistry techniques and some of the commonly used HTS platforms are Roche 454, Illumina, ABI SOLiD, and Helicos. The platforms differ in many respects, including the number of reads per run and the read length attained, the time it takes to run a sample through the pipeline, the costs associated with the sample preparation, sample run, and purchasing and maintaining the sequencer, and the data pre-processing required before statistical analysis can take place.

In addition to traditional applications of large scale DNA sequencing (e.g. genome sequencing and assembly) the HTS technologies are now used extensively in biological assays previously done with microarrays. This earlier technology relies on the reverse complementary hybridization property of nucleic acid molecules to interrogate the DNA sample under study for the presence of molecules of interest by means of millions of carefully chosen single-stranded polynucleotide probes covalently bound to the surface of the chip (microarray). The DNA sample is denatured, giving rise to the starting pool of single-stranded targets. The hybridization of the target to the probe results in the emission of a fluorescent signal and the intensity of the combined emission of many copies of the probe is a measure of the abundance of the corresponding target molecule in the sample. With the development of HTS it

has become possible to assess the composition of the sample under study directly, without relying on designed probes, by sequencing millions of the DNA fragments in the sample simultaneously. Depending on the choice of the platform and the length of DNA fragments in the sample, one might not be able to sequence the entire fragments due to limitations on the length of the sequenced reads. Thus, often only one end of the fragment is sequenced. Some platforms allow for sequencing of both ends of the fragment in what is known as paired-ends (PE) sequencing and if the combined length of the two sequenced ends exceeds the fragment length, then the sequence of the entire fragment is determined.

Some of the aplications of both microarray and HTS include:

- Differential gene expression. The pool of mRNA representing a transcriptome is reverse-transcribed to cDNA and its composition is assessed by microarray hybridization or sequencing. The expression measures are then compared between 2 or more samples.
- Protein-DNA interactions. The starting pool of DNA is enriched for fragments from loci of interaction.
- Copy number variation in tumour cells.
- SNP identification.
- Methylation state of the DNA.

and many others.

For a large number of these applications, there has been a technological shift from microarrays to HTS over the last few years among the scientific community. Some of the advantages of HTS include [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]:

- Lower cost per assay. This does not take into account the cost of the sequencer itself and there are dedicated sequencing facilities for those scientists not wishing to purchase an in-house machine.
- No need to design sample- and assay- specific polynucleotide probes. In practice, this means no apriori knowledge of the composition of the sample is required.
- Greater ability to interrogate repetitive regions of the genome. A single nucleotide difference between two repetitive regions can be detected with direct sequencing and the correct region identified, while the probe cross-hybridization makes this task much more difficult in the microarray setting.
- Digital (count) and not analogue (fluorescence intensity) measure of target presence. The former is seen as less prone to saturation effects and as being able to produce higher resolution results.

- Smaller amount of required starting material.

The big disadvantage of using HTS for the well-established biological assays like differential gene expression has been the lack of accepted statistical analysis methodology. New tools have been developed to deal with both the digital nature of the data and the platform-specific biases it exhibits. In particular, issues of pre-processing, normalization and statistical significance for individual applications are being currently addressed, with no universally accepted pipeline having emerged for most applications. However, major results have been achieved using currently available tools and it is clear that transition from microarray to HTS platforms for most common applications is inevitable.

1.2 Illumina Genome Analyzer

One of the most commonly used HTS platforms right now is Illumina and its Genome Analyzer (GA) line of sequencers. The GA2 sequencer can produce up to 640 million reads of up to 150 bp in length (including 2×150 for paired-end reads) per run. Depending on the application, shorter reads might suffice and the smaller number of reads may still provide sufficient sequencing depth to answer the biological question of interest.

Prior to being sequenced, the DNA sample undergoes library preparation intended to optimize the number of output reads. The process involves attaching asymmetric sequencing adapters to the double-stranded DNA fragments and, depending on the nature of the assay, subjecting the sample to several rounds of PCR amplification. The PCR step is intended to increase the size of the sample when the amount of the starting material is too small. It should be pointed out, however, that the complexity of the library is determined by the starting material and can only be decreased by PCR (due to PCR-specific biases). The library preparation also usually involves the step of *size selection*, in which the DNA fragments in the range optimal for sequencing are isolated. It is important that the initial sample contains a large portion of fragments in that range and this is usually achieved by appropriate fragmentation techniques (e.g. chromatin shearing in case of ChIP-Seq). Finally, the DNA fragments are denatured and single-stranded DNA is loaded onto the sequencer.

The actual sequencing takes place on a planar surface called *flow cell*. This surface is divided into 8 *lanes* that the libraries are loaded onto. Each lane usually contains DNA fragments from a single library, although several libraries (each with its own unique identifying adapter or bar-code) can share a single lane (this is know as 'multiplexing')[3]. The lane is covered with covalently-bound sequencing adapters that are complementary to those attached to DNA fragments during library preparation. After DNA fragments are anchored to the surface of the flow cell, they are amplified by *PCR bridge amplification* to give rise to *clusters* of identical fragments, localized to small neighborhoods on the flow cell surface. The actual sequencing is done by synthesis, employing reversibly-terminated nucleotides with fluorescent tags. This approach creates a reverse-complement of the original single-stranded DNA fragment one nucleotide at a time, the identity of which is established by analyzing

the light emissions, resulting in a sequenced *read*, which is a tag identifying the original fragment. At each sequencing stage, the 4 modified nucleotides are added to the flow cell and are incorporated into the growing reads. Only 1 nucleotide can be incorporated at a time due to the blocking modifications. The images of the flow cell are taken for further processing, the unincorporated nucleotides are washed away and the blocking modifications are removed, followed by the next round of sequencing. Optionally, the other end of the fragment can be sequenced as well when the paired-ends chemistry is employed.

The images of the flow cell are then analyzed to identify the clusters and the identity of the resulting reads. It is the need for a strong signal to be picked up by the image analysis software that necessitates cluster formation in the first place, as a single-fragment signal would be too weak. For each read, the image analysis also provides the base-calling quality scores for each nucleotide. These scores are based on relative strengths of intensities corresponding to 4 different nucleotides and reflect the confidence that can be placed in the identity of the called nucleotide.

Finally, the resulting set of reads is aligned to the genome or some subset of it, such as the transcriptome, for further analysis in most applications (any de-novo assembly being the notable exception). This alignment is done using one of the existing short read aligners: ELAND (unpublished proprietary aligner that is part of the Illumina pipeline), MAQ [11], SOAP2 [12], BWA [13], Bowtie [14] or one of many other available options. Since each cluster represents only 1 strand of the original DNA fragment, the resulting read can be either a subsequence or the reference strand (usually called *positive-strand read*) or it could be a subsequence of the reverse complement of the reference strand (usually called *negative-strand read*). Thus, the output of the alignment program usually contains not only information about the location of the aligned read in the genome but its strandedness as well, which is a crucial piece of information for some applications. The strandedness is also important when aligning the paired-end reads to preserve proper orientation of the fragment ends.

1.3 ChIP-Seq

Chromatin immuno-precipitation (ChIP) is a well-established biological assay to identify the sites of DNA-protein interactions. The most commonly studied types of interactions are the binding of transcription factors and other transcriptional machinery (e.g., polymerase) and DNA packaging (in particular, histone modifications that affect chromatin state).

A ChIP protocol starts with a sample of cells, to which a cross-linking agent (usually, formaldehyde) has been added. The effect of cross-linking is to covalently bind the DNA to the associated proteins. The cells are then lysed, chromatin is isolated and is sheared into smaller fragments using hydro-shearing, sonication or some other technique. This pool of chromatin fragments is then probed with an antibody specifically targeting the protein of interest. This is often done by letting magnetic beads coated with the antibody interact with the chromatin fragments. This use of antibody gives the assay its name. The resulting

chromatin sample is therefore enriched for DNA fragments that are bound to the protein under study. Finally, the cross-links are reversed and the purified DNA is extracted.

The investigator now has a sample of DNA that is enriched for the loci of interest relative to the genomic background. To actually identify the interaction sites, the sample is either applied to a microarray spotted with probes for candidate loci (so-called ChIP-chip experiments) or is sequenced on one of HTS platforms and the resulting reads are aligned to the genome or a subset of it (so-called ChIP-Seq experiments [1, 2]).

The actual identification of interaction sites is done by dedicated tools, called peak-finders that are looking for genomic locations that are enriched in sequenced reads relative to the background. The approaches to peak-finding are described in detail in Chapter 2.

1.4 Properties of Illumina-Sequenced ChIP-Seq data

1.4.1 Chromatin Shearing and Genome Structure

The first step in the library prepartion of ChIP-Seq samples after the cross-linking of the protein is chromatin shearing. It is a well known fact that some regions of chromatin are more open than others in their native state and thus are more amenable to shearing. These regions tend to be overrepresented in the starting DNA sample due to the size selection step, which eliminates very large fragments [15, 4, 16, 7, 10]. One of the examples of this effect is the overrepresentation of promoters of actively transcribed genes in input DNA samples (sheared chromatin that does not undergo IP), and, since the state of chromatin depends on the experimental conditions, some authors find correlation between gene expression and read density in input DNA samples [17, 18]. Thus even the starting DNA sample for ChIP-Seq data is biased relative to genomic background and a proper adjustment should be made to account for it. There is also some evidence that chromatin structure affects the composition of ChIP-Seq samples through mechanisms independent of shearing, though their exact nature remains unknown [18]. There are several different shearing methods available, with some authors suggesting that acoustic fragmentation should be preferred to usual nebulization [19] and others suggesting that digestion with micrococcal nuclease can be used instead of cross-linking for stable protein-DNA interactions [3].

Another commonly observed phenomenon is enrichment at telomeric and centromeric regions of the genome [16, 18]. This effect might be due to overrepresentation of these regions in the original DNA sample or to the existence of TF-binding hotspots in those regions and appears to be independent of PCR amplification step of the library preparation [16].

1.4.2 Enrichment/Antibody

An important feature of the ChIP protocol is that it is an enrichment and not a purification assay. Thus, while it enriches the DNA sample for fragments associated with the protein of interest, it does not result in a sample consisting exclusively or even mostly, of such fragments. In our experience, about 20% or less of the sequenced reads correspond to the signal sites, with the vast majority of reads coming from the background fragments (other authors report similar fractions: 10-20% ([1]), 5%-20% ([2]), 1% ([16]), 15% ([7]); however this fraction might be higher for settings other than transcription factor binding (46% for Pol II [7]). The actual fraction of the reads corresponding to signal depends on the number of binding sites in the genome and the efficiency of the antibody. In any case, the number of signal reads cannot exceed twice the number of fragments bound to protein of interest in the starting DNA sample (each double-stranded DNA fragment can contribute up to 2 reads once deposited onto a flow cell).

The choice of the antibody is an important issue [8]. The ideal antibody should be specific to the protein under study and sensitive enough to pull out most fragments associated with this protein. If the antibody is not sensitive, some potentially important signal will be missed. If the antibody is not specific (e.g., it is targeting an epitope common to some other proteins), then the resulting enrichment signal will be some convolution of several protein-binding profiles.

1.4.3 Alignment Issues

Alignment programs often have to deal with ambiguities when trying to align the non-PE reads to the genome. An ambiguity arises when a read aligns to multiple places with the same alignment score, which is some function of the number of mismatches, insertions, deletions, and sometimes the base-calling quality scores. A decision has to be made about the assignment of such read (often called a *multi-read*) to the genome and the usual strategy is to not assign the read at all if the best alignment is not unique (here 'best' can refer to an interval of alignment scores, e.g. if the investigator is willing to treat 1 and 2 mismatches equally)[20]. Some alternative approaches involve assigning a multi-read to one of its best alignments at random or with the probabilities given by the numbers of unambigously assigned reads in some neighborhood of the candidate locations.

When adopting the standard practice of discarding multi-reads, one encounters the problem of *non-mappable* locations in the genome [16, 7, 17]. We define (strict) mappability $M(l,k)$ to be a function of genomic location l and read length k, such that $M(l,k) = 0$ if the k bp subsequence of reference strand starting at position l or the reverse complement of this subsequence occur anywhere else in the genome, and $M(l,k) = 1$ otherwise (with a caveat that a location with a corresponding k bp subsequence that is a reverse complement of itself is non-mappable since it cannot be determined which end of the fragment, 5' or 3' relative to reference strand, the read represents). In other words, some reads can never be aligned

uniquely to the genome no matter how one defines the alignment scores because their exact match appears in the genome more than once, and therefore the locations of these matches will never have a read aligned to them.

1.4.4 Repeat Regions Enrichment

Enrichment in repetitive regions of the genome, especially for satellite repeats, is often observed in ChIP-Seq experiments [2, 16, 5, 21, 17]. Some of this might be due to the inadequacy of the reference sequence, as the highly repetitive regions may be collapsed into a single copy in the reference assembly due to lack of information for their proper resolution [8, 10]. In practice this may result in read pile-ups along the genome in locations that are ought to be non-mappable but cannot be recognized as such due to assembly deficiencies. Alternatively, this enrichment might reflect the true copy number variation of the sequenced genome relative to the reference sequence [4, 17, 8, 22, 10].

1.4.5 GC Bias

Illumina sequencing technology is a subject to a well-known GC bias, that results in sequenced reads being more GC-rich than the starting DNA sample [23, 19, 17, 18]. This bias is not unique to any particular assay (e.g. mRNA-Seq or ChIP-Seq) but is prevalent in a variety of applications. Some authors have suggested modifications to the library preparation protocols involving a size selection step that aim to reduce the bias, but these suggestions have not been widely adopted yet [19]. Some authors argue that this bias is introduced by the PCR amplification of the sample and that eliminating the PCR amplification step of the library preparation reduces the bias [24], while others mention that, on the contrary, PCR amplification would favor AT-rich fragments [23]. It should also be mentioned that fragments with exceptionally high GC content tend to be under-represented among the sequenced reads, possibly due to single-stranded fragments adopting secondary structures that prevent them from being PCR amplified on the flow cell [22].

1.4.6 PCR Amplification

As part of the library preparation step, the DNA sample is usually subjected to several rounds of PCR amplification. It should be pointed out that there are two independent steps of PCR amplification in the protocol: one during library preparation and the other being the process of cluster formation on the flow cell, with the latter being outside of the control of experimentalist. The purpose of the former step is to increase the amount of starting material for cluster formation and therefore the number of sequenced reads. However, PCR amplification can bias the composition of the sample by preferentially amplifying certain fragments at the expense of others.

One way in PCR artifacts can manifest themselves in the data is as locations with multiple reads of same strandedness aligned to them, in effect producing stacks of reads [19, 16, 24]. In practice, such duplicate reads might also arise at the site of the true interaction due to the IP enrichment step and it might be hard to distinguish these from PCR artifacts. PCR amplification can also be bias composition of the sample in other ways that are harder to detect, for example by affecting the overall GC content of the fragment pool (see discussion in section 1.4.5). More subtle effects are also possible but have not been explored in ChIP-Seq data, to the best of our knowledge.

When the amount of starting material is too small, PCR amplification is required and the approaches mentioned in section 1.4.5 that eliminate it altogether are not aplicable. Therefore for libraries of low complexity the identification of binding sites may largely be driven by PCR artifacts [8].

There is a bias during the sequencing process itself, that favors shorter fragments for sequencing. This might be due to the process of PCR-based cluster building on the flow cell surface as longer fragments are less amenable to bridge formation.

1.4.7 Strand-Specific Sequencing

As mentioned above, for each aligned read the output of alignment software provides its location in the genome as well as its strand (whether it's a subsequence of reference strand or its reverse complement). In the case of point-like binding events, as is often the case with transcription factors, we can expect to see the pattern of positive-strand reads concentrated upstream of the event and negative-strand reads concentrated downstream of the event. This is often used by peak-finding tools to guide the identification of true binding sites. As a simple illustration of this phenomenon, consider the double-stranded DNA fragment 5'-GGATTAC-3' / 3'-CCTAATG-5', with GGATTAC being the subsequence of the reference strand. The single-stranded fragment 5'-GGATTAC-3' will give rise to a 3bp read GTA (reverse complement of TAC) during the sequencing process due to the nucleotides only being added at the 3' end of the growing sequence. Similarly, the other single stranded fragment 5'-GTAATCC-3' will give rise to a 3bp read GGA (reverse complement of TCC). While GGA is a subsequence of the reference strand and therefore is a positive-strand read, GTA is a subsequence of the reverse complement of the reference strand is a negative-strand read. Thus, the only reads that the double-stranded DNA fragment can give rise to are a positive-strand read at the 5' end of the fragment location on the reference strand and a negative-strand read at the 3' end of the fragment location on the reference strand. See Figure 1.1 for a schematic illustration and Figure 3.3 for some examples of read density at true signal sites in a real dataset.

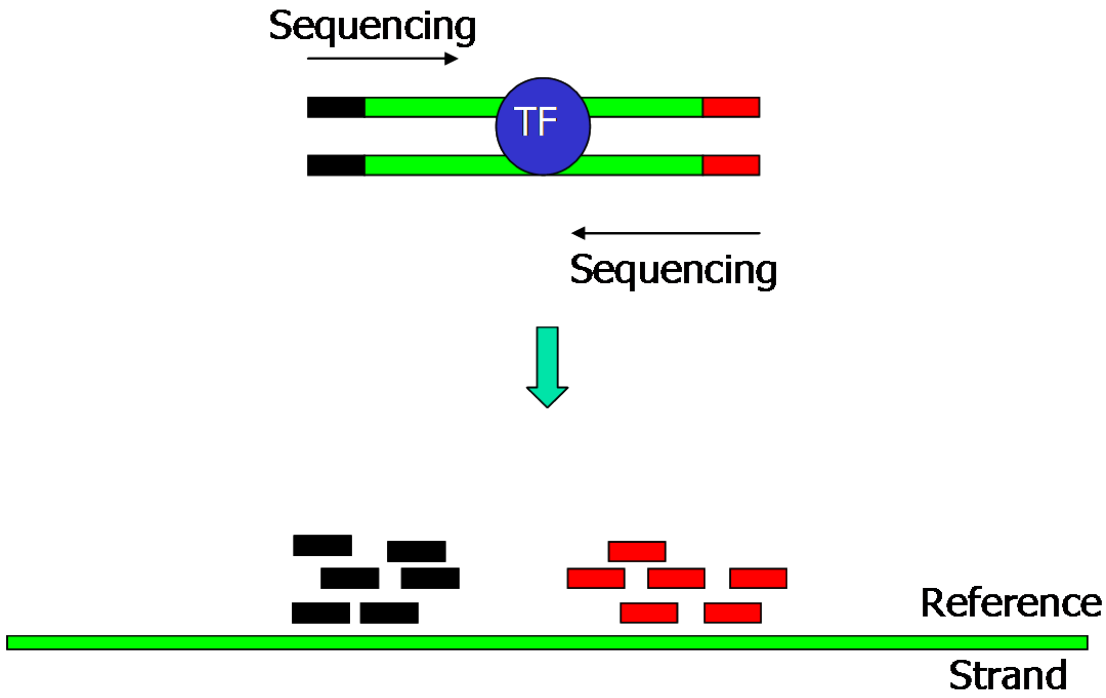

Figure 1.1: Illustration of strand-specific read pile-up in case of punctate interaction events, like transcription factor (TF) binding. Positive-strand reads are colored black, while negative-strand reads are colored red. For illustration purposes, both strands of the original fragment are colored to indicate the fragment end (but not the specific strand) that gives rise to a read.

1.4.8 Biological Interpretation

It is important to keep in mind that the observed results of ChIP-Seq experiments should be interpreted on the level of cell populations (the original sample) and not on the level of individual cells [3]. As a hypothetical example, observing comparable binding signals from transcription factors A and B at the genomic locus X may indicate that the two bind collaboratively OR competitively. The exact nature of their interaction cannot be discerned without further studies. Also, the detected binding events or modifications may or may not be functional and additional information (e.g., gene expression studies) is needed to establish this. Finally, one has to be careful when interpreting the strength of the detected signal as the strength or frequency of the underlying binding event, since the signal is subject to modulation by various biases described above.

Chapter 2

Overview of Current Approaches to Peak-Finding

2.1 Introduction

In this chapter we formalize the conceptual framework of peak-finding by introducing the notions of enrichment measures and enrichment statistics and categorize various peak-finders in terms of this framework. We provide a detailed overview of some common strategies, their relative advantages and disadvantages, and describe the statistical models used by some popular peak-finding tools. We discuss in some detail the different kinds of control samples used in ChIP-Seq experiments, and how they are incorporated into the peak-finding procedure. We also address the important issue of validation in the context of ChIP-Seq experiments and the shortcomings of the currently available validation approaches.

Peak-finding tools attempt to find the sites of protein-DNA interaction of interest based on genome-aligned reads. One has to be careful to distinguish between different types of interactions (e.g. transcription factor binding (TF) vs. histone modifications) as the appropriate methodologies can be quite different. Most currently available tools are designed for optimal performance with TF or other point-like event data, although some tools dedicated to longer events are also available. This work will focus on transcription factor binding applications.

The underlying strategy of all peak-finders is to locate genomic regions that are enriched with reads relative to some background, and to identify the true signal sites amongst those, while filtering out read-rich non-signal sites. The peak-finders look for candidate binding sites along the genome and calculate the enrichment statistic T_i for each candidate site i. Most tools provide a ranking measure for sorting the identified peaks, usually a measure of statistical significance (a p-value or False Discovery Rate (FDR) based on an assumed or estimated null distribution of T_i) or signal strength (the value of T_i itself, such as the number of reads or fold enrichment over a control sample).

2.1.1 One-Sample vs. Two-Sample Approaches

The main distinction between the peak-finding methods derives from their choice of background against which to calculate the enrichment. The so-called *one-sample methods* estimate the background from the ChIP-Seq sample itself, while *two-sample methods* use a second control sample as the background. The first ChIP-Seq publications [1, 2, 15] employed one-sample approaches, motivated by the assumption that the distribution of aligned reads along the genome is uniform, apart from the binding sites. The assumption of uniformity allows the underlying background rate of read occurrence to be estimated from the sample, and the significantly enriched sites can then be identified using appropriate distributional assumptions. However, it was quickly observed that this assumption of uniform background is not tenable [25, 4, 16, 6, 7]. The mappability and GC biases discussed in section 1.4 introduce local variations in the background read rate. Assembly collapse, PCR and chromatin structure effects can all result in non-specific enrichment far in excess of the estimated background rate. Additional non-specific enrichment can occur if the chosen antibody does not possess the desired specificity properties. To what extent these biases affect the final set of peaks is determined by the nature of the data and the choice of the peak-finder. If the binding profile under study consists of a large number of strong sites, then a two-sample approach does not provide much improvement over a one-sample approach, and the control sample may not be necessary [5]. On the other hand, the situations when the profile consists of a few weak binding sites or when the antibody has low specificity present a challenge for a one-sample approach, and a suitable control sample might be required to properly assess the background.

2.1.2 Count-Based vs. Overlap-Based Methods

Another distinction between different peak-finders arises from the choice of the primary enrichment measure, E_b, or enrichment at a single base-pair (bp) level. Commonly, $E_b(l)$ is defined to be either a) the number of reads aligned to location l in the genome (count-type enrichment), or b) the estimated number of aligned fragments overlapping l (overlap-type enrichment). In the former case, some peak-finders further subdivide E_b into $E_b^+(l)$ and $E_b^-(l)$, the numbers of reads aligned to l on the positive and negative strands, respectively. As a variation on the count-based approach to enrichment quantification, the read density can be smoothed using some kernel, producing a more continuous enrichment profile along the genome. The distinction between the two approaches to defining enrichment has to do with the fact that often the data consists of the single-end sequenced reads and the information about the other end of the fragment is missing, while in the case of the paired-ends sequenced reads the information about both fragment ends, and therefore the entire fragment after proper alignment, is present. PE-sequenced data thus provides more information about the binding event, since we expect most of the fragments at the signal site to cover the event, with the most covered base-pair usually used as a point estimate of the event location. Peak-

finders that employ overlap-type enrichment extend single-end reads to full fragments by using either a constant fragment length value (usually, a user-supplied or estimated average fragment length) or by extending reads probabilistically, and then calculate the enrichment.

Both types of primary enrichment measures extend naturally to genomic intervals (e.g., candidate binding sites), often with the count-type interval enrichment measure given by $E_i = \sum_{l \in i} E_b(l)$ (the number of reads in the candidate interval), and with the overlap-type interval enrichment measure given by $E_i = \max_{l \in i} E_b(l)$ (the maximum number of overlapping fragments). Other definitions of interval enrichment measures are also possible, e.g. the number of fragments overlapping the region. These interval enrichment measures or their functions usually serve as the enrichment statistics T_i. It should be pointed out that the count-type enrichment usually necessitates the binning of the genome into intervals of some pre-specified size, while overlap-type enrichment measures can be computed along the genome without preliminary binning, e.g. by looking for islands consisting of overlapping fragments. This latter form of enrichment also provides an intuitive choice of the point estimate of the protein binding site, namely the bp with the highest overlap enrichment.

2.1.3 Strand Information

A common strategy for the peak-finders is to utilize the strand-specific nature of sequencing, as discussed in section 1.4.7. Given that a point-like binding event will be represented in the data as a cluster of positive-strand reads of upstream of the binding site, followed by a cluster of negative-stran reads downstream of the biding site (where upstream and downstream are relative to the position of the event on the reference strand), peak-finding tools attempt to increase the resolution of their search by leveraging this information. This strategy is a required one for overlap-type enrichment methods working on single-end data, since they need to extend each read to a full-length fragment and to do so in a proper direction.

One standard approach to utilizing this information for peak-finders using the count-type enrichment is to shift the read positions towards the candidate binding site, moving positive-strand reads downstream and negative-strand reads upstream. This results in the concentration of signal around the point estimate of the binding event. Most often the shift is done globally by the same amount, although a local estimate of the required shift distance is also possible. In practice, the shift distance most often used is 1/2 of the average fragment length, with the latter quantity being supplied by the user or estimated from the data itself. The peak-finders that use the overlap-type enrichment measures extend each read by some amount (e.g. the average fragment length), transforming the reads into the estimates of the originating fragments. The schematic illustrations of the read-shifting and overlap approaches are shown in Figures 2.1 and 2.2, respectively.

There are several commonoly used strategies for estimating the average fragment length. One is to consider the positive and negative strand-specific enrichment profiles (e.g.,smoothed strand-specific read densities) for a subset of strong candidate regions, estimate the distance

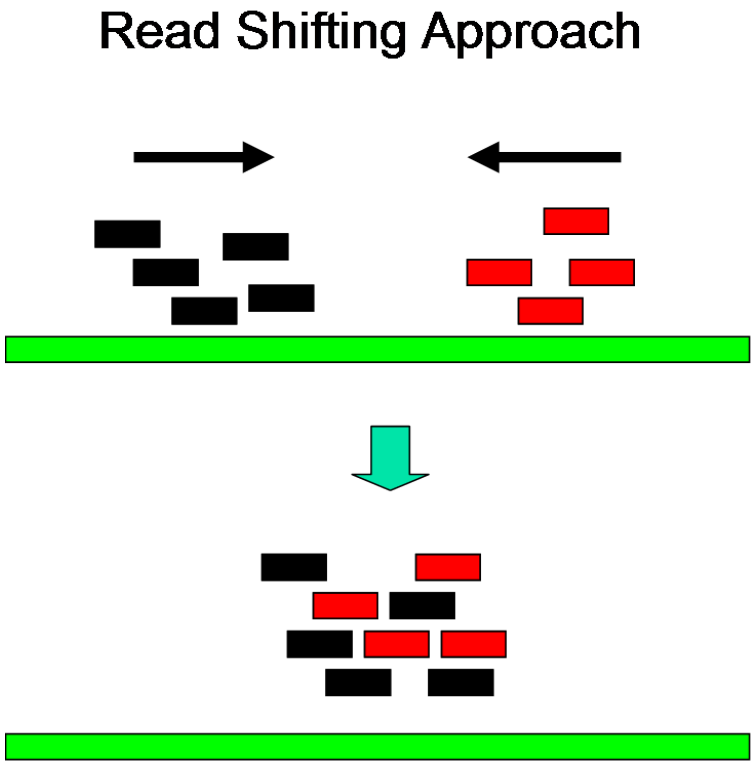

Figure 2.1: Schematic illustration of the read-shifting approach. Positive-strand reads are colored black, while negative-strand reads are colored red. The shift results in higher read density around the putative binding site (cf. Figure 1.1)

between the profiles for each region, and to let the mean or the median of these distances be the estimate of the average fragment length. Another strategy is to shift the strand-specific enrichment profiles relative to each other, until the maximum correlation between the profiles is obtained, with the corresponding shift being the estimate sought.

An important consideration to keep in mind with regard to the distance between the strand-specific profiles is that there are different conventions for reporting the position of the aligned read. In particular, one popular convention reports the 5' position of the match on the reference strand as the location of the alignment. Thus a 20bp read that spans bps 1001-1020 on chr1 will have 1001 reported as its location, regardless of alignment strand. In practice, this means that for the positive-strand reads the reported location is the corresponding fragment end, while for the negative-strand reads it is the (fragment end − read length + 1). A shift or another adjustment by the read length is then necessary if the information about the fragment length is to be inferred from the data or used to guide the peak-finding.

One should also be aware that there is a bias towards the sequencing of the shorter fragments on the flow cell and thus the value of the average fragment length based on the

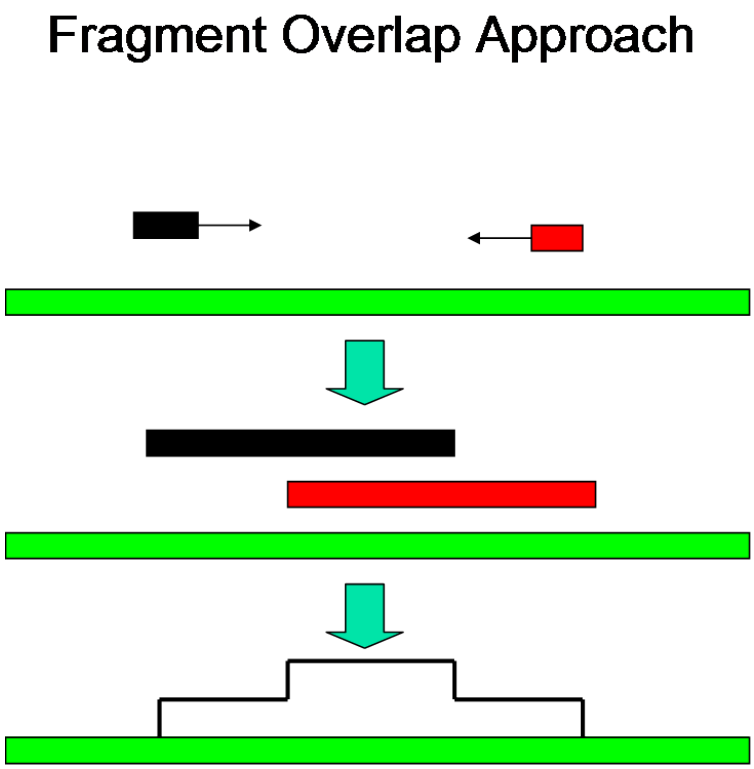

Figure 2.2: Schematic illustration of the overlap-type approach. A positive-strand read is colored black, while a negative-strand reads is colored red. Each read is extended to a full fragment in the 3' direction and the coverage by the resulting fragments is computed along the genome.

size selection step overestimates the average fragment length of the sequenced fragments [4, 20], and should perhaps be lowered when supplied to a peak-finder as an argument. The actual effect of using the overestimate of the average fragment length on the number of called peaks is likely to depend on the extent of the bias in the data set, among other things.

2.1.4 Point Event Identification

For ChIP-Seq applications to point-like events (e.g., TF binding), a bp-level estimate of the event location is often desired. Most peak-finders provide this estimate of the peak 'center' or 'summit', and the rest try to return peaks of narrow width. This center position is usually estimated as a local maximum of the read density (after read shifting) or overlap enrichment profile. Sometimes, this estimate is based on the maximum value of $T_{i,k}$ if $k > 1$ values of the enrichment statistic are calculated for each enriched region. Alternatively, a peak-finder might look for locations along the genome where a predominantly positive-strand read

region transitions into a predominantly negative-strand read region and identify the point of transition as the site of the binding event (cf. Figure 1.1). Restricting the reported size of the binding event region helps the investigator to identify the potential affected genomic targets (e.g., genes) and to conduct motif-searching.

2.1.5 Paired-Ends Data

Another issue the peak-finders face is how to deal with PE data, since most of them have been designed to handle single-end (SE) data sets. For the peak-finders employing the overlap-type enrichment measures, the extension to PE setting is straightforward: the reads are extended to fragments based on the alignment of the two sequenced fragment ends. For the peak-finders using the count-type enrichment measures, the usual strategy is to treat the 2 sequenced ends of the fragment as independent reads.

2.1.6 Duplicate Alignments

The peak-finders also usually try to address the issue of duplicate alignments: the strand-location combinations that occur more than once. This read-stacking could be indicative of the strong enrichment at a true signal site or could be due to various PCR biases. Many peak-finders try to cap the maximum number of alignment copies kept in the data, by either allowing only 1 copy or using a model-based cutoff. The actual extent to which PCR biases affect the number of duplicate alignments cannot be determined directly and could in fact be quite minor. Therefore, any attempt to reduce the number of allowed alignment copies risks diluting the signal in the data and making it harder to identify the true binding sites. This problem is particularly important in the experiments with large numbers of sequenced reads and small genome sizes, where retaining only 1 copy of each alighment might lead to not being able to distingiuish between the signal and the background at all.

2.1.7 Notation

I will employ the following notation in what follows:

- $E_{i,t}$ and $E_{i,c}$ are the interval enrichment measures for candidate binding event i in the treatment and control samples, respectively.
- N_t and N_c are the total numbers of reads in the treatment and control samples, respectively.
- $n_{i,t}$ and $n_{i,c}$ are the numbers of the treatment and control reads, respectively, in a region corresponding to the candidate binding event i.
- $T_{i,t}$ and $T_{i,c}$ are the enrichment statistics for candidate binding event i calculated in the treatment and control samples, respectively.

- G and G_{map} are the size of the entire genome and its mappable portion (sometimes called the effective genome size), respectively.
- NB and Bin are the shorthands for negative binomial and binomial distributions, respectively.

2.2 One-Sample Methods

Some of the published peak-finders that allow for data sets without control samples include ERANGE (based on [1]), FindPeaks [26], MACS [4], USeq [16], SISSRS[27], CisGenome [5], BayesPeak [28], and MOSAiCS [22]. To guide their peak-calling these peak-finders directly model the background read distribution and then assign statistical significance to their candidate peaks using either analytical or simulation approaches.

One popular background model assumes that the read (or fragment) start site process along the genome is Poisson with constant rate $\lambda = N_t/G_{map}$. This model will be referred to as global Poisson or random uniform model, since, under the model, the distribution of read/fragment start sites across mappable bases is uniform, conditioned on the total number of reads/fragments N_t. Under the random uniform model, the distributional results can easily be obtained for both count-based and overlap-based enrichment measures, and their derived enrichment statistics (e.g., the number of reads in a 50bp window or the number of fragments overlapping a given location), and measures of statistical significance (e.g., p-values) can be assigned to the observed enrichments. Alternatively, one can simulate the null distribution of the enrichment statistics by picking N_t positions from G_{map} bases (usually concatenated into a single chromosome pseudo-genome) at random with replacement (without replacement if duplicate alignments are reduced to a single copy). The advantage of the simulation over the analytical approach is that it is more amenable to modifications. For example, it is unrealistic to treat the mappable portion of the genome as contiguous, and if mappability information is available, it can be incorporated into the simulation. Another proposed modification is to sample clusters of reads instead of the individual reads [6]. A cluster might be defined to consist of all reads mapping to the same bp or of all reads in non-overlapping 5bp windows. This sampling strategy allows for a better modeling of non-homogeneity of the background.

Since the background is known to be non-uniform due to the biases described in section 1.4, some authors abandon the global Poisson model in favor of more sophisticated approaches. One popular strategy is to model the background read counts in windows of certain size as following a negative binomial (overdispersed Poisson) distribution. This approach allows a peak-finder to reject more regions at the same p-value than under the global Poisson model, and is in essence just a way to raise the global peak-calling enrichment statistic cutoff. It should be noted that while the NB model may account for more low-level non-specific enrichment than Poisson model, it is still powerless to discard any read-rich artifacts that have enrichment on par with the true signal.

Another approach is to move away from having a single global cutoff value for enrichment statistics and to determine the statistical significance locally instead. For example, the local constant Poisson rate can be estimated from the region of size L that contains the candidate binding site of size l with $l < L < G$. This local rate varies along the genome and a candidate region in a locally read-rich segment will require more reads to be declared a true binding event than a region in a locally read-poor segment of the genome at the same p-value cutoff.

A twist on the negative binomial approach is implemented in BayesPeak, a peak-finder that uses an HMM framework to detect binding sites. BayesPeak models read counts in windows along the genome as coming from negative binomial distribution if the window corresponds to the background and as coming from a mixture of two negative binomial distributions (the signal and noise components) if the window corresponds to a binding event. The read counts in consecutive windows are dependent due to the fragments that span both windows and the actual enrichment status is unobserved. After the model is fit, the windows with posterior probability of enrichment > 0.5 are declared to be the binding events.

A similar twist (but without the HMM framework) is adopted by the peak-finder MOSAiCS, which also models the read counts in background windows as having NB distribution while the read counts in the signal windows follow a distribution that is a mixture of two components: background (NB) and signal (either NB or a mixture of 2 NBs). MOSAiCS adopts a local approach to calculating the significance of enrichment by modeling the means of the underlying negative binomial distributions (both background and signal) as functions of the GC and the mappability content of the window in question. The peaks are called based on the posterior probability of belonging to the class of enriched or background windows. Interestingly, the authors of MOSAiCS note in the manuscript that a 2-NB mixture model for the signal fits best in the case of the treatment ChIP samples, but that a single NB model is sufficient for the 'signal' in the input DNA samples (a type of a negative control sample), indicating that the high-level enrichment in the ChIP-Seq treatment samples consists of the true signal and some read-rich artifacts that are also present in control samples.

In summary, while ERANGE, FindPeaks, USeq and SISSRS use global Poisson model to guide their peak-finding, other methods attempt to deal with the non-uniformity of the background by employing a more sophisticated global model (NB, used by CisGenome, BayesPeak) or by switching to a local model (Poisson for MACS, NB for MOSAiCS). It should be noted that BayesPeak does capture the local effects to some extent, due to the underlying HMM machinery.

The enrichment statistics T_i used by one-sample methods are usually just the interval enrichment measures E_i for the candidate regions (approach taken by ERANGE, FindPeaks, USeq, SISSRS, and MACS) and their null distributions follow from the underlying global (or local) Poisson or negative binomial models. Analytical p-values (possibly adjusted for multiple hypothesis testing) or corresponding FDRs are easily obtained, with analytical FDRs estimated as E_t/O_t where E_t is the expected number of regions exceeding enrichment statistic cutoff t under the null distribution and O_t is the corresponding observed quantity.

Simulation approaches can also be used to obtain p-values and FDRs. Mixture-model approaches (BayesPeak and MOSAiCS) use the posterior probability of a binding event as the significance measure.

2.3 Two-Sample Methods

There are many more published two-sample methods than one-sample methods and all the one-sample methods reviewed above also have the two-sample versions. Some of the published two-sample methods designed for the transcription factor binding site identification are ERANGE, FindPeaks, MACS, USeq, QuEST [29], CisGenome, SISSRS, PeakSeq [7], GLITR[21], SPP [6], T-PIC [30], BayesPeak, MOSAiCS, CCAT [31], PICS [32], Sole-Search [10], and CSAR [33]. Some of these can also be applied to the problem of identification of non-punctate events, and a few methods have been developed that specifically aim to identify the locations of histone modifications (ChIPDiff [34], SICER [9]) or the sites of differential Pol II activity (Poisson Mixture Model [35]). As a side note, we mention that SICER also has a one-sample module that identifies the significantly enriched regions based on the global Poisson background model. These peak-finders targeting non-TF events will not be discussed further, but their basic approaches are similar to some of those presented below.

There are two basic strategies for introducing the information from properly normalized control samples into peak-finding, and often both are employed simultaneously. One strategy is to compare the enrichment at the candidate site directly to the enrichment of the same site in the control sample, incorportating the control information into the enrichment statistic T_i. The significantly larger enrichment in treatment relative to control indicates the sites of true binding events or artifacts that are not picked up by that particular control sample. Another strategy is to use the control sample to deduce the null distribution of T_i and to attach a measure of statistical significance to the called peaks. The basic assumption underlying both of these strategies is that after suitable normalization, the background enrichment profile in the treatment sample is the same as the overall enrichment profile in the control, up to some random (and hopefully negligible) variation.

2.3.1 Types of Controls and their Properties

There are several different types of negative control samples commonly employed in ChIP-Seq experiments:

- Input DNA control. This control sample consists of the native sheared chromatin that does not undergo the immunoprecipitation process but is otherwise subjected to the same library preparation protocol as the treatment sample. This is by far the most widely used type of control. Most often the sample of interest is split, with one

portion providing the input DNA control and the other subjected to the full IP process, although suggestions to re-use input DNA samples have been made (see discussion later in this section).

- Mock-IP control. This control sample consists of chromatin that undergoes the same library preparation as the treatment sample but there is no antibody used during the IP step. The control samples of this type are obtained by splitting the original sample of interest.

- Non-specific antibody control. This control consists of chromatin that undergoes the same library preparation as the treatment sample but the IP is done with an antibody known for binding proteins in a non-specific manner, usually Immunoglobulin G (IgG). The source of the starting material for this control sample is the sample under study, just like for the previous two control types.

- Different tissue type or condition. This control consists of chromatin obtained from a different source (tissue type or condition) that undergoes exactly the same library preparation steps as the treatment sample, including using the exact same antibody.

Before discussing the properties of various types of controls, it should be mentioned that there are different reasons for using controls in the first place. The control sample might represent a biological situation of interest, e.g. if the goal of the experiment is to study differential binding in two tissue types or the same tissue type under different set of conditions. More often the controls are used to capture the sample-specific background, to help in filtering out any non-specifically enriched regions and to boost the significance of the regions with weak signal.

If the background assessment is the primary purpose of the control, it is extremely important that the backgrounds of treatment and control samples be very similar (after suitable normalization, as described in the next section). In particular, input DNA, mock-IP and non-specific Ab controls are usually obtained from the same sheared chromatin as the treatment sample, and therefore should capture any biases or artifacts introduced by the shearing process or inherent in the sample. By definition, a different tissue/condition control sample might not capture this background, and it is sometimes used as a different-treatment sample, with separate controls (e.g. input DNA) supplied for both treatment and different-treatment samples, with the final comparison being done on the sets of detected binding events in the two treatment samples.

Input DNA controls will not capture any biases or artifacts introduced during the IP step, while mock-IP and non-specific Ab controls aim to do just that. Mock-IP controls aim to capture the effects of different factors of the IP protocol, apart from the antibody itself, on the background enrichment and thus would not be able to guard against non-specificity of the antibody. Additionally, there are known problems with mock-IP control samples that result in a very small amount of post-IP starting material [5]. IgG-type controls aim to

additionally capture the effects of non-specific antibody-protein interactions, but it is not clear why these should be the same for immunoglobulin G and the antibody for the protein under study, since the latter is targeting a specific epitope or a set of epitopes. Another approach is to use a mutant tissue as the control, with the protein of interest knocked out, but this might not be feasible or meaningful due to the effects of the knockout mutation on the underlying biology and possibly on the chromatin state. Yet another approach is to use as the treatment sample a mutant tissue where the protein of interest has been fused with a foreign epitope (surface amino acid chain), absent in the wild type. The antibody targeting the foreign epitope can then be used and the wild type tissue employed as the control. Since the wild type lacks the epitope targeted by the antibody, this control sample should be able to account for any antibody non-specificity. In this case the treatment and control samples come from different sources of chromatin (different tissues), which might affect the background.

Does the choice of the control make a difference? As mentioned earlier, if the distributions of enrichment measures for signal and background regions are well-separated, then the control might not even be necessary. Whether any particular choice of control captures the desired background is probably data set-specific, and depends on the nature of the binding event, the quality of the antibody, the consistency of the library preparation process carried out by the technician and many other factors. As a (probably extreme) example, we have come across a data set that had both input DNA and wild type IP controls (the treatment had a foreign epitope fused to the protein of interest). We discovered that running popular peak-finders on this data set results in about 80% reduction of called peaks when switching from input DNA to the wild type control, due to some unforeseen antibody properties. As far as we know, no definitive study has been done to assess the appropriateness of different control types and input DNA controls continue to be widely used.

An interesting point raised in the literature is whether input DNA controls can be re-used. Some authors claim that separate input DNA controls should be used for different tissues or for the same tissue under different cellular conditions [7, 17] while others disagree [21]. The authors of the web-based peak-finder Sole-Search stress that input DNA libraries should be cell type-specific and make several input samples available online for their users, perhaps implying that they do not think that different cellular conditions or shearing protocols will affect the background as mesaured by input DNA controls. As far as we know there have been no claims of notable differences between input DNA samples produced from the same tissue cells under the same conditions and with the same protocols being followed, and no comparisons have been made of the effects of different chromatin shearing techniques on the resulting background enrichment profile and the identified binding sites. It should be noted that any observations made with regards to different input control samples are conditional on the protocols being followed and the biological phenomenon under study. If the treatment sample shows a very strong signal profile, then even significant differences in input DNA samples used as controls might not affect the final set of identified binding regions.

2.3.2 Normalization

Most peak-finders compare the candidate region enrichments in treatment and control samples to assess whether the enrichment is due to the background artifacts or the true signal. Most also utilize the control sample to provide the null distribution of enrichment statistics T_i for the purposes of assigning statistical significance to the called peaks. In both situations, proper normalization is essential. The basic argument behind the normalization approaches is that for the background windows i, $E_{i,c} = \alpha E_{i,t}$, where α depends on sample sizes N_t, N_c and the fraction of the background reads in the treatment sample. Thus α needs to be taken into account if window counts $n_{i,c}$ and $n_{i,t}$ (or enrichment statistics T_i based on them) are to be compared.

A very popular normalization approach is to assume that normalization factor $\alpha = N_c/N_t$, known as the total sample size normalization. In the situations, where each sample consists of reads from a single lane, the term 'lane total normalization' is also used. One way to normalize the treatment and control under this assumption is to subsample a larger of the two samples to ensure that after the subsampling $N_t = N_c$. This approach is taken by QuEST, USeq, GLITR, and Sole-Search and it results in some data loss as reads are discarded. Another approach is to record the total sample sizes and to use them to adjust the count- or overlap-based enrichment measures, for example by comparing $n_{i,t}/N_t$ to $n_{i,c}/N_c$ in a candidate region i. This approach is taken by ERANGE, MACS, and PICS.

Since some fraction of the reads in the treatment sample does not come from the background that the control sample aims to capture, some peak-finders try to base normalization only on the background portion of the treatment sample, instead of the entire sample. They argue that if this is not done, some weak signal might be missed. One way to implement this normalization approach is to assume that for some small k all genomic windows with $n_{i,t} < k$ contain background reads only and then use the ratio $\left(\sum_{i:n_{i,t}<k} n_{i,t}\right) / \left(\sum_{i:n_{i,t}<k} n_{i,c}\right)$ as the normalization factor (CisGenome, SPP). Another approach is to estimate the fraction $0 < \alpha_{BG} < 1$ of background reads in the treatment sample and then to subsample reads to ensure that after subsampling $N_c = \alpha_{BG} N_t$ (CSAR). Other approaches involve fitting a line to points $(n_{i,t}, n_{i,c})$ for windows i that are assumed to represent background (FindPeaks, PeakSeq).

With normalization approaches based on the subsampling of reads from one or both samples, the subsampling can be repeated several times and only the peaks identified in all samplings can be retained. Multiple repeated subsamplings can also be used to obtain a better estimate of the null distribution of the enrichment statistics. Variations on this idea are used by CCAT and CSAR.

Mixture model-based approaches (BayesPeak, MOSAiCS) incorporate control read counts as covariates into the underlying model without normalizing (presumably, the normalization factor will contribute to the intercept term). The peak-finder CSAR normalizes the distribution of enrichment measures in the treatment sample with respect to the distribution of

enrichment measures in the control sample. Finally, some methods do not state explicitly what type of normalization (if any) they adopt (SISSRS, T-PIC).

In practice, for samples with small fractions of true signal reads using total sample size normalization should work as well as normalizing the background portion only. This is usually the case with TF ChIP-Seq data sets. However, large fraction of true signal reads has been reported in Pol II ChIP-Seq study and total sample size normalization might not be acceptable in such settings.

2.3.3 Candidate Region Identification

Some peak-finders produce an initial set of candidate regions that are further processed to determine the final set of identified binding sites. This first filtering step is usually done by employing some pre-specified (ERANGE, FindPeaks, QuEST, SISSRS, GLITR, PICS) or model-based (FindPeaks, MACS, PeakSeq, Sole-Search) cutoffs on the minimum number of reads in a window or the minimum fold enrichment over (normalized) control. Model-based cutoffs are usually derived analytically or obtained in simulations under the global (MACS, FindPeaks) or segment-specific (PeakSeq,Sole-Search) Poisson model or under the global NB model (CisGenome). The segment-specific approach of PeakSeq and Sole-Search tries to take mappability effects into account. Other peak-finders might also have internal first pass filters to reduce the amount of computation required for assessing the significance of candidate binding regions. In addition, some peak-finders (CSAR, PICS) also use pre-specified cutoffs on enrichment statistics or other metrics that are used to filter out candidate regions prior to assignment of statistical significance to the candidates.

2.3.4 Enrichment Statistics and Control

A lot of peak-finders incorporate the (possibly normalized) control into the enrichment statistics T_i in some way. In contrast to one-sample methods, where the usual approach is to have $T_i = E_{i,t}$, i.e. to use the enrichment measures (number of reads in a window or a peak height) as enrichment statistics, most two-sample peak-finders let $T_i = f(E_{i,t}, E_{i,c})$ but use different functions f. Some possible choices of f are listed below.

- $f(E_{i,t}, E_{i,c}) = E_{i,t}/E_{i,c}$, i.e. the fold enrichment approach (SISSRS, CCAT, PICS).
- $f(E_{i,t}, E_{i,c}) = g(E_{i,t} - E_{i,c})$, i.e. the background subtraction approach, followed by a score transformation (SPP).
- $f(E_{i,t}, E_{i,c}) = d((E_{i,t}, E_{i,c}), l)$, where $E_{i,t} = n_{i,t}, E_{i,c} = n_{i,c}$, l is the normalization line fitted to window read counts (as described above) and d is some measure of distance from a point to a line (FindPeaks).
- $f(E_{i,t}, E_{i,c}) = P(X > E_{i,t}|E_{i,t}, E_{i,c})$, where $X|E_{i,t}, E_{i,c} \sim F(E_{i,c}, E_{i,t})$ for some specified choise of distribution F. In this approach, $E_{i,t}$ is modeled as having a distribution

with parameters that depend on the observed value of $E_{i,c}$ (and possibly the observed value of $E_{i,t}$), and the resulting enrichment statistic T_i is a p-value. For example, one can model $n_{i,t}$ as Poisson($n_{i,c}$) random variable, where $n_{i,t}$, $n_{i,c}$ have been suitably normalized (ERANGE, MACS, QuEST, CSAR). Alternatively, one can model $n_{i,t}$ as $\mathrm{Bin}(n_{i,c} + n_{i,t}, 0.5)$ random variable (again, after proper normalization) (USeq, PeakSeq) or $\mathrm{Bin}(n_{i,c} + n_{i,t}, p)$, where the $n_{i,t}, n_{i,c}$ are non-normalized read counts and the normalization ratio is incorporated into p (CisGenome). More sophisticated distributions are also explored (T-PIC).

Yet another approach (GLITR) is to use 3 control samples, each of the same size as treatment (an example of total sample size normalization) and to let one of them be a pseudo-treatment sample and the other two be the control samples for the two treatment samples (one real, one pseudo). Both treatment and pseudo-treatment samples are scanned to produce candidate regions and for each region the vector of two scores (peak height, fold enrichment over control) are obtained. The scores are normalized and the enrichment statistics T_i are calculated as fraction k/n of the closest n neighbors of the candidate binding event in the 2-D score space that come from the treatment sample.

In summary, the information from control sample is used to transform enrichment statistics T_i from simple measures of absolute enrichment $E_{i,t}$ into measures of relative enrichment, either a fold enrichment, difference enrichment, a p-value, or a measure of similarity to other enriched treatment regions.

2.3.5 Statistical Significance

In order to assign statistical significance to enrichment statistics T_i, their null distribution must be known or estimated. There are several ways to address this issue:

- When T_i itself is a p-value, it can be used as a measure of significance directly (possibly after adjusting for multiple hypothesis testing). This is the approach taken by ERANGE, CisGenome, QuEST, PeakSeq, T-PIC and Sole-Search.

- A second control sample (of the same size as the treatment sample) can be run against the original control, and the resulting values of T_i' provide the null distribution. This is the approach that can be taken by QuEST if a second control sample is available (in practice, if the original control sample is large enough to be split). This is also the approach taken by USeq to obtain the empirical FDR value for each p-value cutoff (for this peak-finder T_i's are the p-values). A variation on this general scheme is employed by GLITR which produces points in its 2-D score space (see above) for both treatment and pseudo-treatment samples and calculates T_i for each point as the fraction of treatment points within certain neighborhood of the point. An empirical FDR is obtained for each cutoff value t of T_i as P_b/T_b, where P_b, T_b are the numbers of pseudo-treatment and treatment points, respectively, that have $T_i > t$. Finally, a

similar approach is used by CSAR which merges reads from the treatment and the control and repeatedly samples a pseudo-treatment sample from this joint set of reads, and runs it against the rest of the reads to obtain the null distribution of T_i.

- SISSRS randomly samples genomic regions to obtain the null distribution of T_i, which for this peak-finder is the fold enrichment over control. The authors reason that vast majority of the genome lies in the background regions and will provide a good estimate of the null distribution of fold enrichments.

- Mixture model-based methods (BayesPeak and MOSAiCS) use posterior probability of binding as their significance measure.

- Finally, a very popular technique for obtaining emprical FDR values for various T_i cutoffs is the so-called 'library swap' approach, used by MACS, SPP, CCAT and PICS. ERANGE also uses this technique to provide a single FDR estimate for its final set of called peaks, but not to set cutoffs for T_i. In this approach, the roles of treatment and control samples are exchanged and the resulting distribution of enrichment statistics T_i' is used to obtain an empirical FDR for the enrichment statistic cutoff t as follows: eFDR=$\#(T_i' > t)/\#(T_i > t)$. The intuition behind using this approach is that the distribution of enrichment statistics for background regions should be the same for both samples (after samples have been properly normalized). The authors of the CCAT package prove that under some assumptions, P(region i is not signal $|f(n_{i,c}, n_{i,t}) > t)$ can be estimated as $\#(f(n_{j,t}, n_{j,c}) > t)/\#(f(n_{j,c}, n_{j,t}) > t)$, where f is the significance function such that low values of $f(n_{i,t}, n_{i,c})$ correspond to true signal regions (e.g. $f(x, y) = y/x$, inverse fold enrichment approach). Among these assumptions are a) $N_c = N_{t,BG}$, where $N_{t,BG}$ is the number of reads in the treatment sample that come from the background, and b) P($f(n_{i,t}, n_{i,c} > t|$ region i is signal) << P($f(n_{i,c}, n_{i,t}) > t|$ region i is signal). The interpretation of CCAT's derivation is straightforward in the case of fold enrichment f (P(region i is not signal $|n_{i,t}/n_{i,c} > t) = \#(n_{j,c}/n_{j,t} > t)/\#(n_{j,t}/n_{j,c} > t)$), but is not so clear when dealing with some other enrichment statistics, such as p-values based on distributions defined in terms of control samples.

One problem with the normalization methods that do not try to balance the sample sizes by subsampling is that due to asymmetry of the normalization procedure, the estimated significance can be arbitrarily affected by scaling the sample sizes, depending on the choice of the enrichment statistic. For an illustration of this point, consider a treatment and a control samples with 20M and 10M reads, respectively. Suppose that there are 40 treatment reads and 10 control reads in a candidate region i and that the Poisson model for region counts is assumed, with the rate parameter to be estimated from the control sample. One way to estimate the significance is to adjust the control rate for the ratio of the total sample sizes. In this case, the model states that the observed 40 reads come from Poisson(10/(10/20)=20) distribution. Another approach is to adjust the observed treatment read counts instead of the

control read counts, in which case we assume that the observed adjusted (40/(20/10))=20 counts come from Poisson(10) distribution. The fold enrichment of observed counts over underlying rate is the same, but the statistical significance ascribed will differ. Examples like this can be produced for other p-value enrichment statistics as well that are based on enrichment adjustments. This will pose a problem when using a control sample to estimate the null distribution of enrichment statistics. With the usual approach of leaving treatment read counts alone and normalizing the control read counts, the sample with more reads will tend to produce more called regions at the same p-value enrichment statistic cutoff, sometimes resulting in empirical FDR of 100% for all called peaks (when control sample is much larger than treatment). Similar problems can arise with other types of enrichment statistics as well; for example, the sample with more reads might have a lot of regions with high fold enrichment (but low number of reads) due to very sparse coverage in the other sample. This leads some authors who do not employ subsampling approaches to recommend to the users that they do not use unbalanced samples (MACS).

In general, the authors of peak-finding tools do not attempt to justify rigorously the assumptions underlying their derivations of null distributions of enrichment statistics (with CCAT being the one exception). As a result, any measures of statistical significance reported by these tools need to be taken with a grain of salt and treated more as a ranking device rather than taken at the face value. This pragmatic attitude is often adopted by the tools' authors themselves, when they treat the computed enrichment p-values as enrichment statistics and estimate their null distribution based on, e.g., library-swap approach, indicating that they are not willing to trust the models that produced these p-values in the first place (otherwise uniform distribution for p-values would be assumed). Some tools report several different p-values based on different models, and these p-values are often not very similar to each other.

2.4 Validation and Comparison

There are several ways to assess the performance of the peak-finders on a particular dataset:

1. If the transcription factor under study has a strong canonical motif, one can look for motif occurrence rate in the identified signal regions (or within some small distance of the called peak summits for higher resolution). However, the existence of non-canonical motifs make equating true binding event with motif occurrence impossible. It is also true that not all canonical motifs will show signs of binding due to various other factors affecting the binding process, e.g. the chromatin state.

2. Often sets of qPCR positive and negative regions are available, sometimes from earlier studies. The qPCR assay targets specific genomic regions and compares the PCR-based measures of abundance of the region-containing fragments in two DNA samples (the treatment and control samples in the case of ChIP-Seq). If the region is found to be

significantly enriched in the treatment relative to the control based on the qPCR assay, it is said to be *qPCR positve*, and it is said to be *qPCR negative* otherwise. One has to be careful when dealing with ChIP-qPCR positives, since some regions are selected preferentially during the ChIP process and might give rise to some false positives. In general, the regions are chosen based on some previous knowledge of transcription factor targets, and thus the concern about false positives is not applicable. These qPCR positive regions can be used to assess sensitivity of a peak-finder.

In conjunction with qPCR positives, often a set of qPCR negative regions is used to assess a peak-finder's specificity. However, the negative regions chosen for qPCR validation are often extremely read-poor, thus making any assessment of specificity unreliable. For example, the authors of an early ChIP-Seq study of NRSF binding [1] build an ROC curve for their peak-finding approach using a set of 83 qPCR positive and 30 qPCR negative regions, and this set is used by several other peak-finding publications for the purposes of sensitivity/specificity analysis. However, when we look at how many of the 1kb genomic windows containing these regions have the number of reads that exceeds 10^{-5} p-value cutoff based on the global Poisson model (a modest cutoff, given the multiple hypothesis testing consequences of looking at about 2.8 million such windows), we find that fully 75/83 qPCR postives but only 3/30 qPCR negatives pass this cutoff. In other words, the vast majority of qPCR negatives do not show enough enrichment to even be considered as candidate regions by a one-sample approach and should not be used as a validation set for the two-sample approaches. The numbers are almost identical for some other published NRSF data sets [29] and this set of qPCR verified regions. Similarly, for the STAT1 binding data set from [2], 15/17 qPCR positives and 2/41 qPCR negatives pass the cutoff based on global Poisson p-value of 10^{-5}. Another data set commonly used in peak-finding publications for validation purposes is on the binding of transcription factor CTCF [36] and a set of 82 qPCR positives and 17 qPCR negatives from a different publication has been used to assess the peak-finders' performance on this data set. Under the global Poisson model, 58/82 qPCR positives pass the p-value cutoff of 10^{-5}, compared to 3/17 qPCR negatives, again rendering any specificity analysis based on this set of regions uninformative. The surprisingly large number of non-enriched qPCR positives in this data set is probably due to differences in library preparation, illustrating the danger of validation using data obtained under a different set of conditions (another explanation is that errors have occurred due to coordinate changes between the different versions of the genome). Since most often the primary purpose of using control samples is to filter out the non-specific enrichment artifacts, the ideal set of regions for specificity assessment should consist of such artifacts. Unfortunately, no such data set exists.

3. Gene expression studies are sometimes used in conjunction with ChIP-Seq experiments to identify the transcriptional targets of the protein binding. The utility of using such studies as a validation tool is doubtful, since a true binding event might not be

functional, and the task of assigning a transcriptional target to a binding site is rarely trivial. Gene expression studies, however, are invaluable in narrowing down the list of called peaks to those with an identified effect for further study.

Overall, there is no gold standard data set to validate the performance of peak-finding tools. Computational spike-in experiments have been proposed and carried out based on introducing simulated signal into the real data [16, 5, 31], but the biological spike-ins are sorely needed, that would rely on the true biological signal of known strength.

Two recent studies [37, 38] have compared the performance of different peak-finding tools on some data sets. There are some easy-to-spot qualitative differences between the sets of peaks produced by different tools, the most obvious being the differences in overall numbers and the widths of the reported peaks. The differences in peak width might in principle be an important consideration for the users, since narrower peaks allow for a more precise motif search and identification of potential binding targets. However, since almost all peak-finders produce a point estimate of the binding event (so-called peak summit), one can simply look at the windows of pre-specified length centered on the peak summits for motif-finding, and also to use summits themselves when looking for the nearest potential transcription targets.

The differences in the numbers of reported peaks are more important. The reviews cited above conclude that these differences reflect the peak-finders' internal stringency criteria. In other words, if two peak-finders identify X and Y peaks, respectively, with $X << Y$, then it is usually the case that the X peaks from the first peak-finder are among the Y peaks from the second peak-finder. Users usually can set stringency criteria for peak-finders to be different from the default ones, and thus adjust the number of the reported peaks. How to properly set this criteria in the light of the dubious meaningfulness of the reported statistical significance measures (see section 2.3.5), and whether the additional peaks represent mostly true signal or non-specific enrichment, are open questions to which no simple answers exist. An interesting observation made by the authors of the two reviews is that the relative numbers of peaks produced by different peak-finders are to some extent data set-dependent: a peak-finder X may produce many more peaks than a peak-finder Y for one data set but many fewer for another data set. A very unsettling observation is made by [37] who note that when they compare the number of binding events at two points in their time course data, the ratios of the numbers of identified events between the two points range from 8-fold decrease to 2-fold increase, depending on the choice of the peak-finder.

Overall, choosing the 'best' peak-finder seems to be a data set-specific task that cannot be adequately resolved without some pre-defined validation criteria.

2.5 Summary

In this chapter we have discussed the various approaches to peak-finding. We have identified the modules of the peak-finding procedure (e.g. normalization or candidate region identification) and have compared the published methods based on their approaches to implementing

these various tasks. Of particular interest to us were the statistical underpinnings of the signal site identification, based on method-specific enrichment statistics and their assumed or estimated null distributions.

As detailed above, all peak-finding approahces fall into two basic categories: those that utilize information from a control sample and those that do not. All one-sample methods use global enrichment cutoffs to find candidate regions, with exception of MACS and MOSAiCS. Whether the global cutoffs are based on Poisson, negative binomial or some simulated randomization model, the peak-finders utilizing them fail to account for systematic local read density fluctuations. These fluctuations arise due to the mappability effects, the technological GC bias, the enrichment in open chromatin regions and at assembly collapse sites, the natural copy number variation, and other sources. All that the more sophisticated global models such as negative binomial or simulation-based approaches accomplish is to increase the 'p-value' for the same level of enrichment compared to the global Poisson (uniform) model. These global approaches are powerless to filter out strong non-specifically enriched artifacts and raising the enrichment significance cutoff too high may result in discarding some true but weak signal. MACS is the only one-sample peak-finder to use local cutoffs based on the enrichment level in the encompassing region. However, based on the size of the region chosen to estimate local background enrichment, MACS might be powerless to discard highly localized artifacts or to pick up true signal sites in close proximity to each other or to some locus of non-specific enrichment. MOSAiCS attempts to model background enrichment as a function of local GC and mappability states, features known to affect read density. We propose our own one-sample approach in the next chapter that takes the shape of the read density profile into account and is much better at differentiating between true signal and highly enriched artifacts than the current one-sample methods.

In general, two-sample approaches are preferred to one-sample ones when appropriate controls are available. In order for a control sample to be 'appropriate', it needs to faithfully capture the background of the treatment sample in question. There are different types of control samples and it is not usually clear apriori which ones are the best. Non-specific Ab controls are known to suffer from the problem of very low amounts of starting material for sequencing. Input DNA controls are easily obtained but are insensitive to any biases introduced during immunoprecipitation and cannot account for antibody promiscuity, among other things. Other control types have their own disadvantages in some situations. In the case of very strong and clear signal, the choice of a control sample is unlikely to matter, but then a one-sample method might perform just as well, and a control sample may be unnecessary.

Another distinction that we draw is that between overlap-based and count-based definitions of enrichment. This distinction is somewhat artificial, since one can usually switch between the two, given that most count-based peak-finding tools use read-shifting as a preprocessing step and thus provide an estimate of the fragment length. Overlap-based methods are more easily extended to paired-ends data, while count-based methods are more amenable to modeling techniques, but overall, there is no inherent advantage to either definition.

All surveyed tools use the strand-specific nature of the aligned read information to guide their peak-finding by either read-shifting or extending the reads to full fragments. The ultimate goal of this technique is to increase the read density or coverage in the immediate neighborhood of a binding event, increasing the likelihood of the enrichment statistic in that neighborhood passing the significance cutoff. This use of strand information is universal in peak-finders targetting transcription factor binding events, and some authors explicitly mention that it improves the performance of their method [4]. The utility of these pre-processing steps in the case of non-punctate events of interest, such as histone modifications, is dubious, but is unlikely to be detrimental.

Most peak-finders report the information about the identified signal sites as a collection of regions with attached point estimates of the binding events (peak summits). We find that when dealing with transcription factor data, only the point estimates are of interest, as they can be used directly to look for possible nearby regulatory targets, and the regions of some given length centered on peak summits can be used for motif-searching. The entire reported peak regions themselves are of interest only if a non-punctate event, such as a long-range histone modification, is being studied. The only other situation where the size of the reported peak region might be of interest is when the binding event is not quite point-like, either due to two or more copies of the transcription factor under study binding in tandem or to interactions with co-binding factors. In this latter situation, it would still be preferred that a single point estimate of the binding event is provided (in the case of interfering co-binding factors) or that point estimates of all individual binding events are obtained (in the case of tandem binding). Some peak-finders address this last issue of multi-peaks by attempting to look for sub-peaks within a peak region using some ad-hoc measures of dips in the peak region read density (FindPeaks, QuEST), while others attempt to model the number of binding events in the region and identify the corresponding point locations (PICS). Most, however, simply ignore this issue, although in some cases, there is an implicit assumption that the small size of sliding window used for identifying peaks will prevent multiple nearby peaks from being merged together. In our opinion, to be successful in as many settings as possible, a peak-finder should be able to handle the multi-peak scenario and that the modeling approaches are preferred when estimating the number of individual peaks in a region.

Most peak-finding publications explicitly address the issue of duplicate alignments (strand-location combinations) as possible PCR artifacts and adopt one of the three possible approaches. They either retain all copies of an alignment, restrict all alignments to a single copy, or use some arbitrary or model-based cutoff for the maximum number of alignment copies retained. Keeping only one copy of each alignment for further analysis can be a disastrous course of action if the data set provides very high genomic coverage, as it will dilute the signal to the level of the background noise. Thus, we believe, that in general, either a model-based approach should be employed or, preferably, all alignment copies should be retained and alternative strategies for discarding potential PCR artifacts should be adopted. Such strategies might compare the number of alignment copies at some genomic position to

the numbers at the nearby positions or to use some information about the size of the region corresponding to the event of interest.

A crucial module of any two-sample peak-finder is the normalization of the treatment and control samples. Loosely speaking, the two basic approaches to normalization are a) the use of some scaling ratio when comparing the candidate region's enrichment in the two samples and b) subsampling reads from the larger of the two samples until the read ratio is optimal in some sense (this includes the approaches that subsample the reads to create extra pseudo-treatment samples). The latter approach results in discarding the data and in lowering one's ability to properly estimate the binding profile. However, the use of scaling ratios by non-subsampling approaches suffers from the downstream effects on statistical significance calculations when the sample sizes are unbalanced as discussed in section 2.3.5. To avoid these effects, some authors recommend using treatment and control samples of comparable size; however, a better approach would be to develop the framework for assessing statistical significance that does not introduce ambiguity by asymmetrically adjusting the raw enrichments in the treatment and control samples. One example of such an approach is that used by CisGenome where the comparison is made based on unadjusted read counts with the normalization ratio incorporated into the p parameter of the assumed conditional binomial distribution. However, even this approach would fail to be adequate in the case of large differences in sample sizes if the resulting p-values are treated as enrichment statistics and a more sophisticated approach, e.g. library-swap, is used to assign the significance (it should be noted that this is not an approach taken by CisGenome). In this case, a larger of the two samples will produce more enriched regions at a range of enrichment statistics cutoffs, and if this sample is the control, the statistical significance assigned to a majority of true binding sites might be quite low. This reinforces the point that careful thought should be given to how a null distribution of enrichment statistics is obtained and whether it is a subject to the unwanted effects of the sample size. We also find that the peak-finders that depend on the existence of very large control samples for proper subsampling and statistical significance calculations (e.g., QuEST and GLITR that utilize the concept of a pseudo-treatment sample) are probably not very suitable for the usual scenarios of relatively similar sample sizes.

We discussed at some length in this chapter the current approaches for assigning measures of statistical significance to the identified binding sites. We re-iterate here that most of these are not based on any verified data-generating model and are not properly calibrated. Any reported p-values or FDR levels should not be taken at their face value but rather used as a ranking tool for assessing the relative confidence in the candidate regions.

Conclusions about the superiority of a particular peak-finding approach will have to wait until proper validation data sets become available. At the moment, it appears that the relative performances of the different tools depend on the data set under study, and prior knowledge of the biological process of interest is required to choose the appropriate set of peaks from among those produced by different methods.

Chapter 3

A Proposed Shape-Based Method for Signal-Noise Deconvolution in One-Sample ChIP-Seq experiments

3.1 Introduction

In this chapter we introduce a novel peak-finding strategy for experiments involving trancription factor binding that lack appropriate control samples and some data sets that we test our method on. We identify candidate binding regions by accounting for local read density variations due to GC and mappability effects. We then classify the candidate regions into specifically and non-specifically enriched by leveraging the knowledge of the shape that the read density profile assumes at the true signal sites. We use deduced sets of true positive and true negative enriched regions to demonstrate that our approach is better at removing non-specifically enriched regions from the set of identified binding sites than other one-sample approaches and provides a superior spatial resolution to most examined peak-finders.

3.2 Data Set Description

We describe briefly here some of the data sets that we will be using in this and the next chapter.

3.2.1 Estrogen receptor β

This is a published data set from [39] that investigates the binding of estrogen receptors α (ERα) and β (ERβ) in U2OS (human bone cancer) cell line. Cell lines that have inducible expression of the two estrogen receptors have been created and are called U2OS-ERα and U2OS-ERβ. The goal of the experiment was to see how the binding in the induced cells

is affected by the presence of estradiol (E2), which can serve as a ligand for an estrogen receptor. Eight samples were prepared and sequenced on 8 lanes of the same flow cell, with one lane devoted to each of the 8 samples. The experimental design is summarized in Table 3.1. Four lanes correspond to the four treatment samles of interest (2 tissue types, each with and without E2), while the other four provide the corresponding control samples. Non-specific IgG antibody was used to produce the controls. We will only consider U2OS-ERβ, no E2 sample (referred to as ERβ sample) and its control (sometimes referred to as ERβ BG) in this work.

Tissue	E2 added (Y/N)	Ab	Number of reads
U2OS-ERα	Y	anti-ERα	0.9M
U2OS-ERα	N	anti-ER α	3.8M
U2OS-ERβ	Y	anti-ER β	4.9M
U2OS-ERβ	N	anti-ER β	5.2M
U2OS-ERα	Y	IgG	2.4M
U2OS-ERα	N	IgG	4.8M
U2OS-ERβ	Y	IgG	5.0M
U2OS-ERβ	N	IgG	5.2M

Table 3.1: Experimental design of ERβ data set

3.2.2 PIF3 data set

This is an unpublished data set on the binding of transcription factor PIF3 in *A.thaliana*, a grassy plant. The plant has been mutated to express a hybrid PIF3-Myc protein, where Myc is a transcription factor absent from the wild type (WT) plants. Myc-specific antibody is used in the IP step and the same procedure with a WT plant yields a control sample. Input DNA samples have been obtained as well for both mutant and WT plants, but we will show that they are inadequate at capturing the treatment (mutant) sample background.

There are two *biological* replicates of both mutant (P3M) and WT samples, obtained from different plants. These are referred to as experiment 3 (exp3) and experiment 4 (exp4). For each biological replicate, two *technical* replicates of both P3M and WT IP samples were obtained by running sub-samples of the same library on different flow cells. The organization of samples by flow cell is shown in Table 3.2. The three flow cells were sequenced at intervals of at least several weeks. The technical replicates in this setup (same library, different flow cell) are just one step away from being as closely related as possible (same library, same flow cell) and one fully expects them to show very similar enrichment profiles. See section 4.2 for a discussion of biological and technical replicates.

Flow Cell 1	Flow Cell 2	Flow Cell 3
Exp3 P3M ChIP rep1 (5.7M)	Exp4 P3M ChIP rep1 (5.1M)	Exp3 P3M ChIP rep2 (5.0M)
Exp3 WT ChIP rep1 (5.9M)	Exp4 WT ChIP rep1 (6.5M)	Exp3 WT ChIP rep2 (7.9M)
Exp3 P3M Input (5.6M)	Exp4 P3M Input (3.7M)	Exp4 P3M ChIP rep2 (8.2M)
Exp3 WT Input (6.9M)	Exp4 WT Input (5.9M)	Exp4 WT ChIP rep2 (8.2 M)

Table 3.2: Partial experimental design of PIF3 data set. The numbers of reads for each lane are given in parentheses. Flow cell 3 tends to have more reads as it was done later than the other two flow cells and the sequencer's performance has been brought closer to optimal.

3.2.3 NRSF Monoclonal Ab

Another data set we look at in some depth is the study of the binding of the transcription factor NRSF in Jurkat human T lymphoblast cell line from [29]. The study provides raw reads for two NRSF treatment samples: one immunoprecipitated with a monoclonal antibody and another with a polyclonal antibody. A large sample of input DNA is provided as well. It is not clear how the samples were distributed across flow cells or lanes, but it seems likely that several lanes worth of data were combined for each sample.

3.2.4 Other Data Sets

We will briefly mention some other published data sets, providing references where appropriate. The references can be consulted for the experimental details.

3.3 Motivation

Most of the authors of the peak-finding tools described in chapter 2 highly recommend using the control samples for background assessment. However, as discussed in section 2.3.1, the proper choice of the type of the control sample to use is rarely obvious, and in our own work we have encountered situations when some of the commonly employed types of controls failed to capture the background artifacts that they were meant to filter out.

We have discovered multiple examples of regions enriched in the treatment sample relative to IgG control but not representing the true binding signal in ERβ data set. An example of such region is shown in Figure 3.1. This 3kb-long region on chr12 contains 97 reads in the treatment sample and 39 reads in the control, resulting in enrichment p-values of ≈ 0 under the global Poisson model (the treatment and control global Poisson read count rates for 3kb bins are 5.45 and 5.04, respectively). After adjusting for the lane totals, the region enrichment of treatment over control is more than 3-fold, yet it is not due to the actual binding event. This claim is based on the fact that the region lacks both the pattern of highly concentrated clusters of strand-specific reads on the opposite sides of the true binding site and the canonical ERE binding motif, and is found to be similarly enriched in other ChIP-Seq data sets, e.g. the data sets on binding of transcription factors GABP [29] and STAT1 [2]. These latter data sets were collected using Jurkat and HeLa S3 cell lines, respectively.

We also found an extreme example of how misleading a widely used control type can be in the PIF3 data set. Running the routine data analysis with the commonly used peak-finders identifies only about 20% of the number of peaks found when the IP WT control is used instead of the input DNA. It is clear that in this particular data set, the input DNA control cannot adequately capture the background, and relying on it would lead to erroneous conclusions.

These observations have motivated us to develop a functional way to deconvolve signal and noise in the ChIP-Seq treatment samples in the absence of a suitable control sample. In this setting the challenge is to find a set of classifying features that can be used to separate the true binding sites from artifactual enrichment. Most one-sample peak-finders described in Chapter 2 use raw enrichment measures for classification, e.g. declaring any region with $> k$ reads a true binding site and all regions with $\leq k$ reads to be noise. This approach might sometimes be adequate (as mentioned in section 2.1.1, in the presence of very strong signal and no strong noise, one- and two-sample approaches produce very similar results), but could be very misleading if the overall signal is weak relative to the background enrichment. Figure 3.2 shows a plot of counts in genomic 1kb bins in a treatment sample vs. an input DNA sample from the PIF3 data set. As seen from this plot, using a raw enrichment measure (e.g. number of reads in a bin) as a ranking device would lead to the set of the top detected peaks consisting entirely of artifacts (as the regions also showing extreme enrichment in control samples must be). In general, the deficiency of signal/noise classification approaches based on simple enrichment measures alone is the reason why many authors strongly recommend

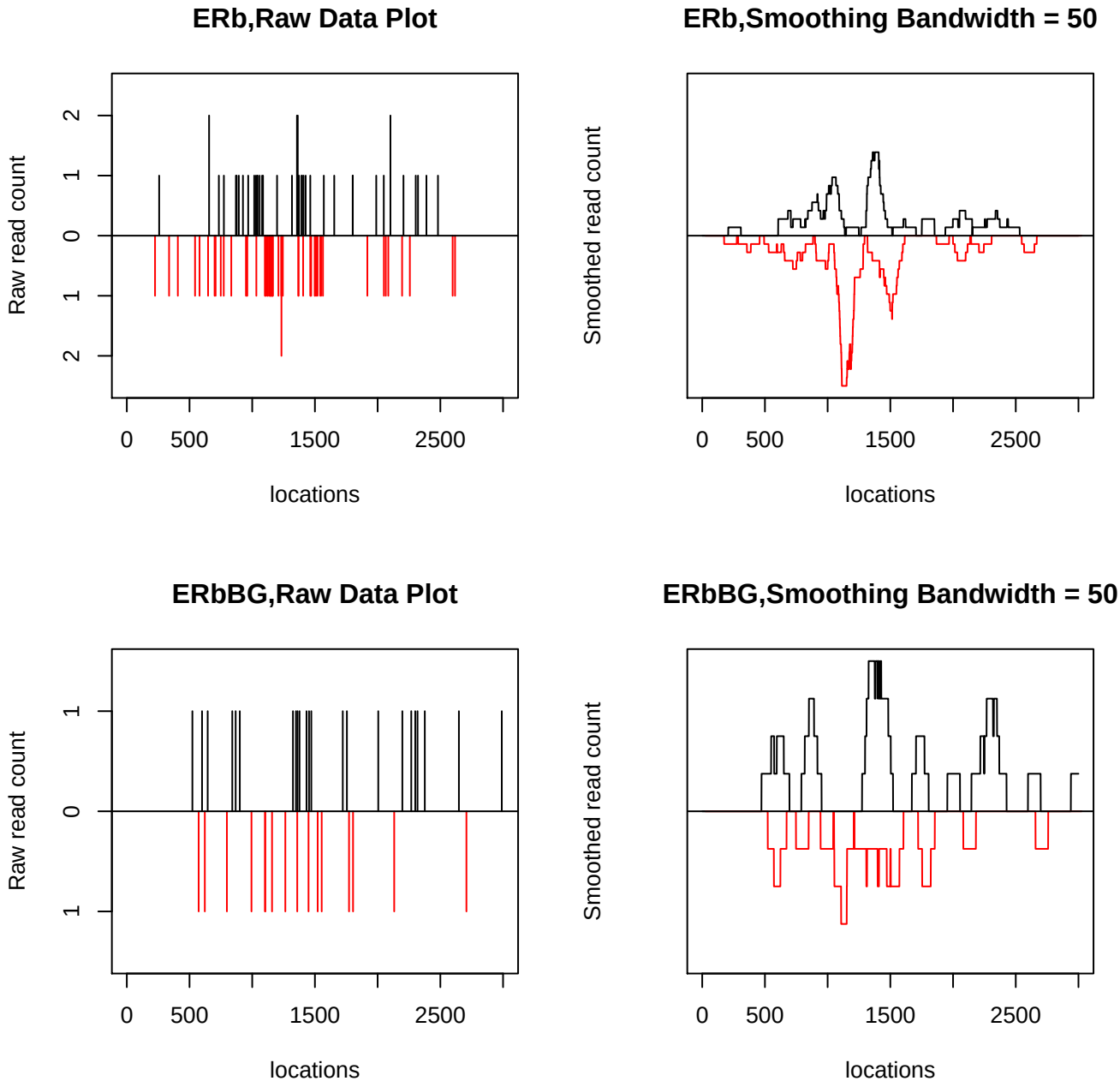

Figure 3.1: An example of a non-specifically enriched region in ERβ dataset. Black lines show poistive-strand (raw and smoothed) reads and red lines show negative-strand reads. ERβBG is the control IgG sample. The origin along the x-axis corresponds to position 13141501 on chr12 (hg18 assembly)

.

the use of the control samples.

We propose to use our knowledge of the peak shape that the read density at the true binding event site takes to develop a classifier to distinguish between signal and noise. Some examples of raw and smoothed read densities at binding sites supported by gene expression data for ERβ data set are shown in Figure 3.3. All of them show the previously described pattern of clustered reads of opposite strands upstream and downstream of the putative binding site. When smoothed, the densities show the characteristic strand-separated peak shape (SSPS). For comparison, various types of non-specific enrichment are shown in Figure 3.4. The top panel of Figure 3.4 shows an enriched region that might correspond to the binding of a different TF (GABP is shown to bind here [29], although in a different tissue type) and is probably a region of open chromatin. The second panel from the top shows an example of a 'spike' artifact - multiple reads mapping to very few locations in close proximity to each other. Some spikes are experiment-specific and some (like this one) seem to occur in several (but not all) of the examined ChIP-Seq experiments. The third panel from the top shows an example of a hyper-enriched region that is found in all of the examined data sets. Usually these types of artifacts are found to occur in centromeric and telomeric regions, although this one is not (it is located on chr1, 44392501-4439400, hg18 assembly of human genome). Finally, the bottom panel shows an artifact (found to occur in multiple examined data sets) that could possibly be mistaken for a true binding event based on the overall peak shape. However, it is wider than the peaks in Figure 3.3 and the two strand-specific peaks are not as separated. We will leverage these two attributes of the shape of the observed read density to filter out such artifacts. Interestingly, the regions in the bottom two panels have been annotated in multiple cell lines with an H3K4Me3 mark and have been classified as DNAseI hypersensitive sites as per UCSC Genome Browser annotation.

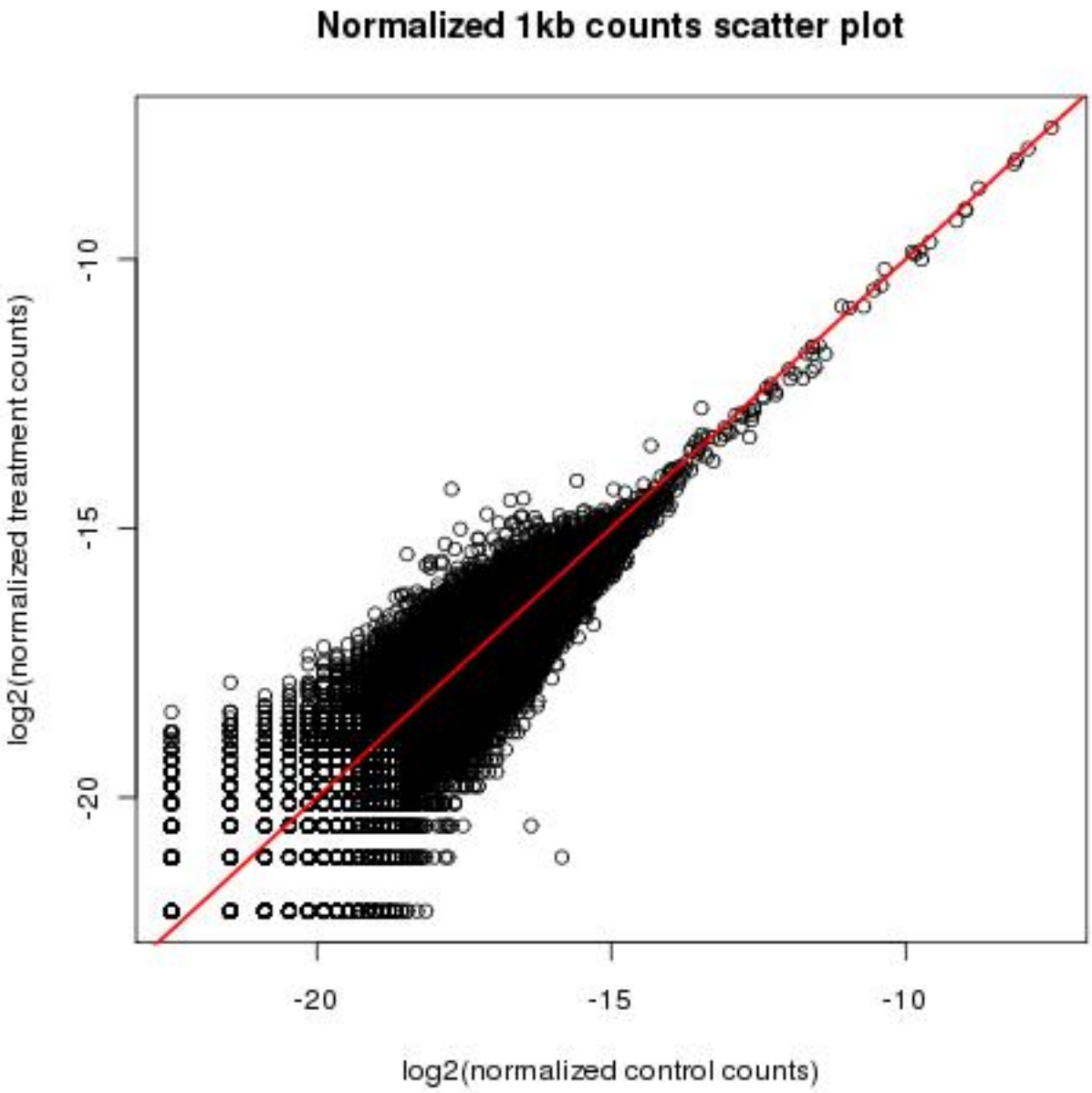

Figure 3.2: Scatter plot of treatment (mutant) vs. control (WT) read counts in genomic 1kb bins for PIF3 experiment, divided by lane totals (on $\log_2$ scale). Red line corresponds to equal normalized counts.

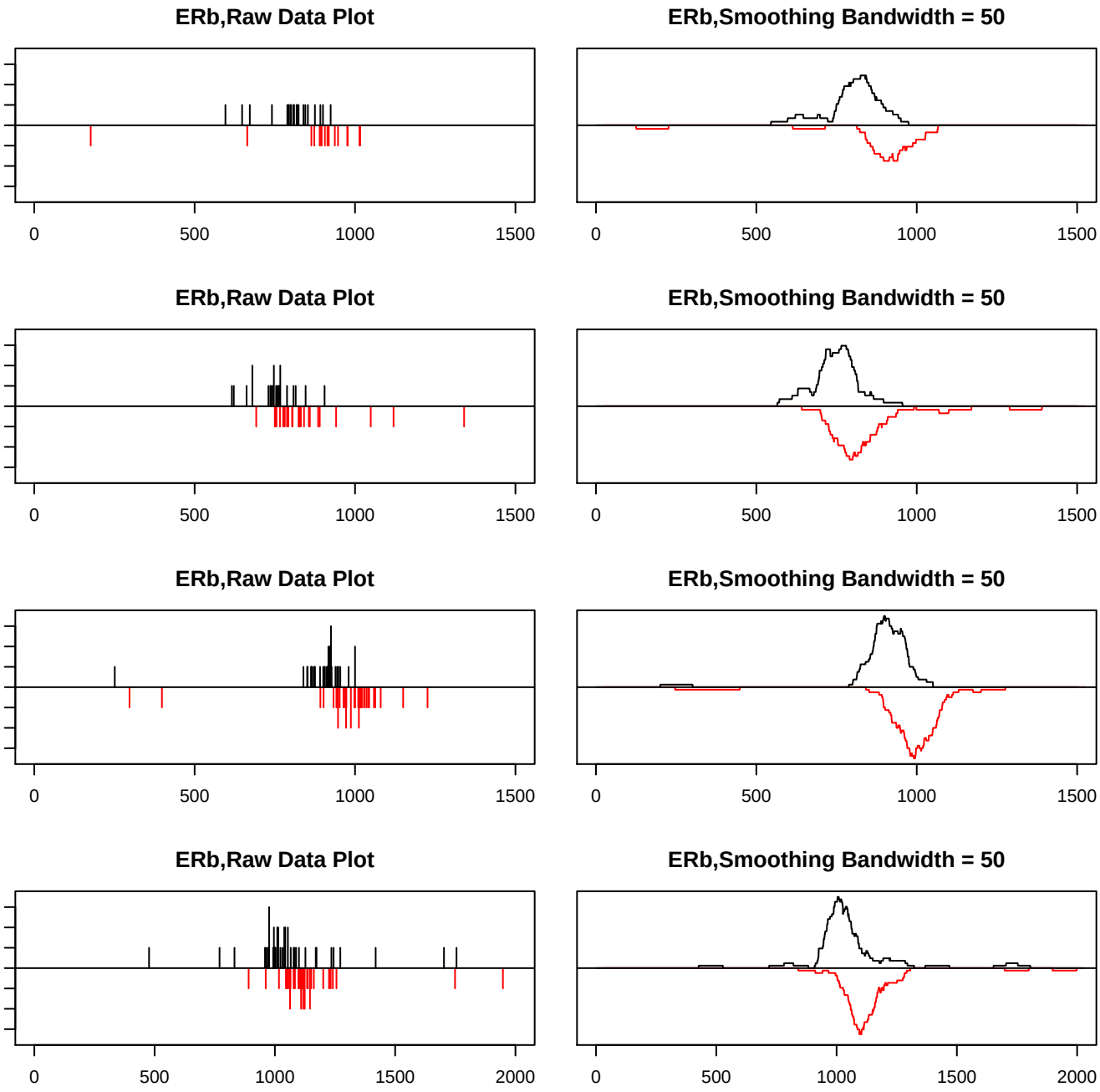

Figure 3.3: Some examples of read density at the true binding sites in ERβ data set (confirmed by gene expression data). The densities show the characteristic strand-separated peak shape.

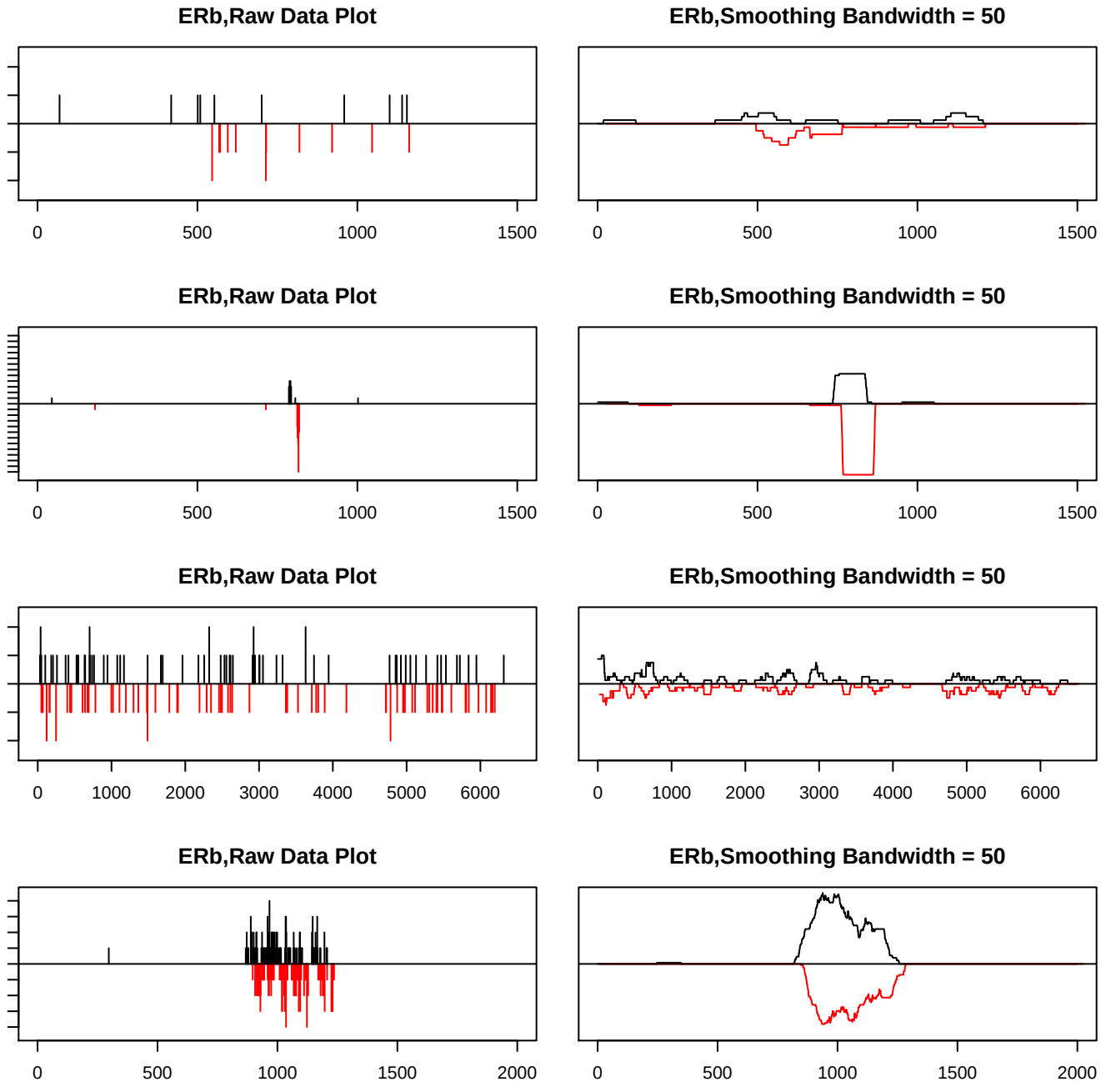

Figure 3.4: Some examples of read density at artifactual regions in ERβ data set.

Our approach to solving the problem of signal-noise deconvolution in the absence of control samples is to classify enriched regions based on whether or not they have SSPS. Briefly, we

1. Model the read start site process that would give rise to such a shape.
2. Identify candidate (enriched) regions.
3. Fit the model to the candidate regions.
4. Assess the model goodness-of-fit.
5. Reduce the information on the model fit in each region to a set of classifying features.
6. Estimate the set of the true binding events based on these features using an iterative classifier.

It should be pointed out that the approach presented here is not suitable for non-punctate events, such as histone modifications or tandem binding of transcription factors that do not manifest themselves as narrow peaks. Also, it is conceivable that there might be sites of punctate non-specific enrichment that show the same read density pattern as the true binding events. One obvious example of this would be the non-specific precipitation of proteins other than the one under study due to a promiscuous antibody. The developed method depends on the existence of the separation between read density profiles of the population of true binding sites and the population of artifactual enrichment sites and if the separation does not exist (as is the case with the very wide or very narrow true binding events or an excess of point-like artifactual enrichment), it will perform poorly.

3.4 Model Description

We assume a point process model for the read (fragment) start sites along the genome. Specifically, for a genomic position t, let X_t^+, X_t^- be independent random variables corresponding to the number of positive and negative strand reads at t. Let N be the number of the binding events along the genome, with point binding locations $b_1, \ldots, b_N$. Then we model strand-specific read counts at location t as $X_t^s \sim \text{Poisson}\left(\lambda^s(t) = \lambda_{BG}^s(t) + \sum_{i=1}^{N} \lambda_i^s(t)\right)$, where $s \in \{+, -\}$. λ_{BG}^s is the background rate function, conceptually representing the strand-specific rate at location t in absence of any binding events. Since the background process is assumed to be unbiased with respect to strand, we have $\lambda_{BG}^+ = \lambda_{BG}^- = \frac{1}{2}\lambda_{BG}$, where λ_{BG} is the overall read background rate. We do not assume that λ_{BG} is constant throughout the genome but we do assume that it is smooth enough that we can treat small genomic regions as having constant region-specific background rate. λ_i^s is the strand-specific contribution to

rate at t due to a binding event at b_i. Since the data-generating process involves a step of fragment size-selection, $\exists\ k$ s.t.

$$|t - b_i| > k \Rightarrow \lambda_i^+(t) = \lambda_i^-(t) = 0$$

Of course $\lambda_i^+(t) = \lambda_i^-(t) = 0$ trivially if t and b_i are on different chromosomes. The rate contribution due to a binding event is strand-specific and the nature of the data dictates that given $|t - b_i| = d < k$, we should have $\lambda_i^+(t) > \lambda_i^-(t)$ if $b_i > t$, and $\lambda_i^+(t) < \lambda_i^-(t)$ if $b_i < t$. We choose to ignore the effects of mappability at this bp-level resolution, but they can conceivably be incorporated with indicator variables.

We want to estimate the number of binding events N and their locations $b_i, i = 1, \ldots n$ based on the observed values x_t^+, x_t^- of X_t^+, X_t^-. In order to do this, we first identify an initial set of enriched regions (this process is detailed in section 3.5) and then fit the model to the observed read counts in the region under the assumption that the region contains $n = 1$ binding events, using a parametrized functional form for λ_j^s, where j is the index of the associated binding event. The assumption of only 1 binding event in a region can be relaxed as discussed below.

We model the strand-specific read rate due to a binding event as a symmetric peak-like shape that is the same for both strands. This is motivated by the observed behavior of strand-specific read rates, with read density increasing and then decreasing in proximity to a binding event site (see Figure 3.3). Thus for a location t in proximity of a binding event, our model has $\lambda^+(t) = \lambda_{BG}^+(t) + \frac{a}{c} S\left(\frac{t-b}{c}\right)$ and $\lambda^-(t) = \lambda_{BG}^-(t) + \frac{a}{c} S\left(\frac{t-b-e}{c}\right)$, where S is some normalized peak shape function with center parameter 0 (centered on the origin), width parameter 1 (the actual meaning of 'width' depending on the shape) and height normalization requirement $S(0) = 1$. In the expressions above c and $\frac{a}{c}$ are the binding event-specific measures of peak width and height, respectively, while e is the measure of the distance between the centers (b and $b+e$) of the two strand-specific peaks. In this setup the point binding location corresponds to $t = b + (e/2)$. A schematic illustration of our model is presented in Figure 3.5. An examination of multiple binding sites from several data sets has shown that while there are some stochastic deviations from symmetry for strand-specific peaks (small 'bumps' that are evident in Figure 3.3), there are no gross systematic ones.

Sometimes, multiple binding events can affect the local read density. In such cases we model the strand-specific read count rates as $\lambda^+(t) = \lambda_{BG}^+(t) + \sum_{i=1}^{n_t} \left(\frac{a_i}{c_i} S\left(\frac{t - b_i}{c_i}\right)\right)$ and $\lambda^-(t) = \lambda_{BG}^-(t) + \sum_{i=1}^{n_t} \left(\frac{a_i}{c_i} S\left(\frac{t - b_i - e_i}{c_i}\right)\right)$, where n_t is the number of binding events contributing to the read density at t.

We have investigated the issue of the read rate shape further through the use of simulation of the data-generating process. For the purposes of the simulation, we have defined the genome to be a 10kb region with a point location of the binding event b=5kb. We have simulated the chromatin shearing along a single copy of the genome by assuming that breaking

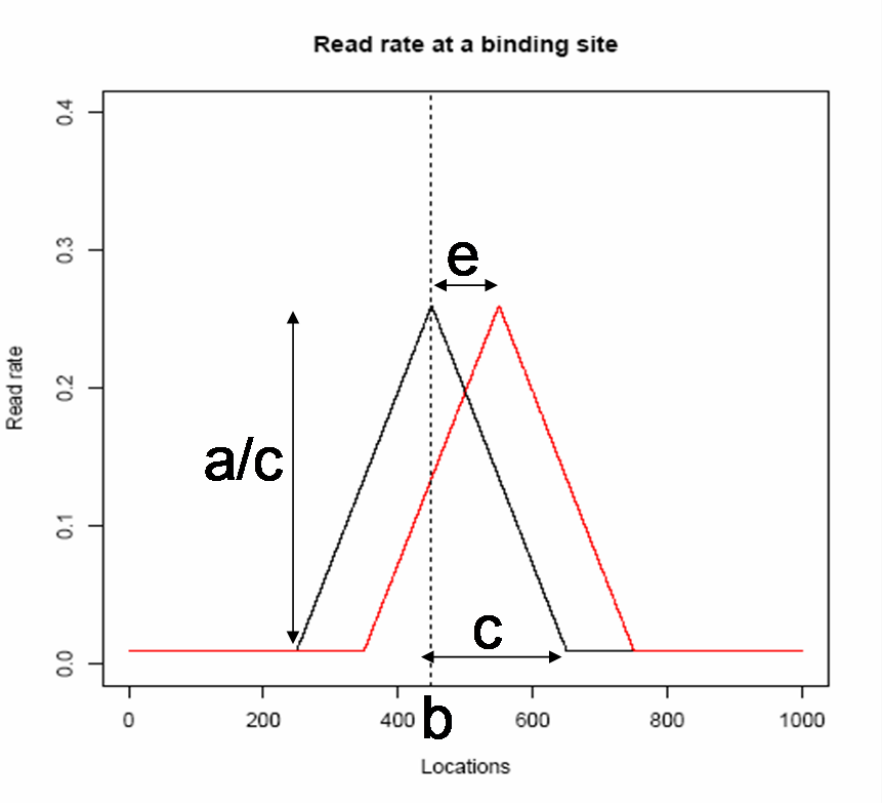

Figure 3.5: Illustration of our shape model in the case of triangular shape. $\lambda^+(t)$ is plotted as a black curve and $\lambda^-(t)$ is plotted as a red curve. Since $\lambda^+_{BG} = \lambda^-_{BG}$, the background rates (flat lines outside triangles) overlap in the plot.

events are independent with P(break at t)=$1/L$, where L is the average fragment length and have collected the fragment covering b, repeating this for many copies of the genome. We proceed to select the fragment lengths according to a size selection specifications and look at the distributions of the positive- and negative-strand reads (the two ends of the fragments) around the point binding site.

There are several changes that need to be done to this basic setup to make it approximate the behavior of the real data. First of all, under the scenario of a fixed point binding site, one would have all of the positive-strand reads upstream of the binding event and all negative-strand reads downstream (cf. Figure 1.1), whereas in the real data the ranges of strand-specific read clusters overlap, with the size of the overlap on the order of 50-100bp or so, depending on the data set. For this reason, in our simulations we allow the binding event to take place anywhere within some distance d of b, where we can allow the probability of binding at $t \in (b-d, b+d)$ to be uniform or to decrease as $|b-t|$ increases, in which case we use the triangular probability distribution. Letting $d = 40$bp, $L = 100$bp and setting size

selection limits at 50-250 bp (reasonable settings for the data sets we have examined), we run the simulations as described above and plot the number of reads at each location, normalized by the number of simulations. The results for the uniform and triangular distribution of the location of the binding event in $(b-d, b+d)$ are shown in the top panel of Figure 3.6. The strand-specific read density profiles seem to be mostly symmetric but with heavier tails in the direction away from the binding site, corresponding to longer fragments. As mentioned in section 1.4, the sequencing procedure favors shorter DNA fragments and thus we do not tend to observe these heavy outward tails in the real data, apart from the very deeply sequenced regions. The bottom panel of Figure 3.6 shows the read density profiles under the assumption that none of the fragments in the 200-250bp range are detected.

Motivated by the observations from the real data and simulations, we choose a triangular shape for strand-specific read counts rates, $S(x) = \max(0, 1-|x|)$. With this shape function, the parameter c corresponds to the half-base of the triangle.

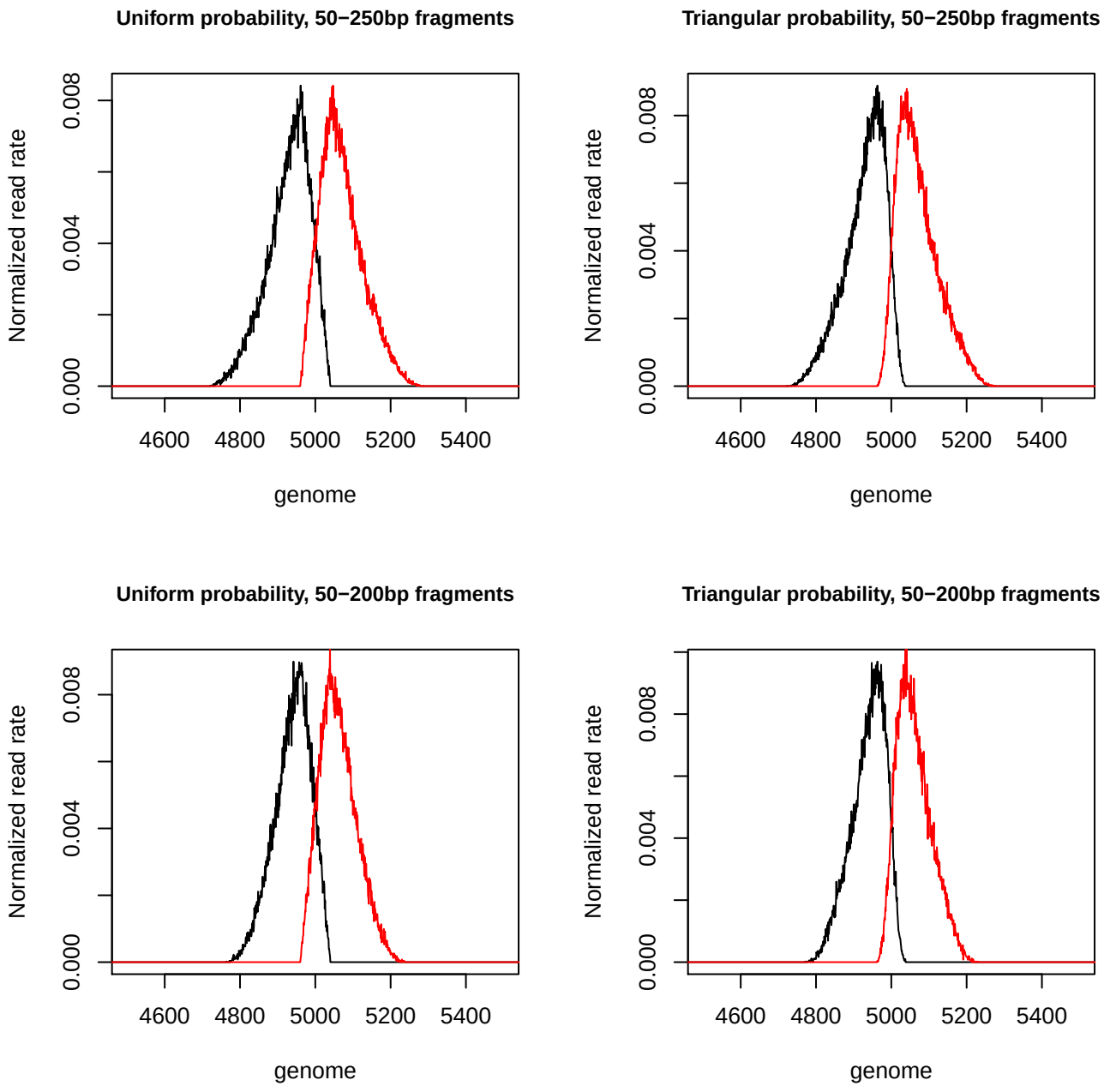

Figure 3.6: Read density profiles obtained in simulations. Positive- and negative- strand profiles are shown in black and red, respectively.

Starting with a set of candidate regions and their estimated background read rates λ_{BG} (see section 3.5), we fit the model separately to each candidate region R. The maximum likelihood estimates of region-specific shape parameters a, c, b, e under the above assumptions about the underlying Poisson point process are obtained by maximizing

$$l(a,b,c,e) = \log\left(\prod_{t\in R} P(x_t^+, x_t^-|a,b,c,e,\lambda_{BG})\right) = \sum_{t\in R}\left(x_t^+ \log(\lambda^+(t)) - \lambda^+(t)\right) + \sum_{t\in R}\left(x_t^- \log(\lambda^-(t)) - \lambda^-(t)\right)$$

with the expressions for $\lambda^s(t)$, $s \in \{+,-\}$ as given above. The actual maximization is done by using *optim()* function of statistical computing environment R [40] and the results of the fit are used to classify the candidate regions into true binding events and arifacts as described in section 3.6.

This approach generalizes easily to the situation where $k > 1$ binding events are assumed to influence the read density in the candidate region. In such cases the maximimum likelihood estimation reduces to maximizing

$$l(\{a_i\}_{i=1,\ldots,k}, \{b_i\}_{i=1,\ldots,k}, \{c_i\}_{i=1,\ldots,k}, \{e_i\}_{i=1,\ldots,k}) = \log\left(\prod_{t\in R} P(x_t^+, x_t^-|a,b,c,e,\lambda_{BG})\right) = \sum_{t\in R}\left(x_t^+ \log(\lambda^+(t)) - \lambda^+(t)\right) + \sum_{t\in R}\left(x_t^- \log(\lambda^-(t)) - \lambda^-(t)\right)$$

where we use the expressions for $\lambda^s(t)$, $s \in \{+,-\}$ given above for the case of mutliple binding events. The rates due to the individual binding events are superimposed and the combined rate due to overall binding is convolved with the underlying background rate.

3.5 Identifying Candidate Regions

Fitting the model to all genomic regions of certain size with at least 1 read in it is computationally prohibitive, so a filtering procedure is required to identify an initial set of candidate regions. Naturally, these regions should be read-rich, but the constant enrichment cutoff might not be appropriate as the genomic enrichment profile is subject to various local biases as detailed in section 1.4. Some of the bias-inducing features one might consider in declaring a region to be a candidate are GC and mappability content (the two major influences on the low-level background noise), as well as the features based on the genome and epigenome annotation, e.g. proximity to telomeric/centromeric regions, DNAse I hypersensitive sites, satellite repeats, and other features that result in very high levels of non-specific enrichment. Our approach is to let the classification based on the shape-model fitting deal with discarding the hyper-enriched artifacts, and to only adjust for the low-level background enrichment variation when identifying the set of the candidate regions.

The dependence of read counts on mappability and GC content can be seen in Figure 3.7 for the ERβ treatment sample. We take all 1kb genomic bins along the genome (except for the non-mappable ones) and break them into 100 categories based on the quantiles of GC and mappability content. For each category, the average read count is calculated and its ratio to the global average is plotted on $\log_2$ scale. The categories with higher GC and mappability content show higher average counts illustrating the dependence of the global background enrichment profile on these features ($> 99.9\%$ of all 1kb genomic bins represent background). We conclude that the approach taken by many one-sample peak-finders in choosing their candidate regions based on a global cutoff (usually, derived from the global Poisson model) will result in the over-representation of the high-mappability, high-GC regions among the candidates, possibly at the expense of some weak signal.

We propose to adjust the enrichment in 1kb bins along the genome for GC and mappability content and to take the bins that exhibit post-adjustment enrichment as our set of the candidate binding sites. This approach significantly reduces the number of regions under consideration for the binding site detection, and increases the representation of the weak signal regions in the set of moderately-enriched candidates at the expense of the background regions with non-specific GC- and mappability-driven moderate enrichment. This first pass will not dispose of the hyper-enriched artifacts, which will be dealt with by the shape model-fitting step. The choice of 1kb bin size as a basis for the binding event identification was motivated by the observation that all of the information (i.e. reads) pertaining to a single binding event is normally contained in the interval of that length due to the fragment size selection, and, at the same time, this interval is usually small enough to preclude incorporating information from multiple binding events that would compromise our ability to fit a 1-peak model.

We assume that for each 1kb bin j, the underlying background rate at each location t in bin j is constant: $\lambda_{BG}(t) = \lambda_{bp,j}$, where bp indicates bp-level read count rate. Letting $\lambda_{BG,j} = 1000\lambda_{bp,j}$ be the overall read occurrence rate in bin j and X_j be the random variable representing the number of reads in bin j, our model is $X_j \sim \text{Poisson}(\lambda_{BG,j} = \lambda_0 F_{m,j}(F_{GC,j})^k), k > 0$, where λ_0 is the underlying overall rate that incorporates information about the total number of reads in the sample, $F_{m,j}$ is the fraction of mappable locations in the bin, and $F_{GC,j}$ is the fraction of GC bases in the bin. According to this model, the background rate is linearly proportional to the sample size and the number of mappable locations in the bin. The functional form chosen for the dependence on GC is not theory-driven but rather aims to capture the relationship empirically. The model is fit by standard Poisson regression techniques to the set of consecutive non-overlapping 1kb bins and, separately, to the set of consecutive non-overlapping 1kb bins with 500bp offset from the chromosomal start positions. For each set of bins, p-values based on Poisson distributions with estimated bin-specific background rates $\widehat{\lambda}_{BG,j}$ and observed read counts x_j are caclulated. Bins passing some statistical significance cutoff are retained as candidate binding sites, and the overlapping bins between the two sets are merged to yield the final set of non-overlapping candidate regions of variable length.

This simple Poisson regression model captures the global dependence of read counts on GC and mappability quite well as can be seen in Figure 3.8. Similar results are seen in all of the data sets we have examined. This indicates that we estimate the bin-specific background rate successfully, although the Poisson distributional assumption might not fully capture the complexity of the read count distribution, as discussed below. One consequence of employing the mappability-based offset $F_{m,j}$ is that fitting the model to the set of low-mappability hyper-enriched artifacts (possible assembly collapse sites) results in very small estimated rates and therefore extremely high average ratio of observed counts to estimated rates for the lowest mappability bin categories. It also bears pointing out that the model is fit to a mixture of backgound and signal bins and the fit might be driven by the hyper-enriched regions.

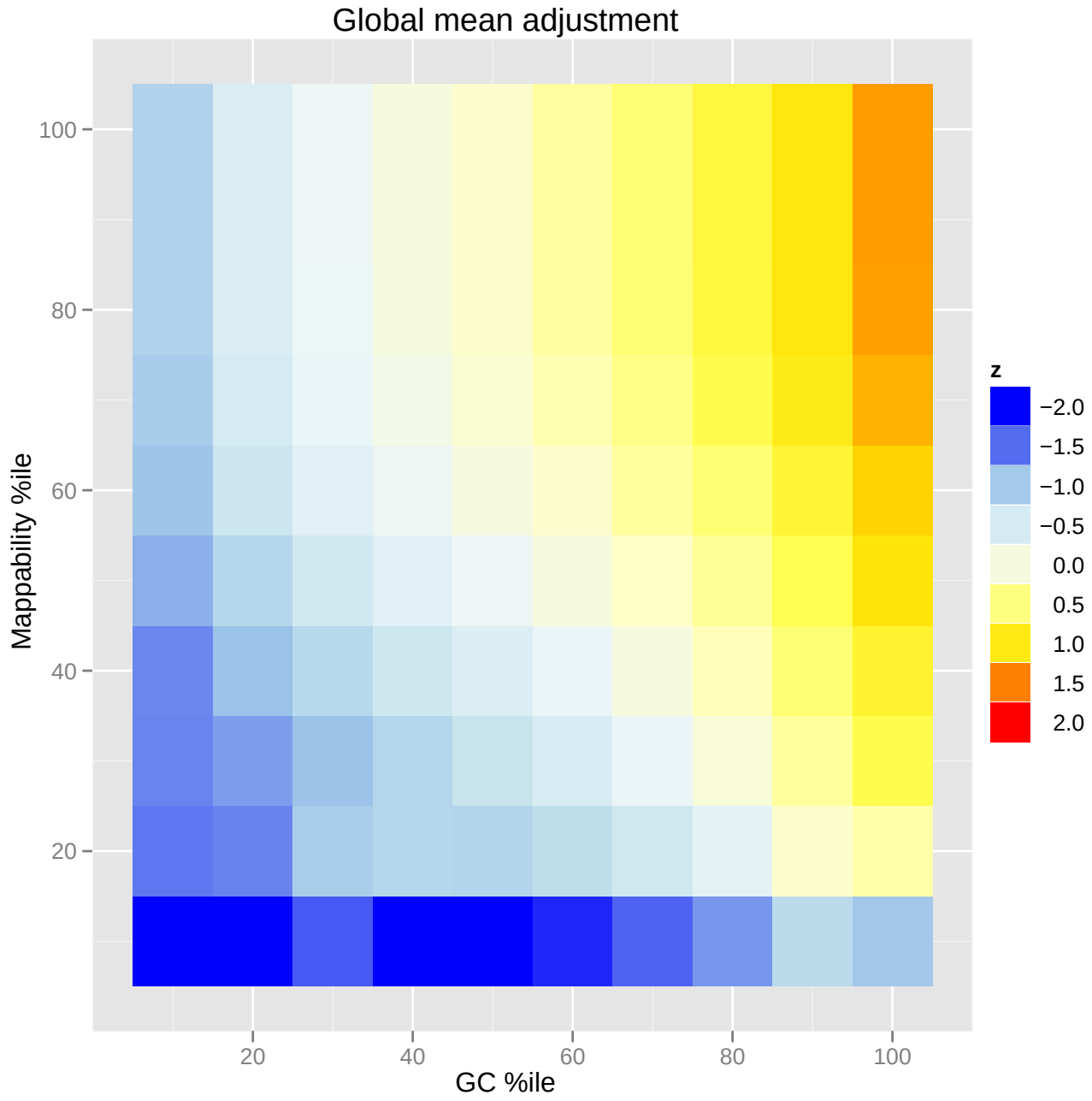

Figure 3.7: The effects of GC and mappability content on read count (1kb bins) for the ERβ treatment sample. Genomic bins are broken into 10 categories based on deciles of mappability content and 10 categories based on deciles of GC content, then crossed to yield 100 total bin categories. For each category, $\log_2(\text{mean}(x_i/M_g))$ is plotted, where x_i is the read count of bin i in that category and M_g is global average of read counts in 1kb bins. Non-mappable bins have been excluded for the plot.The read counts increase as GC and mappability contents increase.

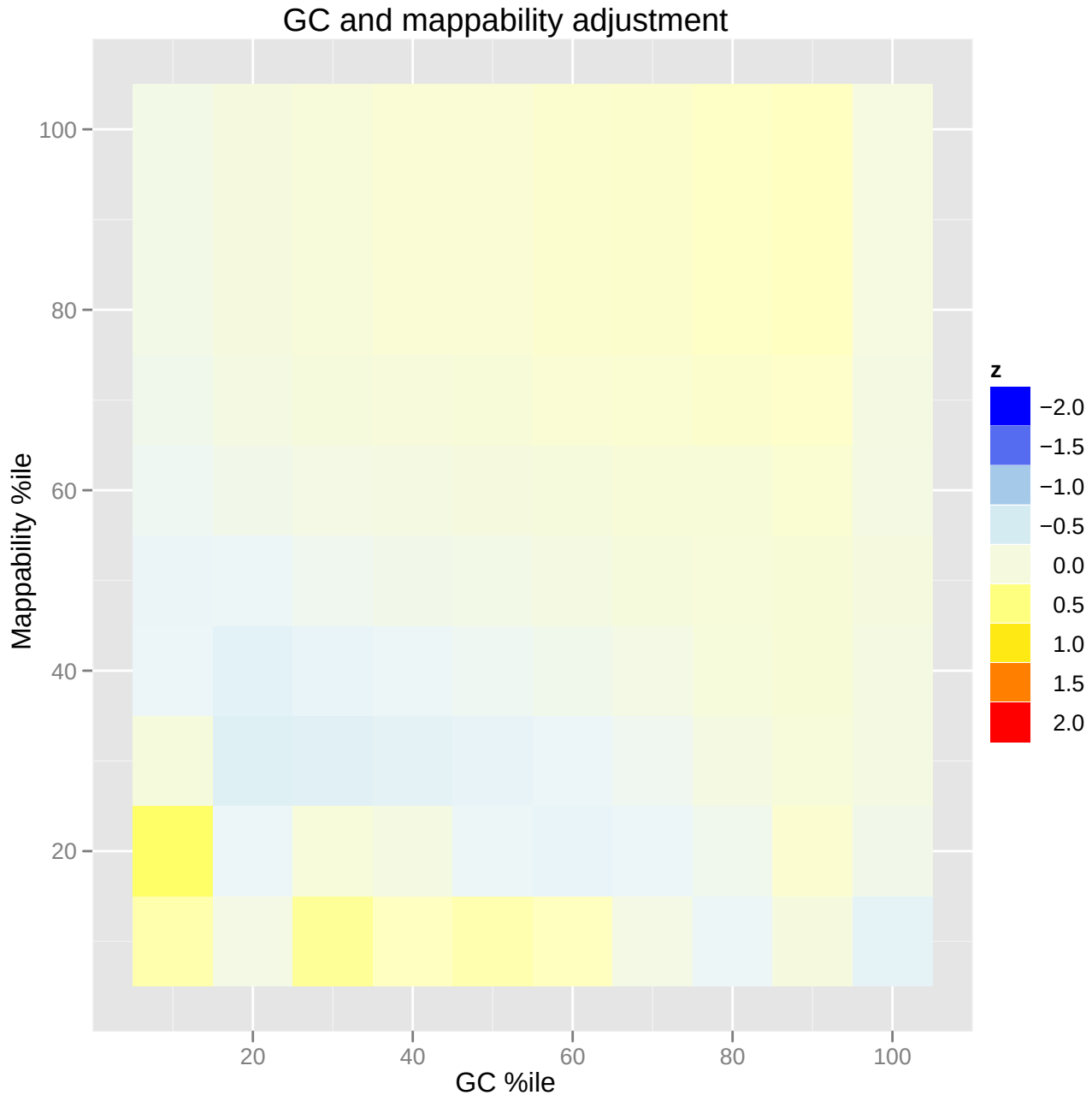

Figure 3.8: The results of model adjustment for GC and mappability effects on read count (1kb bins) for the ERβ treatment sample. Genomic bins are broken into 10 categories based on deciles of mappability content and 10 categories based on deciles of GC content, then crossed to yield 100 total bin categories. For each category, $\log_2(\text{mean}(x_i/\widehat{x_i}))$, is plotted, where x_i is the read count of bin i in that category and $\widehat{x_i}$ is the estimated rate. Non-mappable bins have been excluded for the plot. The dependence on mappability and GC content is much less pronounced than in Figure 3.7.

To the best of our knowledge, the only other one-sample peak-finder that attempts to model local enrichment directly is MOSAiCS [22]. The authors use an overlap-based approach by extending reads to the average fragment length L and model the number of fragments overlapping locations in bin j as $y_j \sim \text{NB}(a, a/\mu_j)$, where $\mu_j = E(y_j) = exp\left(\beta_0 + \beta_M \log_2(F_{m,j}) + \beta_{GC} Sp(F_{GC,j})\right)$ and Sp refers to a B-spline model with knots at the quartiles of the distribution of GC content. Unlike our approach, MOSAiCS uses bins of size 50bp and their definitions of F_m, F_{GC} are somewhat different from those employed by us. For each location t its MOSAiCS GC score is the fraction of GC content in the window $(t - L, t + L)$, with mappability score defined analogously, and the average scores across locations in bin j serve as $F_{m,j}, F_{GC,j}$. The overdispersion parameter is assumed to be the same for all bins.

This overlap-based approach is not directly comparable to our count-based method, since different quantities are being modeled and different bin sizes are employed. However, we can try to use MOSAiCS model with bins of size 1kb and using an approximation $x_j \approx y_j \frac{1000-L}{1000+L}$ as well as substituting our definitions of $F_{m,j}, F_{GC,j}$ into the MOSAiCS model formula. The reasoning behind the approximation $y_j \approx x_j \frac{1000+L}{1000-L}$ is that for 1kb bins and constant fragment length $L < 1000$, the fragments due to the reads outside of the bin itself (i.e. positive-strand reads upstream of the bin and negative-strand reads downstream of the bin) constitute $2L/(1000+L)$ of the total number of overlapping fragments, under the assumption of uniform distribution of fragments in the neighborhood of the bin. Therefore the fraction x_j/y_j of fragments due to the reads inside of the bin is $\frac{1000-L}{1000+L}$. We justify using our definitions of $F_{m,j}, F_{GC,j}$ by assuming that the mappability and GC contents of the regions flanking the bin are not drastically different from those of the bin itself. We will use $L = 100$ in our comparison and will round y_j to the nearest integer.

The only other one-sample approach that tries to avoid using a global significance cutoff for selecting candidate binding sites is that of MACS [4], that uses the average per-bp read count in the surrounding region of several kbs in size as the estimate of the local background rate. Following the MACS approach we can estimate the background rate in a 1kb bin by the per-kb read count in the surrounding 5kb window.

To assess how well these different approaches model the background rate, we divide the genomic bins into 100 categories based on GC and mappability content and for each category i obtain the average adjusted rate on $\log_2$ scale: $\log_2\left(\frac{1}{N_i}\sum_{n=1}^{N_i}\frac{x_{i,n}}{\widehat{x}_{i,n}}\right)$, where N_i is the number of bins in category i, and $x_{i,n}, \widehat{x}_{i,n}$ are the observed read count and the estimated rate for bin n in category i, respectively. The boxplots of the adjusted rates for different background models are shown in Figure 3.9. Our Poisson regression model is an improvement on the global Poisson model and seems to perform better than MACS-style local average. Our approximation to the MOSAiCS approach results in a lot of under-adjusted counts. Interestingly, if we change the response variable in the MOSAiCS model from the number of fragments overlapping a bin to the number of reads in a bin, the performance of this

estimator is superior to our own model ('MOSAiCS new' in Figure 3.9). This improvement is achieved through the use of more parameters in the model relating read counts to GC and mappability (4 parameters in MOSAiCS-style model vs. 1 parameter in our model). Thus, while we feel that our adjustment is sufficient, a better one is possible and can be achieved by using a modified version of the MOSAiCS approach.

Another major difference between our model and the MOSAiCS-style approach is the distributional assumption on the read counts. We model the background read count in 1kb bins as Poisson, while the modified MOSAiCS approach uses the negative binomial (overdispersed Poisson) distribution. Since the model is fit to a mixture of the low-background and the hyper-enriched (signal + artifacts) bins, there will be overdispersion present in the data. To see this, we can scan the genome for 1kb bins that have the same mappability and GC content and look at the mean and variance of the read counts in those bins. Under both models, such bins have identical background rates that depend only on the mappability and GC content of the bin. To avoid dealing with undesired low-mappability effects, we restrict ourselves only to bins with 100% mappable bases. Using this approach with ERβ data, we choose 49 values of GC content $0 < GC_i < 1$ such that there are at least $n_i \geq 10$ 1kb bins in the human genome with GC content $= GC_i$ and mappability $= 1$. We ignore the possible relationship between the enrichment and the number of bins with specific mappability and GC content, since there is no reason to suspect this. Figure 3.10 shows the observed variances for the 49 chosen sets of bins as well as the variances after discarding the top 5% of read counts (to guard against the outlier effects) plotted vs. the mean. Variances obtained from the MOSAiCS-style model and our Poisson regression model (for this latter, mean=variance) are also plotted. We can see that outliers do have a large effect on the variance and that the negative binomial approach tends to over-estimate the variance of read counts after outliers are removed, while our approach tends to under-estimate it for large values of the mean.

Ultimately, the choice of the negative binomial or Poisson distribution does not matter too much, as the only purpose to which we apply this background modeling is to identify an inital set of candidate regsion. In practice, we use our Poisson model with Bonferroni-adjusted p-value cutoff of 0.1 to pick the candidate region set. We find that the resulting set is often larger than the final sets of binding sites obtained from running various peak-finders, indicating that we do not suffer from being overly conservative in this step of the procedure. We use the obtained estimate of the background rate in the model-fitting procedure described above.

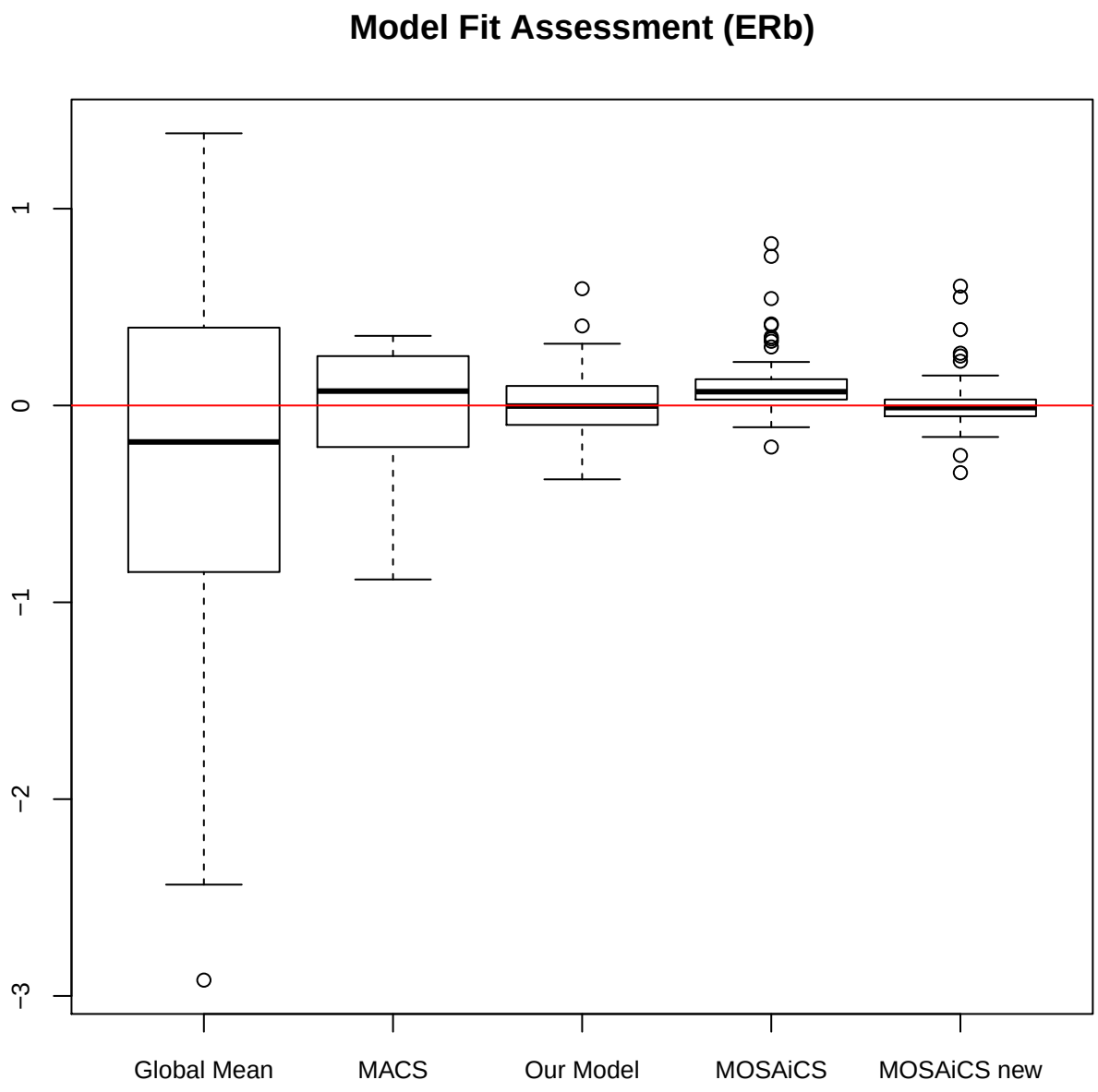

Figure 3.9: The adjusted rates for ERβ data with various background-modeling approaches ($\log_2$ scale). Global mean refers to global Poisson model. See text for description of MOSAiCS new. Red line corresponds to the adjusted rate of 1 (0 on $\log_2$ scale), the desired behavior. Global mean adjustment results in a large number of over-adjusted regions due to a small number of hyper-enriched regions.

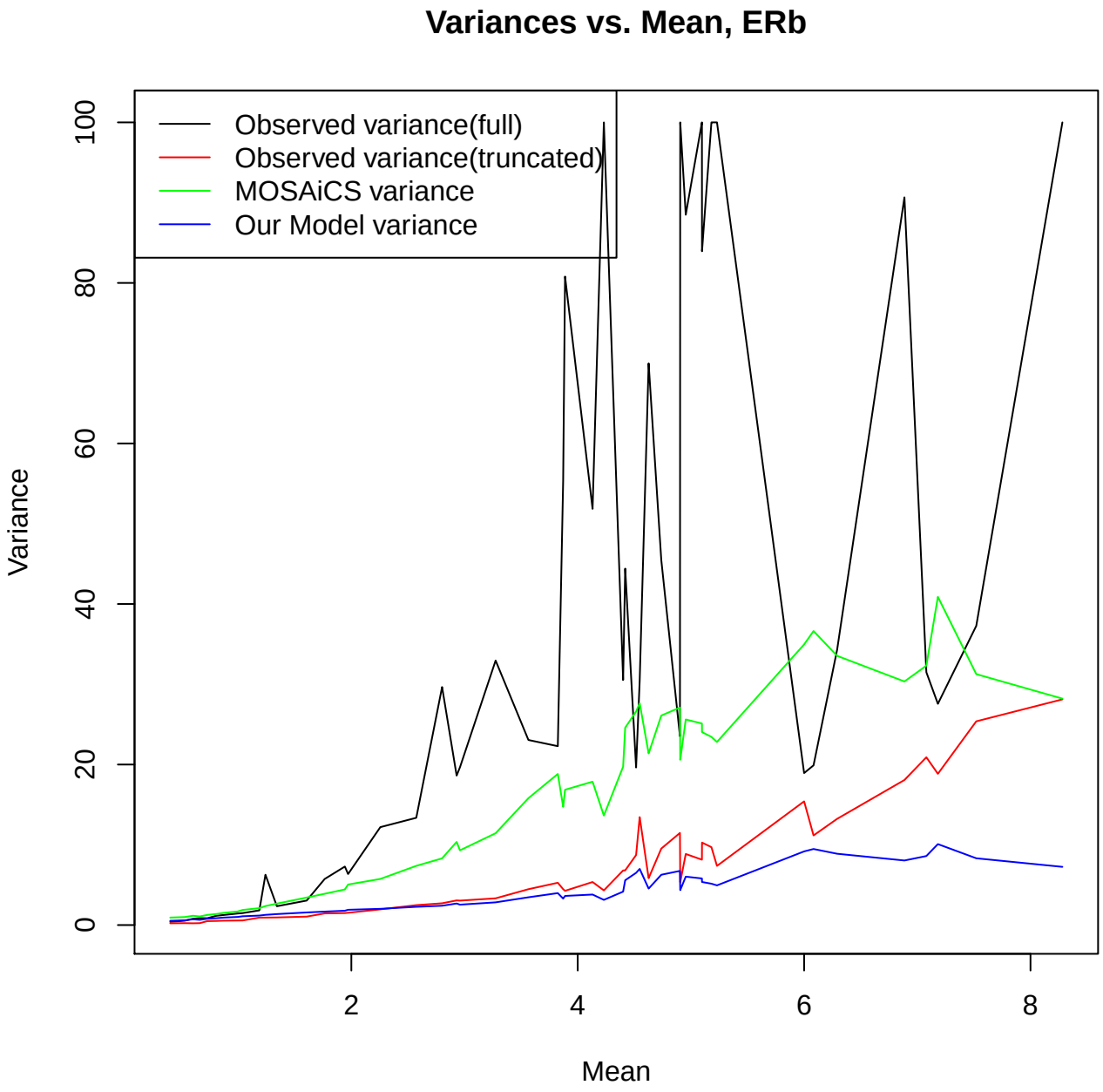

Figure 3.10: Poisson vs. Negative Binomial models. 49 sets of 1kb GC content are obtained that occur 10+ times in human genome. Variances of read counts in each set are calculated before (black) and after (red) discarding top 5% of read counts from each set (to get rid of outliers) and are plotted vs. the means of read counts in each set. All values are truncated at 100. Variances for each set under MOSAiCS-style model (green) and our model (blue) vs. the mean are also plotted. Truncated variances are much lower than raw variances and MOSAiCS-style approach over-estimates non-truncated variance, while our approach under-estimates the variance for higher values of the mean.

3.6 Classification

Having fit the peak model to candidate regions, we need to separate the hyper-enriched artifacts from the true binding sites. In order to do this, we introduce a correlation-based measure of goodness-of-fit of the model to the region, (r^+, r^-). Letting x_R^+ denote the vector of observed positive-strand read counts for positions in some candidate region R, and $\widehat{\lambda}^+$ denote the vector of estimated rates at those positions (where we plug in maximum likelihood estimates $\widehat{a}, \widehat{b}, \widehat{c}, \widehat{e}$ into $\lambda^+(t)$ to obtain the estimated rate at t), we define $r^+ = \text{cor}(Sm(x^+), \widehat{\lambda}^+)$, where Sm is some smoothing function, used to transform observed read counts into a more smooth read density. We use the running mean with window size of 100bp as our choice of Sm. r^- is defined analogously. The reason for introducing these correlation measures is to reduce the number of false positives among the declared binding sites - the regions that do not have the characteristic peak shape but produce peak-like parameter estimates during the model-fitting.

The aim of the classification is to assign the status S to each candidate region, either 1 or 0, corresponding to the true binding site or artifactual enrichment, respectively. We rely in our classification on the fact that the values of parameter estimates and the introduced correlation measures differ between these two classes of candidate regions. Some examples of this are shown in Figures 3.11-3.14, where we plot the distributions of some potential classifying features for the sets of deduced true binding sites and artifacts in two of the examined data sets. We observe that the true binding sites have higher values of r^+ and r^- and that they also have characteristic values of $\widehat{c}$ and $\widehat{e}$, allowing us to use these features in an attempt to separate true binding regions from artifacts among the candidate regions. There are a few deduced true binding regions that have somewhat low values of correlation measures or show uncharacteristic shape parameter estimates. Upon inspection, these regions are revealed to be multi-peaks, an example of which can be seen in Figure 3.15 and should be captured by allowing more than 1 peak in the fitted shape model.

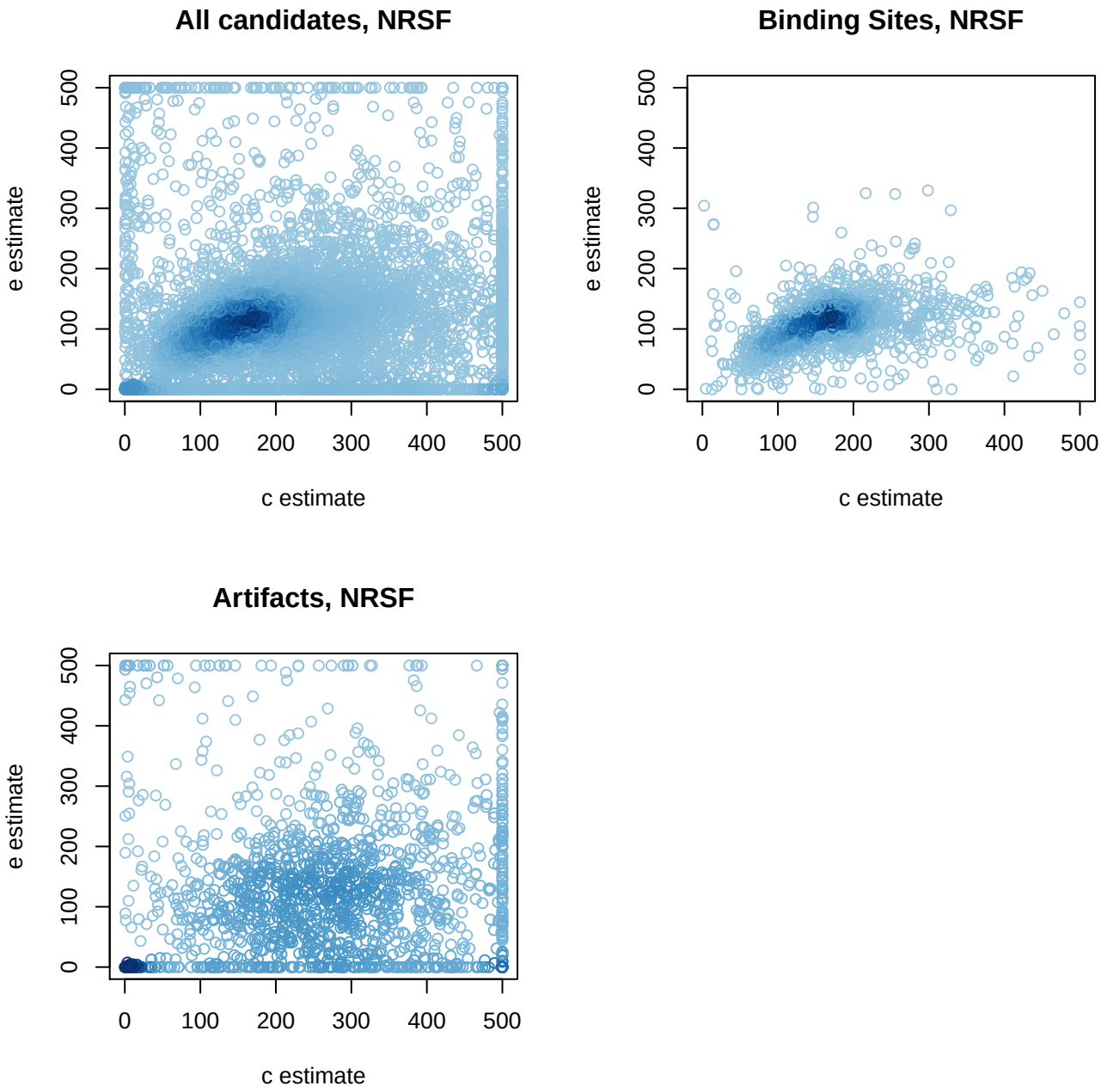

Figure 3.11: Joint distribution of $\hat{c}, \hat{e}$ for monoclonal Ab NRSF data set from [29]. All candidates refers to the set of candidate regions identified using the Poisson regression model. A subset of binding sites was deduced as the set candidate regions with at least 20 reads and a motif occurrence. A subset of artifacts was deduced as the set of candidate regions with fold enrichment < 1 over input DNA control (after adjusting for lane totals).

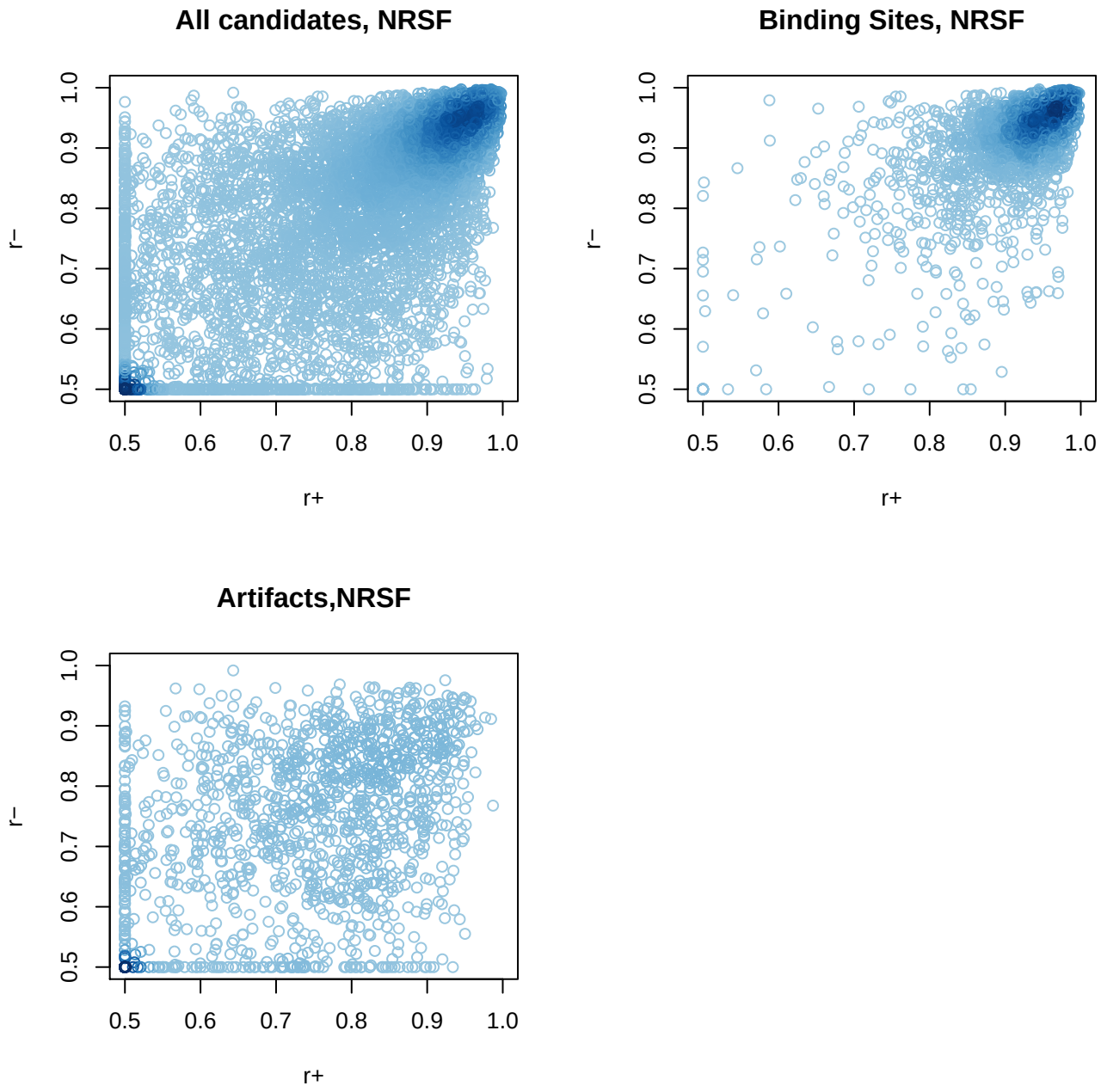

Figure 3.12: Joint distribution of r^+, r^- for monoclonal Ab NRSF data set from [29]. All candidates refers to the set of candidate regions identified using the Poisson regression model. A subset of binding sites was deduced as the set candidate regions with at least 20 reads and a motif occurrence. A subset of artifacts was deduced as the set of candidate regions with fold enrichment < 1 over input DNA control (after adjusting for lane totals).

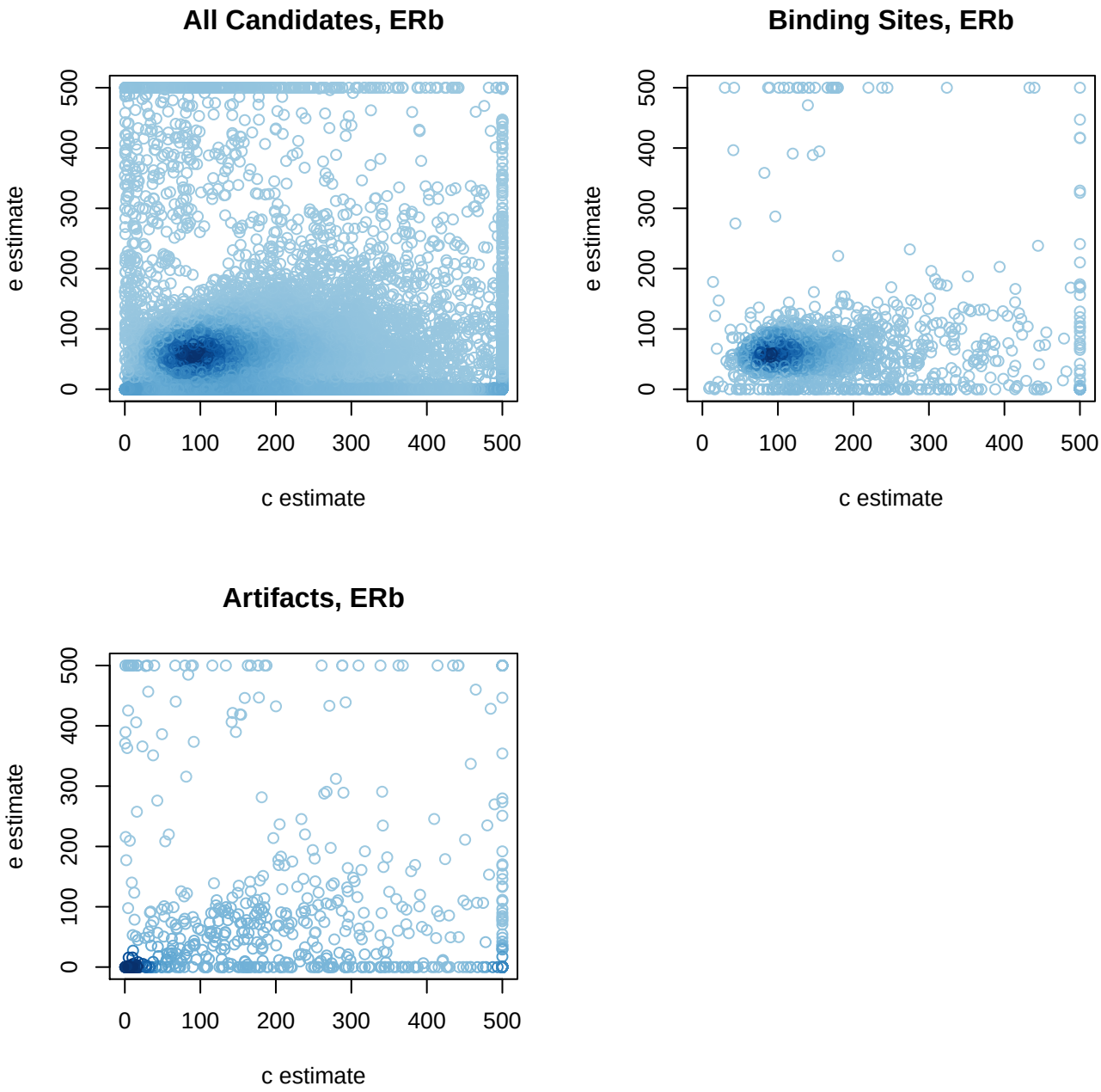

Figure 3.13: Joint distribution of $\hat{c}, \hat{e}$ for ERβ data set. All candidates refers to the set of candidate regions identified using the Poisson regression model. A subset of binding sites was deduced as the set candidate regions with at least 20 reads and fold enrichment > 5 over the IgG control (after adjusting for lane totals). A subset of artifacts was deduced as the set of candidate regions with at least 20 reads, fold enrichment < 2 over IgG control (after adjusting for lane totals) and detected enrichment in datasets on binding of other transcription factors (NRSF from [1] and STAT1 from [2]).

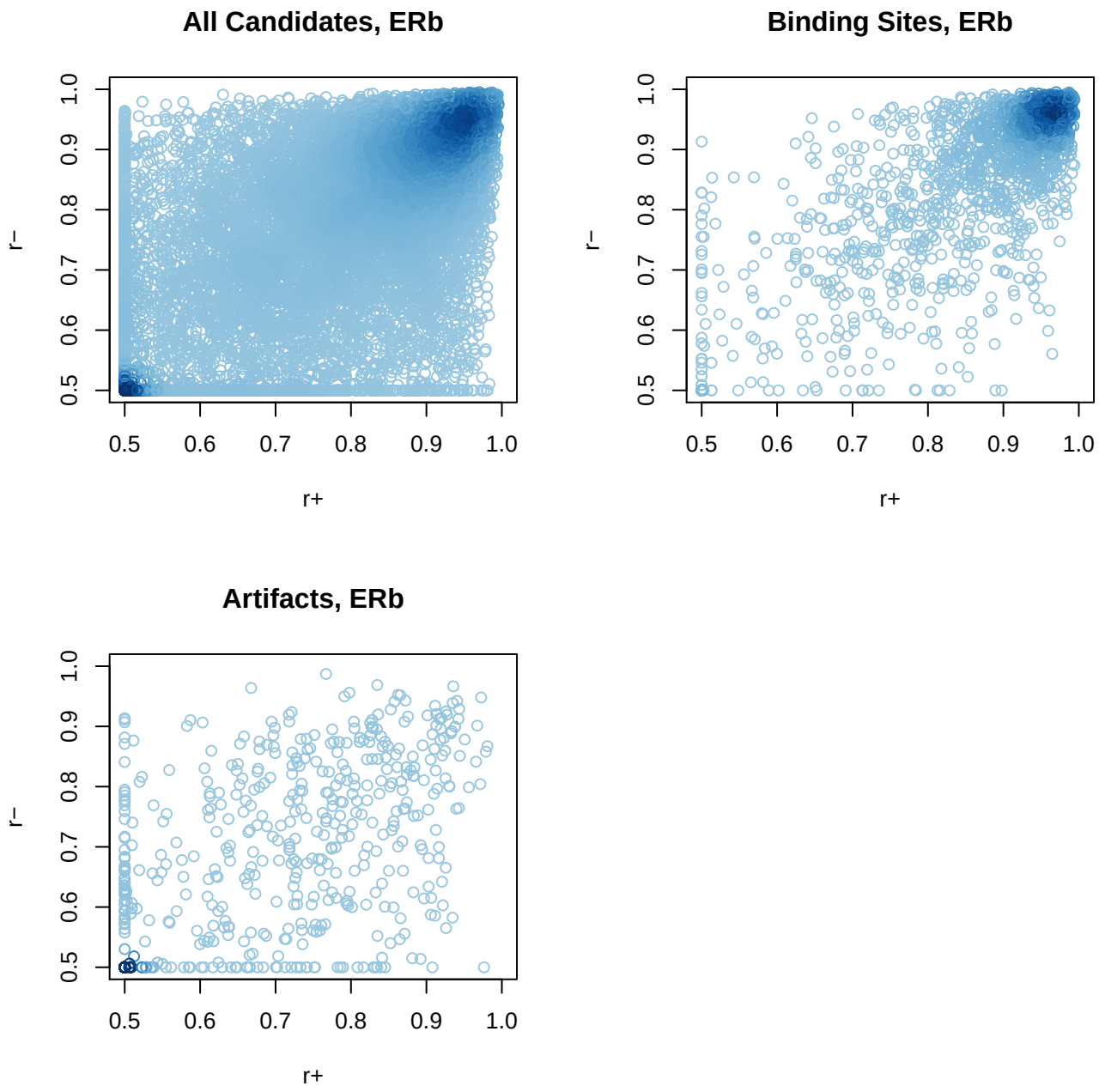

Figure 3.14: Joint distribution of r^+, r^- for ERβ data set. All candidates refers to the set of candidate regions identified using the Poisson regression model. A subset of binding sites was deduced as the set candidate regions with at least 20 reads and fold enrichment > 5 over the IgG control (after adjusting for lane totals). A subset of artifacts was deduced as the set of candidate regions with at least 20 reads, fold enrichment < 2 over IgG control (after adjusting for lane totals) and detected enrichment in datasets on binding of other transcription factors (NRSF from [1] and STAT1 from [2]).

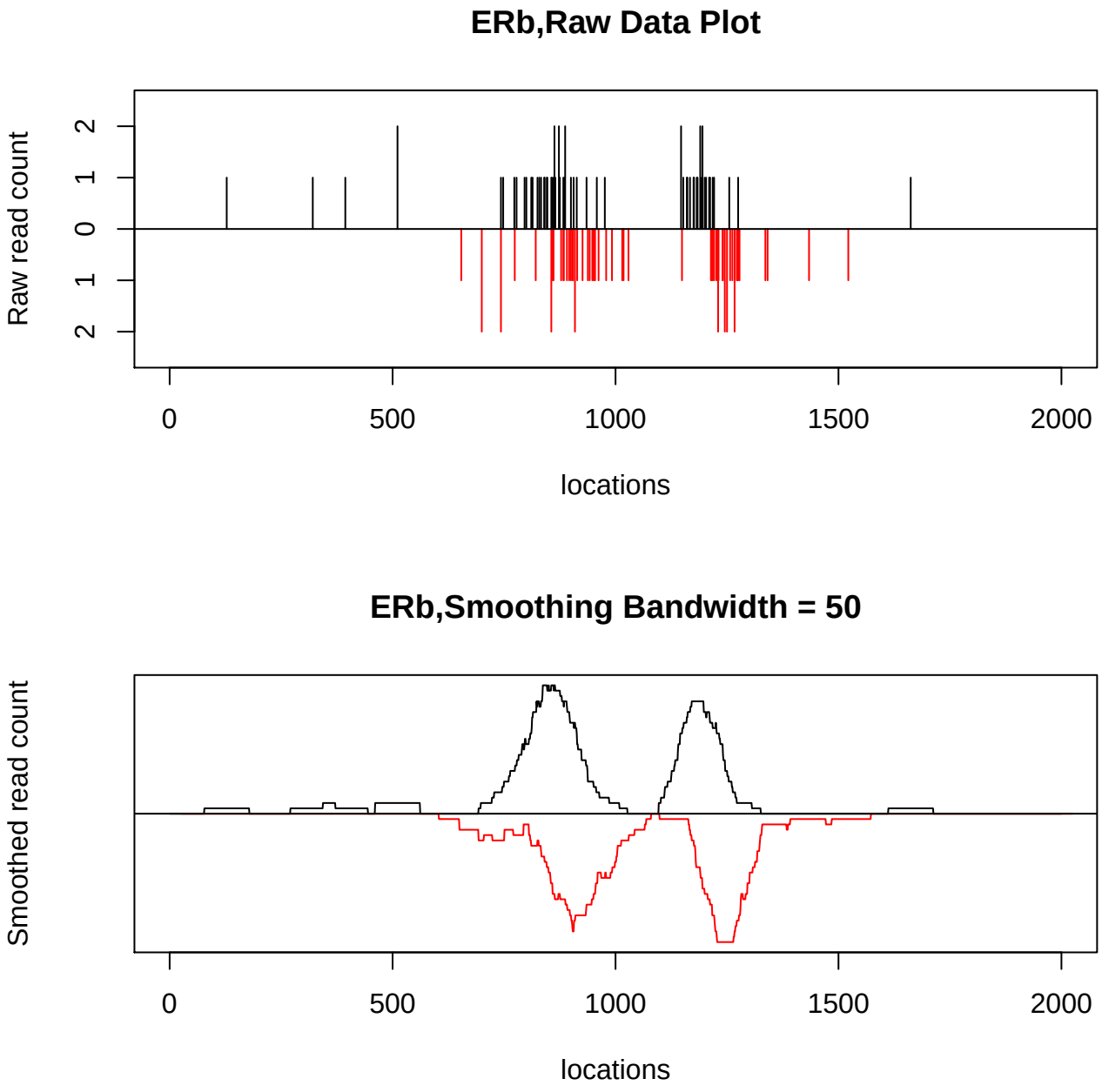

Figure 3.15: An example of a region with 2 peaks in close proximity to each other.

We use a set of classifying features F and for each candidate region R estimate $\mathrm{P}(F_R|S = 1)$ and $\mathrm{P}(F_R|S = 0)$. Using these likelihoods of observed features under the signal and background models, we assign status $S = 1$ to region R if $\widehat{\mathrm{P}}(F_R|S = 1) > \widehat{\mathrm{P}}((F_R|S = 0)$, i.e. if the likelihood ratio is > 1. Based on the observations above, we choose $F = \{\widehat{a}, \widehat{c}, \widehat{e}, r^+, r^-\}$ as the set of our classifying features. We introduce $\widehat{a}$ as one of the classifying features since it is a measure of the overall enrichment and for the true binding sites should be related to the estimated half-wdith of the peak $\widehat{c}$. In essence, we would like to avoid the situations where a small number of widely-spaced reads in a region gives rise to some reasonable values of width and strand separation.

The classification is done via an iterative algorithm that proceeds as follows. Let $R_1, R_2, \ldots, R_n$ denote the n candidate regions. At each iteration i of the algorithm these regions will be divided into an estimated set of true signal sites R_g^i and an estimated set of artifacts R_b^i (referred to as the sets of 'good' and 'bad' regions, respectively). An iteration-specific classification function S_i will assign the signal status to each candidate region and thus $R_j \in R_g^i \Leftrightarrow S_i(j) = 1$.

1. *Initial classification.* All candidate regions with $\min(r^+, r^-) > 0.7$ are assigned status $S = 1$, with the rest of the regions assigned status $S = 0$. This produces the inital estimates of the two region sets, R_g^0 and R_b^0.

2. *First Iterative Step.* Let $Z(i) = 1$ if i is odd and $Z(i) = 2$ otherwise. We divide the set of classification features F into two subsets, $F_1 = \{\widehat{a}, \widehat{c}, \widehat{e}\}$ and $F_2 = \{r^+, r^-\}$. For each candidate region R_j we let F_{1,R_j} denote the set $\{\widehat{a}_j, \widehat{c}_j, \widehat{e}_j\}$ and F_{2,R_j} denote the set $\{r_j^+, r_j^-\}$. During iteration i, we use candidate regions $R_j \in R_g^{i-1}$ to estimate the joint density $p_g(F_{Z(i)}) = p(F_{Z(i)}|S = 1)$ by the empirical density $\widehat{p}_{g,i}(F_{Z(i)}) = p_i(F_{Z(i)}|R \in R_g^{i-1})$. If $Z(i) = 1$ we let $p_i(F_1|R \in R_g^{i-1}) = d(\widehat{a}_j, \widehat{c}_j|R_j \in R_g^{i-1})d(\widehat{e}_j|R_j \in R_g^{i-1})$, where d is a non-parametric kernel-smoothed density estimate and the factorization of the joint density is explained below. If $Z(i) = 2$, we let $p_i(F_2|R \in R_g^{i-1}) = d(r_j^+, r_j^-|R_j \in R_g^{i-1})$. The joint density $p_b(F_{Z(i)}) = p(F_{Z(i)}|S = 0)$ is estimated analogously. For each candidate region R_j we calculate $p_i(F_{Z(i),R_j}|R \in R_g^{i-1})$ and $p_i(F_{Z(i),R_j}|R \in R_b^{i-1})$ and, if the former is greater than the latter, we let $S_i(j) = 1$ and we let $S_i(j) = 0$ otherwise. This produces new estimates of the sets of good and bad regions, R_g^i and R_b^i.

   This procedure is repeated for a maximum of T=20 iterations or until convergence is reached. In this step, the convergence means that for some i_C, $i > i_C \Rightarrow R_g^{i+2} = R_g^i$, i.e. the convergence is achieved for both classifications running in parallel: odd iterations based on F_1 and even iterations based on F_2. There is no implication that $R^{i+1} = R^i$ and in practice we do not see both parallel classifications converge to the same estimate.

3. *Second Iterative Step.* Let $R_g^{1,T}, R_g^{2,T}$ denote the final sets of good regions from the first iterative step based on F_1 and F_2, respectively, and analagously define $R_b^{1,T}, R_b^{2,T}$. For clarity purposes, we will label the iterations during this step starting with 1 again.

Define a new initial estimate of the set of good regions $R_g^0 = R_g^{1,T} \cap R_g^{2,T}$. During iteration i of this second iterative step, we use candidate regions $R_j \in R_g^{i-1}$ to estimate the joint density $p_g(F) = p(F|S = 1)$ by the observed empirical density $\widehat{p}_{g,i}(F) = p_i(F|R \in R_g^{i-1})$, where we let $p_i(F|R \in R_g^{i-1}) = d(\widehat{a}_j, \widehat{c}_j|R_j \in R_g^{i-1})d(\widehat{e}_j|R_j \in R_g^{i-1})d(r_j^+, r_j^-|R_j \in R^{i-1})$, using the same notation as in the first iterative step. The joint density $p_b(F) = p(F|S = 0)$ is estimated analogously. The re-classification procedure is analogous to that from the first iterative step, except that now instead of using different feature sets F_1 and F_2 in consecutive iterations, we use the same full set of features F in every iteration. This process is repeated for a maximum of T=20 iterations or until convergence and the final classification sets R_g^T, R_b^T are used as our estimates of the sets of signal and artifact regions, respectively.

4. *Allowing multiple peaks per region* We can also allow more than 1 peak to occur in the region. An example of a multi-peak region in ERβ data is shown in Figure 3.15. A single-peak model would fit this region poorly; in particular, correlation values will be low and the region will be classifed as an artifact. To avoid this, we fit a multi-peak model (with $k = 2$ peaks, but k can be increased) to all regions that have been classified as artifacts based on a single-peak fit. For each candidate multi-peak region R_j we require that for each constituent peak $i = 1, \ldots, k$ we have $\widehat{p}(F_{R_j,i}|S = 1) > \widehat{p}(F_{R_j,i}|S = 0)$, where $F_{R_j,i} = \{\widehat{a}_i, \widehat{c}_i, \widehat{e}_i, r_j^+, r_j^-\}$ and where the estimates of $p(F|S = 1)$ and $p(F|S = 0)$ are obtained as in the second iterative step above using the final single-peak classification sets R_g^T, R_b^T. Essentially, we require that in order for us to declare the region to contain k peaks, every one of those peaks needs to be classified as 'good' based on how its classification features compare to those for the final set of single-peak regions. The correlation measures r^+ and r^- are common to all peaks in the region and measure how the observed read rates agree with the convolution of the background rate and superimposition of individual binding profiles.

A few words about the choice of the joint density factorization are in order. For computational reasons, we choose not to calculate the non-parameteric estimate of the entire 5-dimensional feature density. We choose to factor the joint density $p(\widehat{a}, \widehat{c}, \widehat{e}, r+, r-) = p(\widehat{a}, \widehat{c})p(\widehat{e})p(r^+, r^-)$ to capture the dependence relationship between the two correlation measures (intuitively, true signal regions will have high values of both stranded correlations while the artifacts will not) and between measures of signal strength a and peak width c (the greater the signal strength, the wider the peak is likely to be).

3.7 Summary of the proposed method

Our proposed method consists of the following steps:

1. Candidate region identification

(a) Bin genome into 1kb bins, record the number of bins in each bin.

(b) Assume that for each bin j, the read count $X_j \sim \text{Poisson}(\lambda_{BG,j})$, with $\lambda_{BG,j} = \lambda_0 F_{m,j}(F_{GC,j})^k$, with $F_{GC,j}$ the fraction of G's and C's among the nucleotides in the bin and $F_{m,j}$ the fraction of mappable bases.

(c) Run Poisson regression of bin read counts on bin GC content, using bin mappability content as an offset.

(d) For each bin j, use the estimated read rate $\widehat{\lambda}_{BG,j}$ to calculate the minimum value c_j of c, where $P(X_j > c) < p$, for some specified p-value p. In practice we let $p = 0.1/n$, where n is the number of genomic 1kb bins being tested, i.e. p is the Bonferroni-adjusted p-value of 0.1.

(e) Keep the bins with observed values of read counts $x_j \geq c_j$. These are the bins that show enrichment in excess of their estimated background rate.

(f) Repeat the procedure for a new partition of the genome into 1kb bins, where the bins are offset from the previous partition by 500bp. This is meant to account for any events that happen at the breakpoints of the original partition.

(g) Merge the two sets of enriched bins to produce a set of non-overlapping candidate regions. For candidate regions that have resulted from the overlap of two or more enriched bins, use the weighted mean of bin-specific estimated read rates to obtain an estimated rate $\widehat{\lambda}_{BG,j}$ for the region.

2. Model Fitting

(a) Assume that the strand-specific read density in the candidate regions corresponding to the true binding events is a Poisson point process, with strand-specific rates given by $\lambda^+(t) = \lambda^+_{BG}(t) + \frac{a}{c} S\left(\frac{t-b}{c}\right)$ and $\lambda^-(t) = \lambda^-_{BG}(t) + \frac{a}{c} S\left(\frac{t-b-e}{c}\right)$, where $S(x) = \max(0, 1 - |x|)$ is a triangular shape function.

(b) Fit the model to each candidate region independently, plugging in the read rates $\widehat{\lambda}_{BG,j}/2$ estimated from Poisson regression as the strand-specific background rates λ^s_{BG}.

(c) For each candidate region, calculate the strand-specific correlation measures r^+ and r^-, given by $r^s = \text{cor}(Sm(x^s), \widehat{\lambda}^s)$, where x^s and $\widehat{\lambda}^s$ are the strand-specific vectors of read counts and estimated rates for locations in the candidate regions, and Sm is a smoothing function. We use a running mean in 100bp windows as our choice of Sm.

3. First Step of Classification

(a) Classify each candidate region as good if $\min(r^+, r^-) > 0.7$ and bad otherwise.

(b) Iteratively re-classify regions into good and bad, alternating between classifications based on feature sets $\{\widehat{a}, \widehat{c}, \widehat{e}\}$ and $\{r^+, r^-\}$. At each iteration i, use the classification of candidate regions into good and bad from the previous iteration $i-1$ to obtain non-parametric density estimates $\widehat{p_g}(F_i)$ and $\widehat{p_b}F_i$ of the feature set for iteration i for the good and bad candidate regions, respectively.

(c) Use the likelihood ratio $\widehat{p_g}(F_{i,R_j})/\widehat{p_b}(F_{i,R_j})$ to re-classify the candidate regions R_j into good and bad. We use the likelihood ratio cutoff of 1 in our classification.

(d) Iterate until convergence or until the maximum of 20 iterations is reached. Odd and even iterations will converge to two different classifications.

4. Second Step of Classification

(a) Let the regions classified as good under both the final odd and the final even iterations of the first classification step be the initial set of good regions for this step. The rest of the regions is declared bad.

(b) Iteratively re-classify regions into good and bad based on the full feature set $F = \{\widehat{a}, \widehat{c}, \widehat{e}, r^+, r^-\}$. At each iteration i, use the classification of candidate regions into good and bad from the previous iteration $i-1$ to obtain non-parametric density estimates $\widehat{p_g}(F)$ and $\widehat{p_b}(F)$.

(c) Use the likelihood ratio $\widehat{p_g}(F_{R_j})/\widehat{p_b}(F_{R_j})$ to re-classify the candidate regions R_j into good and bad. We use the likelihood ratio cutoff of 1 in our classification.

(d) Iterate until convergence or the until the maximum of 20 iterations is reached. The final classification is used to identify the set of true signal sites among the candidate regions

5. Allowing for Multi-Peaks

(a) For $K = 2, \ldots, N$ fit a model with K peaks to all candidate regions classified as bad under the final classification from the model with $1, \ldots, K-1$ peaks allowed. N stands for the maximum number of allowed peaks in a candidate region, and we use $N = 2$ in our classification. For each region, record the peak-specific shape parameters $\widehat{a}_k, \widehat{c}_k, \widehat{e}_k$ for $k = 1, \ldots, K$ and the common correlation measures of goodness of fit r^+, r^-.

(b) Using the final classification from the 1-peak model, obtain non-parametric density estiamtes $\widehat{p_g}(F)$ and $\widehat{p_b}(F)$ as before.

(c) For each region, calculate K peak-specific likelihood ratios $\widehat{p_g}(F_{P_{j,k}})/\widehat{p_b}(F_{P_{j,k}})$, where $P_{j,k}$ is the k-th peak in region R_j.

(d) Require that the likelihood ratio exceed some cutoff (we use cutoff=1) for each peak in the region in order for the region to be declared a true signal site with K binding events.

We end up with a set of candidate regions classified as enriched due to signal and containing between 1 and N instances of binding events.

3.8 Validation

We validate the performance of the method described above on two datasets. One is the ERβ data set that contains a number of artifacts and the other is the monoclonal antibody NRSF data from [29]. This last data set was chosen as being representative of the high quality published data sets.

3.8.1 ERβ

We compare the performance of our one-sample approach to that of MACS, SISSRS, QuEST, ERANGE, CisGenome and PICS. When both one-sample and two-sample approaches are supported by a peak-finder, we try both and report the results. In order to properly evaluate a peak-finder's performance the sets of true positives and true negative regions are needed. As discussed in section 2.4 , for all of the examined ChIP-Seq data sets, including this one, no meaningful set of true negatives exists. Meaningful in this context refers to a set of known read-rich artifacts as opposed to true negative regions showing no enrichment. Moreover, in the case of ERβ, identifying a good set of true positives is also challenging as there are a variety of binding motifs, with canonical ERE motif present at less than a half of all binding sites.

We address these issues by deducing a set of true binding events and artifacts from the data as follows. To be declared a true positive, the candidate region must

1. Show significant enrichment in ERβ sample.
2. Show fold enrichment > 2 over IgG (non-specific antibody) control.
3. Be at most 1500 bps in length (to exclude long regions of non-specific enrichment)
4. Not show significant enrichment in any ChIP-seq samples obtained from two other examined data sets, on binding of transcription factors NRSF [1] and STAT1 [2]. This is done to exclude the non-specifically enriched regions that might show up in unrelated experiments.

To be declared a true negative, the candidate region must

1. Show significant enrichment in ERβ sample.
2. Show fold enrichment < 2 over IgG control

3. Show significant enrichment in either one of the two data sets mentioned above or the IgG control sample.

This approach results in a set of 1,689 deduced true positive (TP) and 610 deduced true negative (TN) regions. Table 3.3 summarizes the results of running various peak-finders on the ERβ data set. In general, two-sample peak-finders perform better at filtering out our defined true negative regions. This is at least partially due to the requirement that fold enrichment for these regions is < 2. There is great variability in the total number of peaks called by peak-finders, reflecting their internal stringency criteria. Interestingly enough, our approach performs better than PICS, which is the only other method that utilizes peak-shape explicitly (by assuming that the shape is that of t distribution with 4 degrees of freedom).

Peak-Finder	Number of Peaks	% of TPs retained	% of TNs retained
Our approach	9,027	85	19
SISSRS (1-sample)	25,970	99 (85)	92 (39)
ERANGE (1-sample)	6,782	76 (76)	23 (23)
MACS (1-sample)	31,030	99 (87)	88 (33)
CisGenome (1-sample)	9,637	87 (86)	43 (39)
SISSRS (2-sample)	14,059	90 (81)	11 (5)
ERANGE (2-sample)	6,176	75 (75)	4 (4)
MACS (2-sample)	25,788	99 (88)	24 (1)
CisGenome (2-sample)	4,452	57 (57)	3 (3)
QuEST (2-sample)	12,063	59 (58)	26 (23)
PICS (2-sample)	12,367	80 (77)	65 (27)

Table 3.3: The results of running various peak-finders on ERβ data. The numbers in parentheses are percentages for the top 9027 peaks, the number of binding sites identified by our approach.

To assess how different 1-sample peak-finders perform we build receiver operating characteristic (ROC) curves based on the number of our defined true positives and true negatives among the top k peaks according to different methods, where we vary k to build the actual curves. The top peaks are determined according to the peak-finders' own measures of significance and in our approach by the likelihood ratio $\widehat{p_g}(F_R)/\widehat{p_b}(F_R)$. The results are shown in Figure 3.16 and illustrate that taking the shape of read density at the candidate binding site into account improves the ability to differentiate between specific and non-specific enrichment.

There are some regions that our approach misses among the defined TPs and some that it retains among the defined TNs. After manual inspection of the missed TPs, we believe most of them to represent either experiment-specific artifacts, weak multipeak signals (where

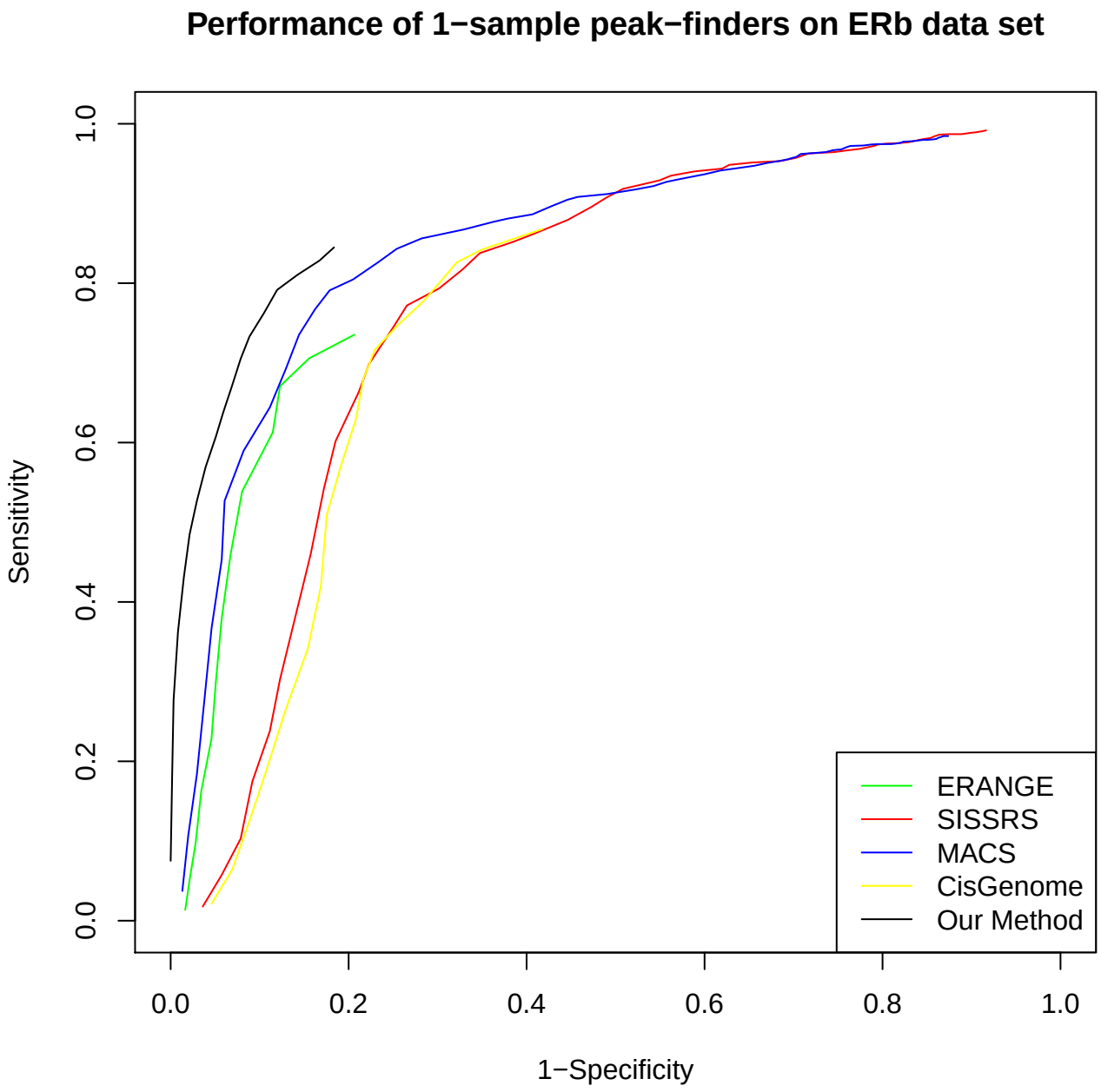

Figure 3.16: Performance assessment of various 1-sample peak-finders on ERβ data set.

neither 1- nor 2-peak model fits well) or non-punctate binding events. An example of such missed region is shown in Figure 3.17. The retained TNs tend to show the strand-separated peak shape that we expect from the sites of true binding events and might represent true binding events that happen in regions of open chromatin (which would explain the enrichment in other ChIP-Seq experiments) or non-specific enrichment that shows the same read density profile as a true binding event. An example of such region is shown in Figure 3.18.

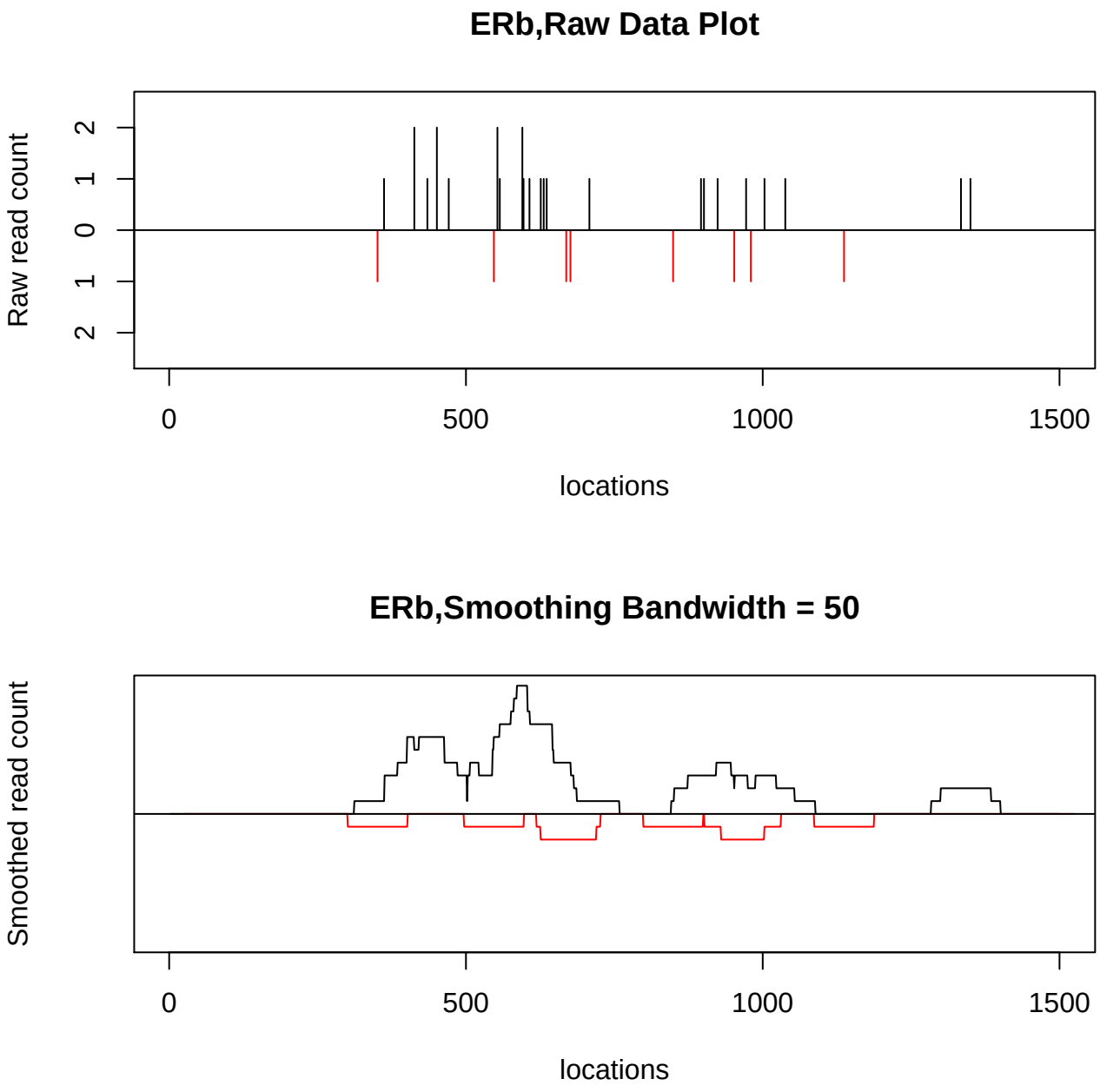

Figure 3.17: An example of a deduced true positive region missed by our one-sample approach in ER β data set.

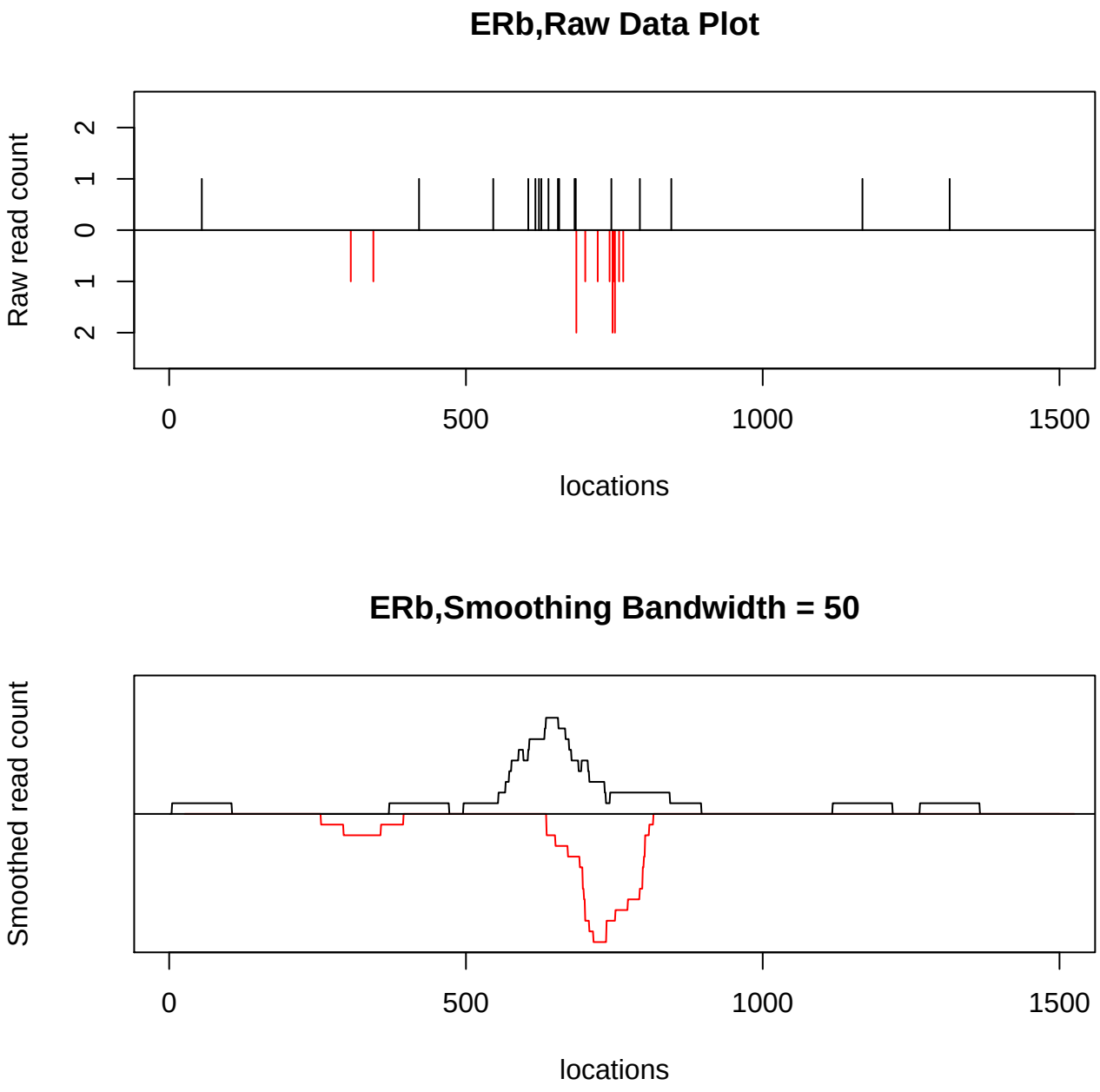

Figure 3.18: An example of a deduced true negative region retained by our one-sample approach in ERβ data set.

3.8.2 NRSF

In order to assess the performance of peak-finders on this data set, we declare a set of true positives to be genomic regions that

1. Show significant enrichment in NRSF treatment sample.
2. Show fold enrichment > 2 over input DNA control.
3. Contain the canonical binding motif.

To come up with a good set of read-rich true negative regions, we scan the genome for regions that

1. Show significant enrichment in NRSF treatment sample.
2. Show fold enrichment < 2 over input DNA control.
3. Do not contain the canonical binding motif.
4. Also show treatment over input DNA control fold enrichment < 2 for a different NRSF binding data set from [1]

This approach results in the set of 1,906 deduced true positive and 691 deduced true negative regions. Table 3.4 summarizes the results of running various peak-finders on this NRSF data set. In general, two-sample peak-finders perform better at filtering out our defined true negative regions, but not by much. This probably reflects very strong absolute enrichment in those regions. Again, we seem to perform better than PICS, the other shape-based approach to signal-noise deconvolution. It should be pointed out that the relatively small fractions of true positive regions among the top 2973 peaks for both 2-sample MACS and SISSRS peak-finders should not reflect poorly on these tools, since this fraction is subject to a tool-specific system of peak ranking.

We build an ROC curve, similar to that presented in ERβ data set analysis. The results are shown in Figure 3.19 and once again we see that our performance is superior to other one-sample methods at high values of specificity.

Just as was the case with ERβ data, there are some true positive regions that we miss and some true negative regions that we retain. Manual examination reveals that the vast majority of missed TPs appear to have shapes that are not fit well by the model (e.g. heavy tails). An example of a missed true positive region is shown in Figure 3.20. Similarly, manual examination of the retained TNs reveals that a lot of them seem to be 'universal peaks' - regions that are enriched and show the characteristic strand-separated peak shape in multiple ChIP-Seq experiments, including at least one instance of input DNA control sample. An example of such retained true negative regions is shown in Figure 3.21, where the smoothed read density is plotted for this region in several ChIP-Seq data sets in different cell lines.

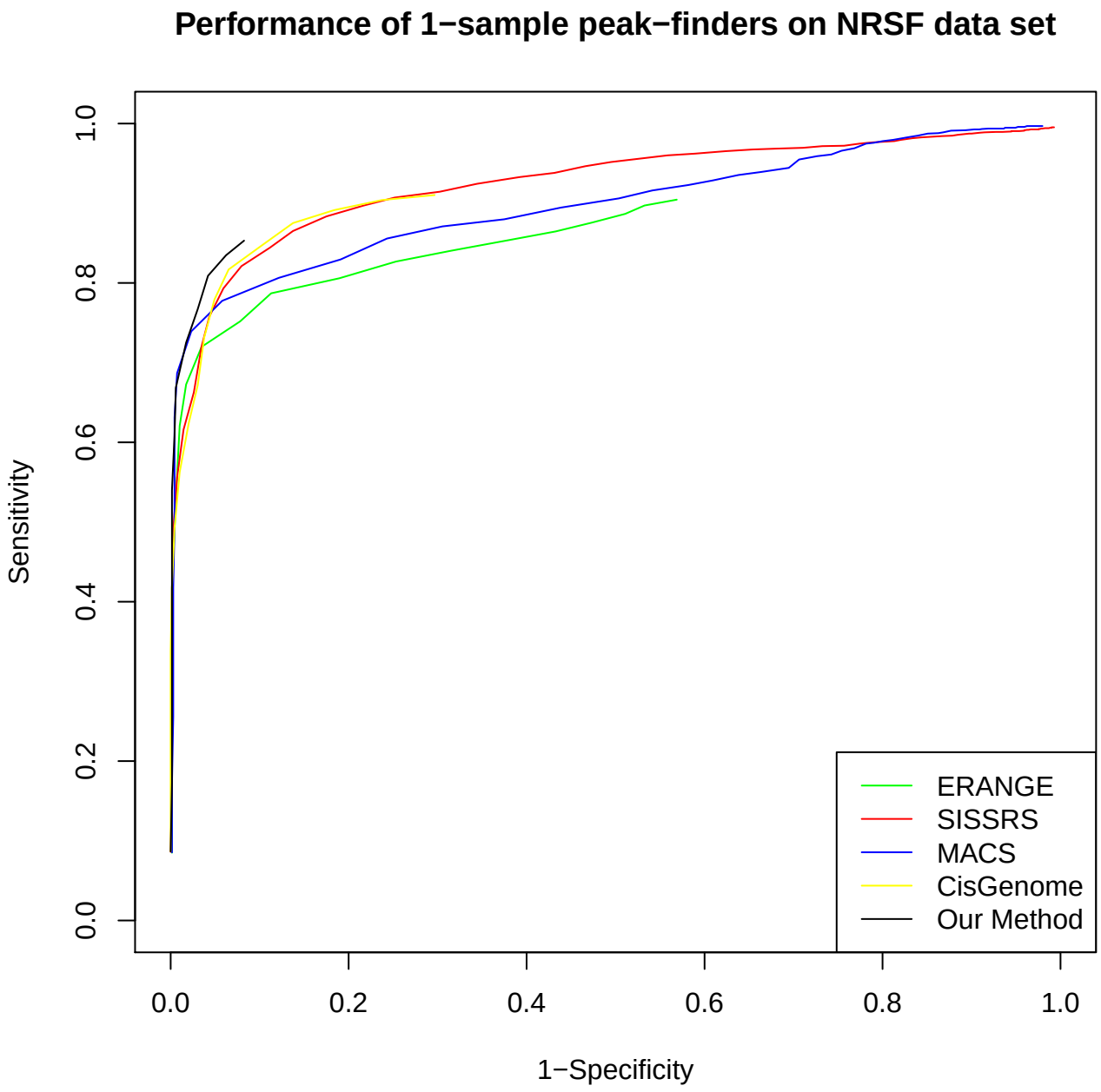

Figure 3.19: Performance assessment of various 1-sample peak-finders on monoclonal antibody NRSF data set.

Peak-Finder	Number of Peaks	% of TPs retained	% of TNs retained
Our approach	2,973	87	10
SISSRS (1-sample)	17,465	99 (86)	99 (13)
ERANGE (1-sample)	4,365	91 (84)	59 (30)
MACS (1-sample)	20,471	99 (87)	98 (29)
CisGenome (1-sample)	3,441	91 (89)	31 (18)
SISSRS (2-sample)	7,354	99 (30)	11 (11)
ERANGE (2-sample)	4,365	91 (84)	59 (30)
MACS (2-sample)	8,076	99.0 (30)	9 (8)
CisGenome (2-sample)	2,616	87 (87)	5 (5)
QuEST (2-sample)	4,876	90 (87)	14 (11)
PICS (2-sample)	3,688	76 (60)	10 (8)

Table 3.4: The results of running various peak-finders on monoclonal antibody NRSF data. The numbers in parentheses are percentages for the top 2973 peaks, the number of binding sites identified by our approach. ERANGE has produced the same set of peaks under both the 1-sample and 2-sample models.

The origin of such behavior is unclear. The 'universal peaks' might represent some true TF-binding hotspots or, more likely, locations of non-specific point-like enrichment, perhaps very narrow open chromatin regions. Such artifacts cannot be filtered out using the shape-based approach that we introduced here and control samples are required to fully account for such events.

We have also compared the spatial resolution achieved by our method, compared to the other peak-finders, on the set of motif-containing peaks. For each peak-finder, we obtain the set of motif-containing regions that are identified as binding sites by both our method and that peak-finder. To avoid any ambiguity, we further reduce this set of regions to only those that are identified as single peaks by both approaches. On this common set of motif-containing peaks, we compare the motif occurrence rate within some distance d of the method-specific summits (point estimates of the binding sites). For our method, we define the point estimate of the binding event as $\widehat{b} + (\widehat{e}/2)$ (after shifting the negative-strand reads downstream by the read length). Figure 3.22 shows the comparison of spatial resolution of our method compared to some other 1-sample peak-finders, while Figure 3.23 shows the comparison to some 2-sample peak-finders. We outperfrom CisGenome, ERANGE, and QuEST, while MACS seems to provide an even better resolution. SISSRS outperforms us for smaller values of d, although we do better for larger ones. Finally, we perform essentially identical to PICS, which indicates that the two shape-based approaches come up with the same point estimates of the binding events. Examination of several peaks where MACS exhibits higher resolution than our method shows that this is due largely to some asymmetry

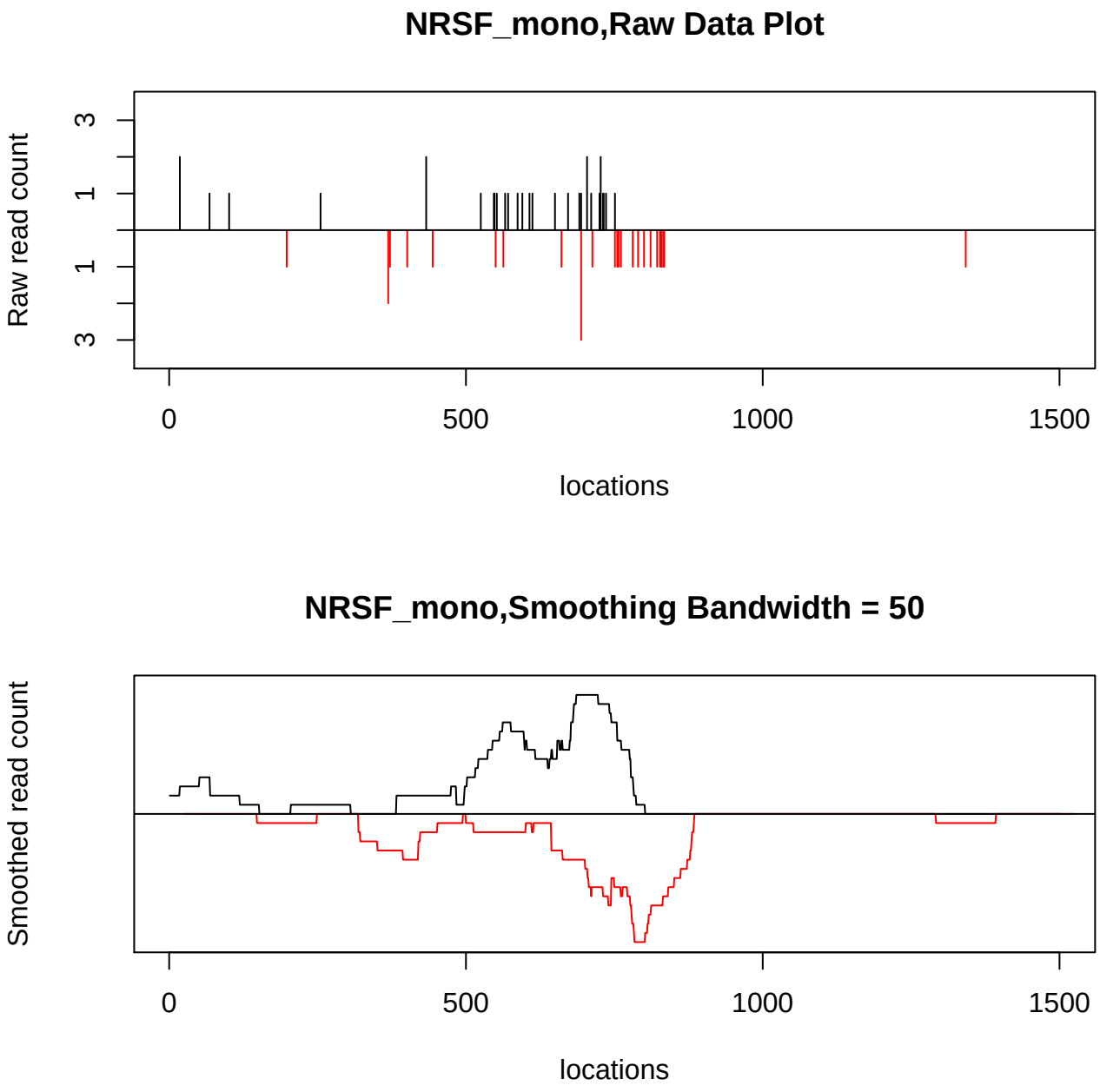

Figure 3.20: An example of a deduced true positive region missed by our one-sample approach in NRSF data set.

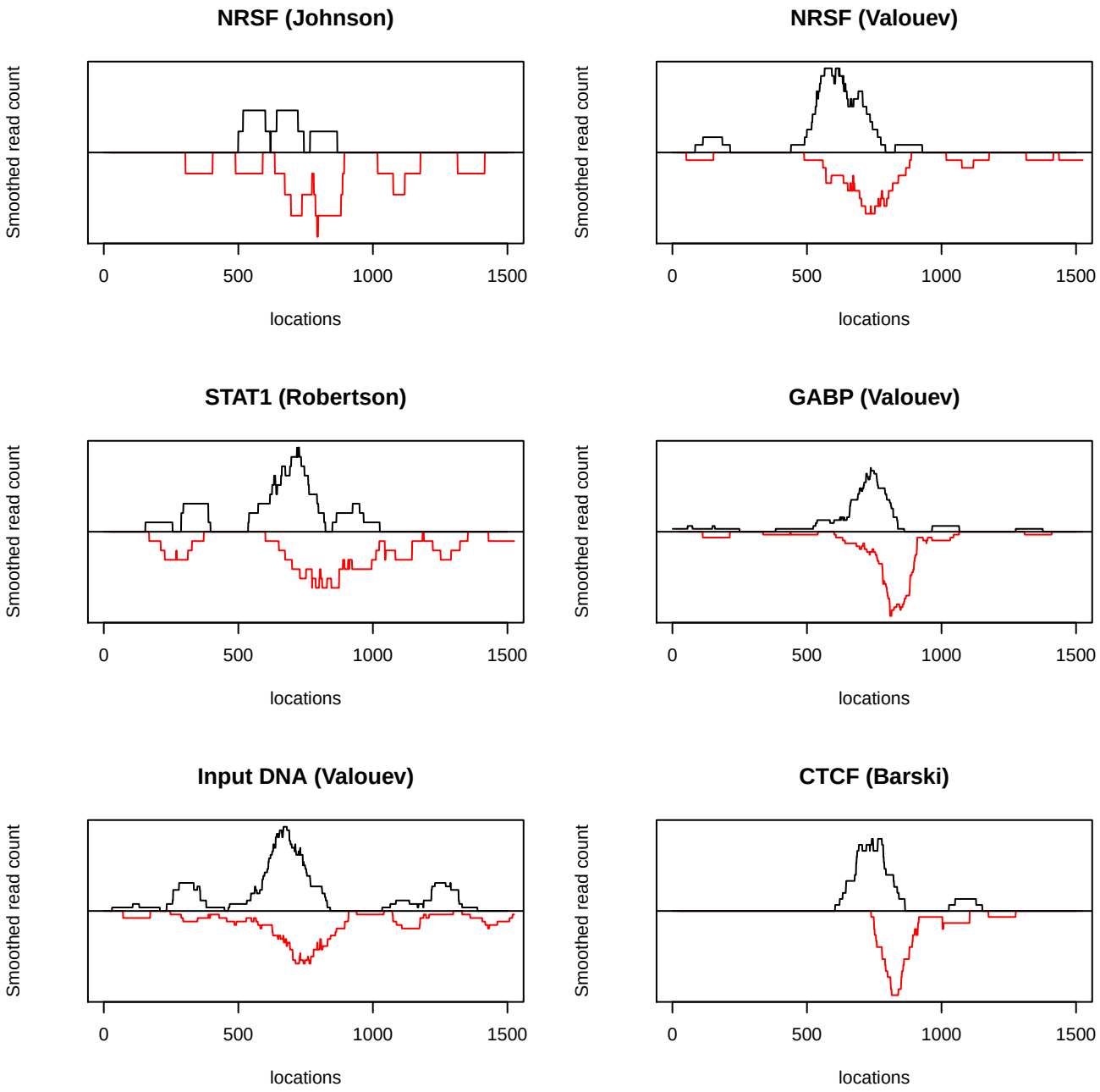

Figure 3.21: An example of a deduced true negative region retained by our one-sample approach. The smoothed read density in the region is shown in several published ChIP-Seq data sets (the data set considered here is NRSF (Valouev)).

in one or both strand-specific peaks. In general, out of all the motif-containing regions identified by both MACS and our method, 87% result in the method-specific summits within 20 bps of each other and 96% result in the summits within 30 bps of each other.

3.9 Summary of Comparisons

The shape-based method presented here is shown to be superior to other one-sample approaches in identifying and discarding hyper-enriched artifacts. A truly meaningful comparison to the two-sample approaches awaits a validation data set with true negative regions that are not defined based on the fold enrichment over the control sample.

Our spatial resolution is superior to many other peak-finders, although we are outperformed by MACS. The lower spatial resolution of our method is a result of occasional asymmetry in the strand-specific profile that our approach is not robust against.

A major drawback of our method is its inability to correctly classify the regions corresponding to true binding events that show read density patterns that are different from those of the majority of the binding regions or that deviate from the modeled triangular read count rate behavior. All of these regions are picked up by the other one-sample peak-finders due to high raw enrichment. Thus, if the data set contains very strong signal and few, if any, strong artifacts, other one-sample peak-finders may produce better overall results. On the other hand, if the signal profile is weak and there are many non-specifically enriched regions, our approach is likely to prove superior. The notable exception to this claim are the cases of data sets with a promiscuous antibody, where most artifactual enrichment will be indstinguishable from true signal based on the read density profile and all one-sample peak-finders will perform poorly.

3.10 Modifications to the Procedure

We have tried several different versions of shape-based region classification. One area for improvement that we have explored is the set of classifing features F that we use. Omitting the goodness-of-fit measures r^+ and r^- from F has resulted in a much higher rate of true negative region retention, as expected. Similarly, omitting $\widehat{a}$ led to both fewer retained true positive regions and more retained true negative regions. We have also tried out some modified features, such as $\widehat{a}/\widehat{c}$ (peak height), but the improvement over current procedure was inconclusive.

The classification procedure presented above is based on the likelihood ratio $\widehat{p}(F|S = 1)/\widehat{p}(F|S = 0)$, with candidate regions classified as true binding events if this ratio is > 1. We have explored an alternative approach, in which regions are classifed as signal if the posterior probability $\widehat{p}(S = 1|F) > 0.5$. With this alternative approach we employ the same iterative algorithm as before but at each iteration i we also estimate $p(S = 1)$ as

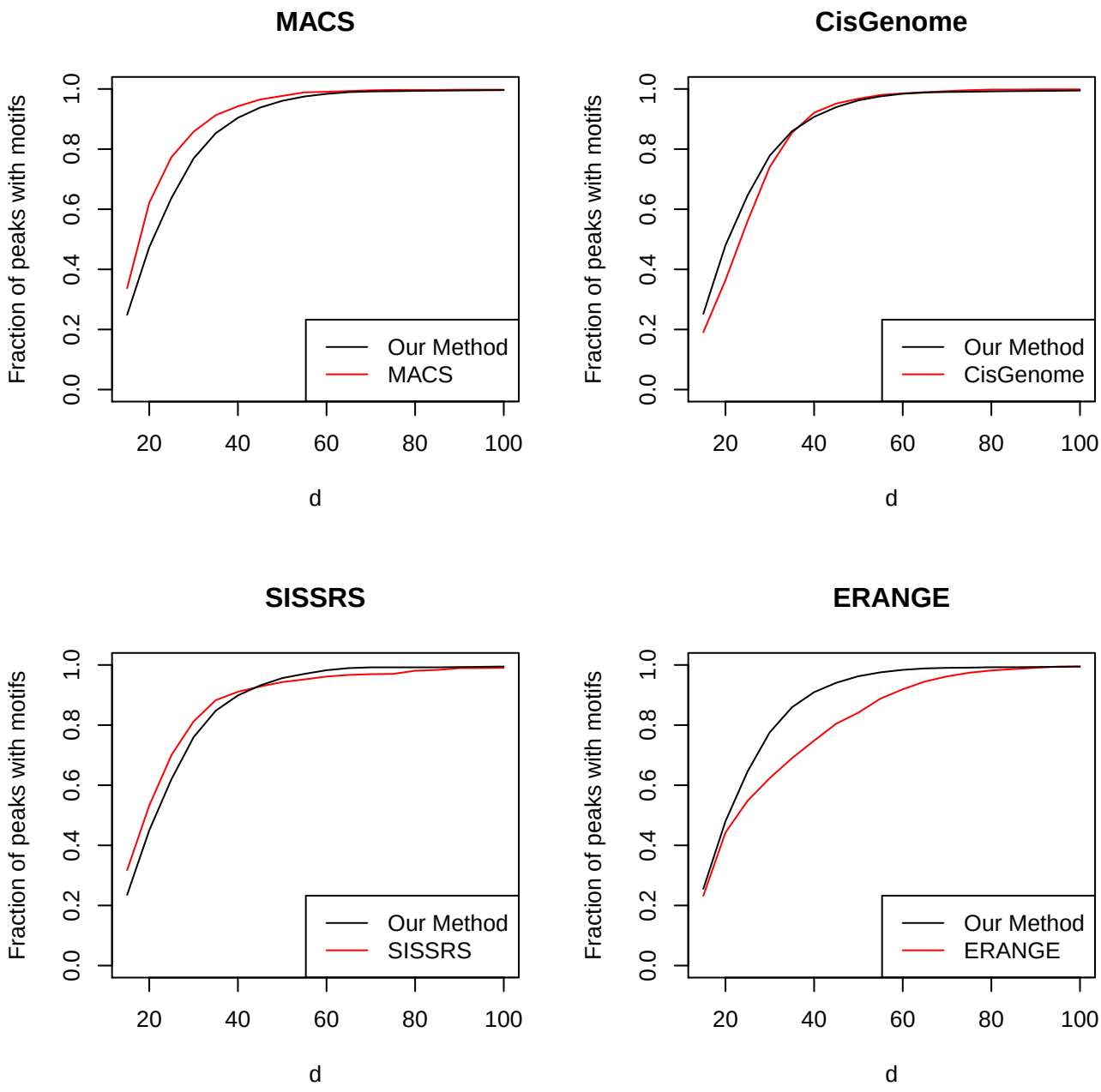

Figure 3.22: Spatial resolution of various one-sample peak-finders. The fraction of common motif-containing peaks with the motif within d bps of the summit is plotted vs. d.

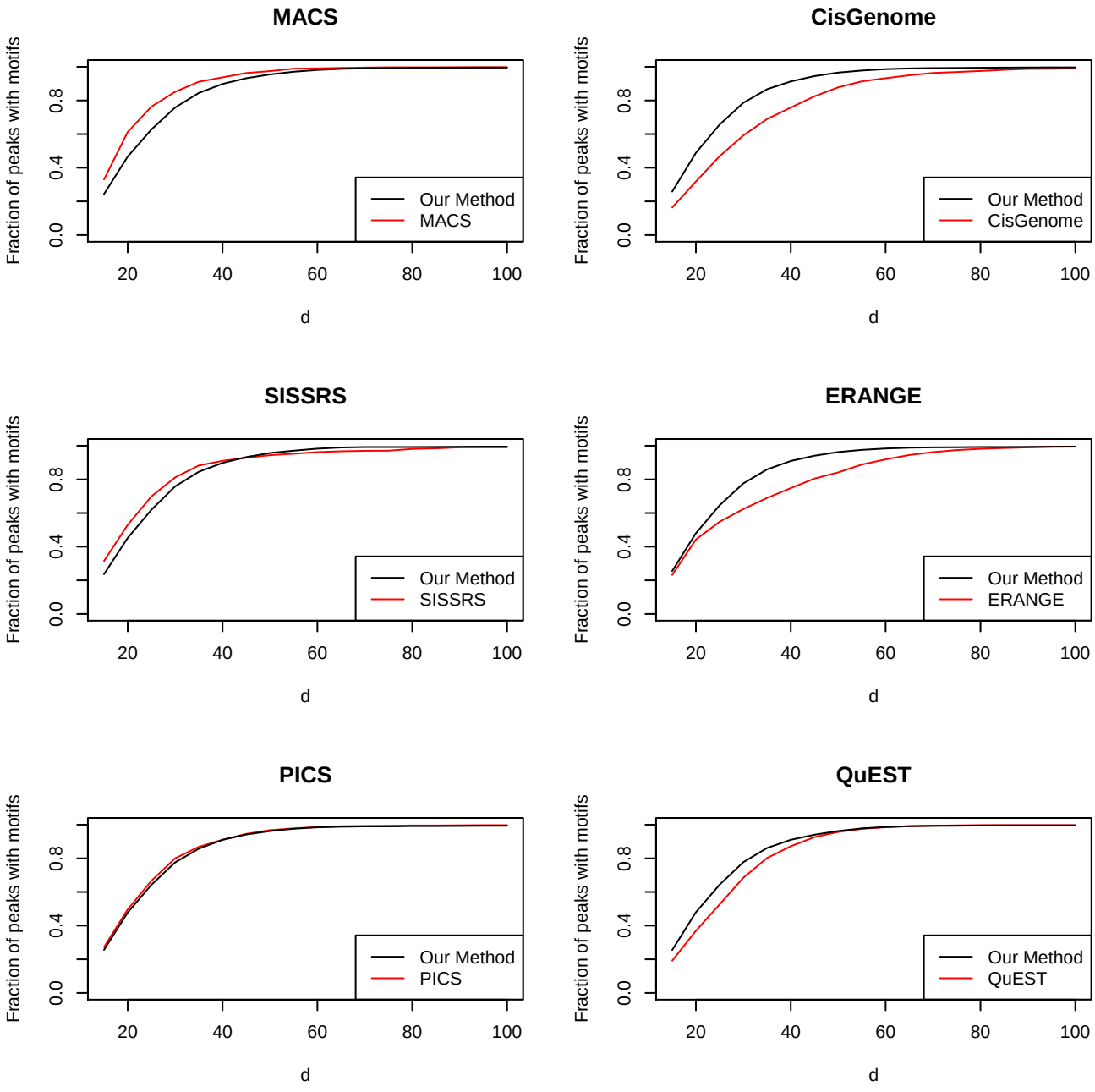

Figure 3.23: Spatial resolution of various two-sample peak-finders. The fraction of common motif-containing peaks with the motif within d bps of the summit is plotted vs. d.

$\widehat{p}_i(S=1) = \#(R_g^{i-1})/n$ and $p(S=0)$ as $1 - \widehat{p}_i(S=1)$, which allows us to estimate the posterior probability $p(S=1|F_i)$ as $\widehat{p}_i(S=1|F_i) = \widehat{p}_i(F_i|S=1)\widehat{p}_i(S=1)/\widehat{p}_i(F_i)$, where $F_i = F_{Z(i)}$ during the First Iterative Step and $F_i = F$ during the Second Iterative Step and $\widehat{p}_i(F_i) = \widehat{p}_i(F_i|S=1)\widehat{p}_i(S=1) + \widehat{p}_i(F_i|S=0)\widehat{p}_i(S=0)$. We find that this alternative approach peforms somewhat better in the data sets that show wider, less peaky signal, e.g. GABP data set from [29], but performs worse with our ERβ data set. What determines which of the two approaches performs better and whether they can be combined into a single procedure remains to be seen.

Chapter 4

Combining Information from Replicate Samples

4.1 Introduction

In this chapter we discuss the important issue of combining data from replicate samples. We discuss different kinds of replicates common in the ChIP-Seq literature and the standard approaches used to integrate data across replicates. We develop several diagnostic plots for assessing whether the standard assumption of Poisson variance holds and observe that the assumption can break down even for technical replicates due to flow cell-specific sequence composition effects. In particular we find that the difference in the relationship of read density and GC content between replicates is the driving force behind the observed departures from the standard model.

4.2 Technical and Biological Replicates in ChIP-Seq Data

An important issue in the analysis of ChIP-Seq data is the incorporation of information across replicate samples. Replicate studies are often done in biology to

1. guard against the effects of experiment-specific artifacts and assess reproducibility of the obtained results,

2. quantify the uncertainty associated with the measure of interest,

3. increase the amount of available data. This is particularly pertinent to ChIP-Seq experiments since a single lane of data might not yield enough information for proper enrichment profile assessment.

In the context of ChIP-Seq, the term "replicates" refers to separately sequenced samples interrogating the same enrichment profile. Replicates are classified into 'technical' and 'biological' based on the nature of their relationship. If the replicates originate from different cell populations representing the same biology and undergo the same library preparation protocol, they are referred to as 'biological' replicates. If a single library is split into 2 samples to be sequenced on different lanes of the same or a different flow cell, the samples are referred to as 'technical' replicates. Other kinds of replicate samples are also possible: starting with the same cell population, the investigator might use different chromatin shearing approaches, antibodies or library prep protocols. These samples are usually referred to as 'technical' replicates as well, and the degree of similarity between different kinds of technical replicates might be quite variable.

The natural question is how to combine the information from separate replicate samples (lanes of sequenced reads). One common approach is to pool reads from technical replicates of the same sample into a single super-sample [2, 7, 22]. The intuition behind this approach is that technical replicates should not be too different from each other and the combined sample will provide a better estimate of the enrichment profile by boosting the signal at the weak binding sites. Some authors also combine reads from biological replicates following the similar line of reasoning [21]. If replicate-specific artifacts are a major presence, then combining the reads will dilute the enrichment at these sites. Whether this will suffice to exclude such artifactual regions from the set of identified binding sites is likely to depend on the particular data set under study.

Another common approach is to identify a set of binding events or peaks for each replicate sample separately and then to intersect the invidual peak lists to obtain the final set of peaks [1]. This is a rather conservative approach and might prevent the investigator from successfully detecting the weaker binding sites.

More sophisticated approaches exist as well. The R package DESeq [41], originally designed for RNA-Seq data analysis, models the variation of normalized gene counts in replicate samples using a negative binomial distribution. The authors of the package mention briefly that this approach can be extended to ChIP-Seq data as well, using pre-defined regions of interest as 'genes' and looking at 'differential expression' in treatment relative to control.

4.3 Consistency of Technical Replicates

4.3.1 Data Set Description

To illustrate the two basic approaches to combining replicates and to explore some issues that arise when analyzing replicate ChIP-Seq data we will use the data set on the binding of transcription factor PIF3 in *A.thaliana*, described in section 3.2.

Table 4.1 shows the numbers of peaks identified by the peak-finder MACS for the samples described above. Replicate-specific peaks were identified using WT IP samples from same

flow cell as controls for P3M IP treatment samples. There is a large difference between the number of peaks identified using pooled reads from technical replicates and the number of peaks identified in both replicate-specific analyses.

Peak set	Number of peaks	% motif occurrence
Exp3 rep1 peaks	376	51%
Exp3 rep2 peaks	415	53%
Exp3 common peaks	201	63%
Exp3 pooled reads peaks	696	44%
Exp4 rep1 peaks	1,151	27%
Exp4 rep2 peaks	2,145	22%
Exp4 common peaks	590	40%
Exp4 pooled reads peaks	2,517	22%

Table 4.1: Numbers of peaks identified by MACS in arabidopsis data set. "Common peaks" are obtained by intersecting the replicate-specific peak sets. "Pooled reads peaks" are obtained by pooling technical replicates of treatment and control samples for the same biological (exp3 or exp4) replicate.

4.3.2 Standard Poisson Model and its Fit

It is common in the RNA-Seq literature to assume that for technical replicates the underlying normalized read rate for each gene is the same in all replicates [42]. A popular approach is to model read counts X_{ij} in gene j for replicate sample i as Poisson($\lambda_j L_i$), where L_i is a tuning or normalization constant (normally, the total number of aligned reads for that sample). It is assumed that $X_{i_1 j}$ is independent of $X_{i_2 j}$ and that X_{ij_1} is independent of X_{ij_2}. The intuition behind the first assumption is that $N_j >> \max(X_{i_1 j}, X_{i_2 j})$, where N_j is the total number of fragments from gene j in the pool of fragments from which samples i_1 and i_2 are drawn. The intuition behind the second assumption is that usually $L_i >> \max(X_{ij_1}, X_{ij_2})$. Sometimes, negative binomial (over-dispersed Poisson) distribution is used as the model for gene read count variability. We can extend this approach to ChIP-Seq data by considering read counts in genomic bins of a given size instead of the genes.

To assess how well the Poisson model fits the observed data for a pair of technical ChIP-Seq replicates, we divide the genome into 1kb bins and fit the model to read counts in those bins. Under the model, the counts X_{ij} in bin j for replicate sample $i \in \{1, 2\}$ folow Poisson($\lambda_j L_i$) distribution, where we let L_i be the lane total (number of aligned reads) for replicate i. Estimating λ_j by maximum likelihood yields $\widehat{\lambda}_j = \sum_i X_{ij} / \sum_{i,j} X_{ij}$. This has the intuitive interpretation of reads in replicate bins being drawn from the same fragment pool

in proportion to their respective total sample sizes. For each bin j we can now compare the observed read counts X_{1j}, X_{2j} to their expected values $\widehat{\lambda}_j L_1$, $\widehat{\lambda}_j L_2$ to obtain a χ^2 statistic with 1 degree of freedom and the corresponding p-value. If the model fits the data well, we should observe uniform distribution of the resulting 117,574 (number of non-overlapping 1kb bins in the *A.thaliana* genome) p-values. Figure 4.1 shows the histograms of p-values for the 4 pairs of technical replicates. There is enrichment for low p-values in 3 out of 4 pairs, indicating some lack of fit.

The observed lack of fit might conceivably result from bins with low read counts, where χ^2 approximation may not hold. However, stratifying the p-values by the quartiles of the average read counts across replicates shows that lack of fit is present across the entire range of read counts. Figure 4.2 shows this for exp3 P3M replicates; note that the first quartile of total counts is 57 reads, a rather large number. This can also be seen in M-D plots (Figure 4.3) where we plot $D_j = \log_2(X_{1j}/L_1) - \log_2(X_{2j}/L_2)$ vs. $M_j = (\log_2(X_{1j}/L_1) + \log_2(X_{2j}/L_2))/2$, i.e. the difference vs. the mean of bin read counts on $\log_2$ scale. The loess fit lines are different for different pairs of replicates but seem to be similar for pairs of replicates on same flow cells. This indicates that there might be a flow cell effect behind the observation of lack of fit.

Table 4.2 shows the numbers of genomic 1kb bins (out of 117,574) with χ^2-derived p-values < 0.01 common to different pairs of technical replicates. Under the model assumptions, we expect each pair of replicates to produce about 1,200 bins with p-values < 0.01 and 3 out of 4 pairs of replicates show enrichment for such bins by a factor of 2-4 (exp4 P3M replicates show considerably better concordance than the other 3 pairs). Few of the bins with small p-values are common between pairs of replicates: out of roughly 10,000 bins that show lack of fit (p-value < 0.01) in at least one pair of replicates, only 420 show lack of fit in at least 2 pairs, 21 show lack of fit in at least 3 pairs and 6 show lack of fit in all 4 pairs. Thus, the lack of fit is not mostly due to some genomic regions that tend be inherently more variable (the few low p-value bins that are common to all replicate pairs are very read-rich and are located in telomeric and centromeric regions).

	Exp3 P3M	Exp3 WT	Exp4 P3M	Exp4 WT
Exp3 P3M	2,356	194	33	57
Exp3 WT	194	4,092	64	99
Exp4 P3M	33	64	1,318	33
Exp4 WT	57	99	33	2,072

Table 4.2: Numbers of common low p-value bins. The entry in cell (i, j) shows the number of bins with p-value < 0.01 in both replicate pairs i, j. The diagonal entries (i, i) are total numbers of bins with p-value < 0.01 in replicate pair i.

The Poisson model specified above can be reparametrized as

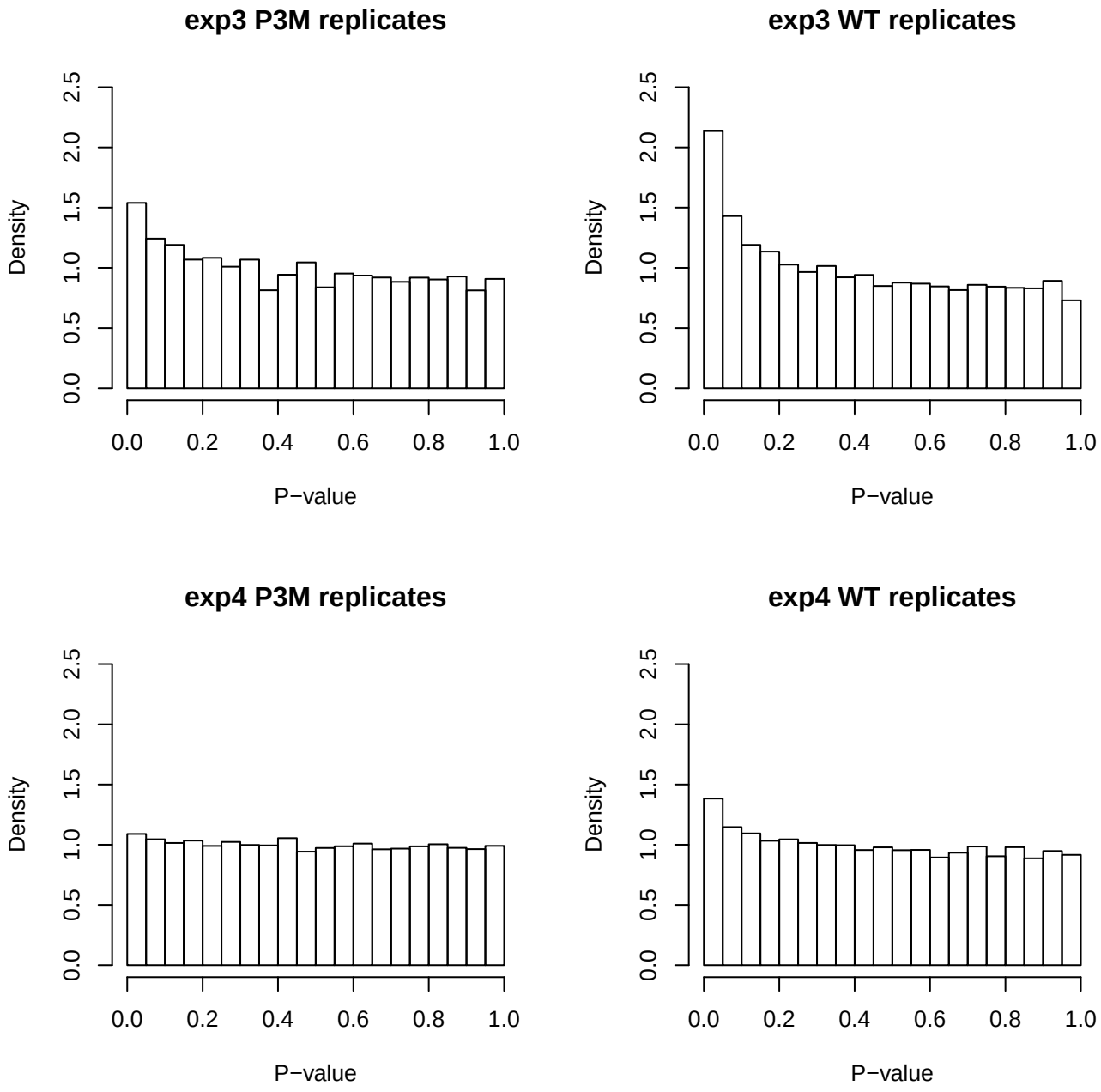

Figure 4.1: P-values for χ^2 statistics of goodness-of-fit of Poisson model to replicate pairs. Excess of small p-values for 3 out of 4 pairs of technical replicates (all except exp4 P3M replicates) indicates lack of fit.

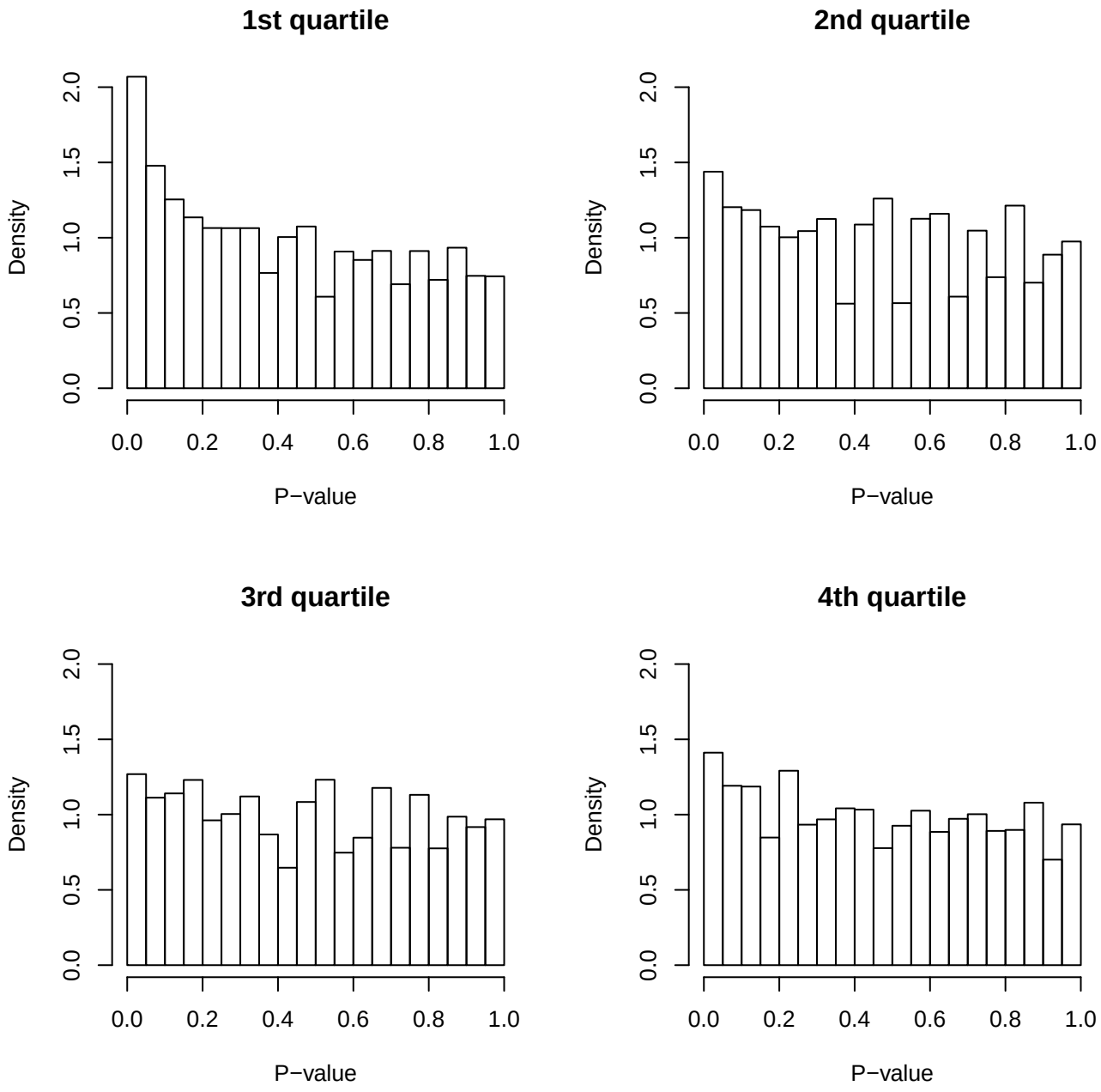

Figure 4.2: P-values for χ^2 goodness-of-fit statistics of Poisson model to exp3 P3M technical replicates, stratified by quartiles of total $(X_{1j} + X_{2j})$ counts. The pairs of exp3 P3M and exp4 WT similarly show enrichment for low p-values across the quartiles of total counts.

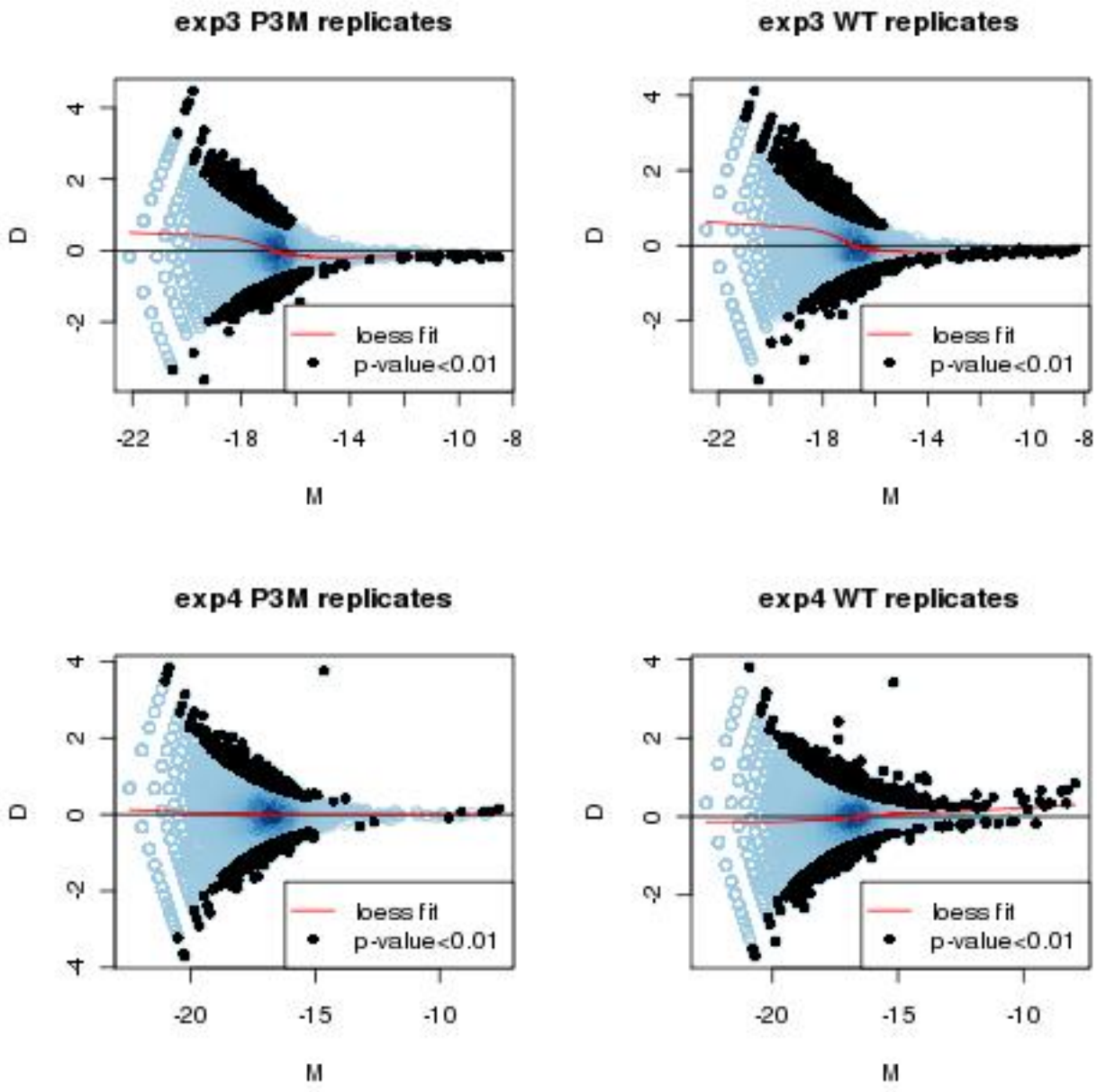

Figure 4.3: M-D plots for 1kb bin read counts for technical replicates. Low p-value bins occur throughout the range of average bin counts.

$$X_{ijk} \sim \text{Poisson}(\exp(\log(L_{ik}) + \eta_{jk}))$$

where $\eta_{jk} = \log(\lambda_{jk})$ and the extra subscript k refers to the library and is suppressed when we consider pairs of technical replicates (samples from the same library). This log-linear model can be fit simultaneously to all 8 lanes under consideration here, and we use the resulting fitted values to produce a set of χ^2 statistics (on 4 degrees of freedom) and corresponding p-values for 1kb genomic bins. We can also introduce a flow-cell covariate into the log-linear model to try to take flow-cell effects into account: $X_{ijk} \sim \text{Poisson}(\exp(\log(L_{ik}) + \eta_{jk} + fc(i,k)))$, where $fc(i,k)$ is the parameter for the flow cell containing replicate i of library k . We can compare the χ^2 statistics (2 d.f.) and corresponding p-values from this model to the one with no flow-cell effects. There are 3,858 bins with p-value < 0.01 in the model without flow-cell effect and 1,560 such bins in the model with flow-cell effect (we expect 1,176 or so if the model is correct). Thus, flow cell effects account for a large portion (but not all) of the observed lack of fit. Even after accounting for flow cell effects, there seem to be some lane-specific effects. These effects cannot be meaningfully estimated in the current setting but an investigation might be possible in the future based on samples multiplexed across several lanes.

4.3.3 Mean-Variance Relationship

We find the following plot to be very informative in the investigation of the relationship of technical replicates in the data. Let

$$M_j = (X_{1j} + X_{2j})/(L_1 + L_2)$$
$$V_j = \frac{(X_{1j}/L_1 - X_{2j}/L_2)^2}{1/L_1 + 1/L_2}$$

be the measures of mean and variance, respectively, for the replicate read counts in bin j, normalized by lane totals. Under the Poisson model assumptions stated above, $E(M_j) = E(V_j) = \lambda_j$. In particular, M_j is the maximum likelihood estimate of λ_j and $V_j = 0$ iff $X_{1j}/X_{2j} = L_1/L_2$ or $X_{1j} = X_{2j} = 0$ (thus V_j is large when the read count ratios are very different from the ratio of normalizing constants). Also, if we let $X'_{1j} = kX_{1j}, , X'_{2j} = kX_{2j}$ then $M'_j = kM_j$, but $V'_j = k^2 V_j$. Thus increasing bin counts by some factor $k > 1$ without corresponding change in lane totals results in higher variance-to-mean ratio for the same bin count ratio.

If the Poisson model assumptions hold, the scatter plot of V_j vs. M_j should be roughly linear, following the line $y = x$. For greater readability, we plot the values (M_j, V_j) after binning based on the quantiles of M_j. With this procedure, the point corresponding to the highest quantile of M_j is always an outlier (it contains the super-enriched telomeric and centromeric artifacts, mentioned in section 1.4) and is omitted from the plot. The M-V plots for the 4 pairs of replicates and different bin sizes are shown in Figures 4.4-4.7. There are

departures from the expected behavior for the same three out of four replicate pairs and the departures become more prominent as bin size increases. There is noticeable overdispersion relative to the Poisson model but the shape of the overdispersion varies depending on the pair of replicates. Two out of four replicate pairs exhibit a strange 'cubic' shape pattern while another shows a 'quadratic' pattern which might perhaps be modeled with the negative binomial distribution (which accomodates a quadratic mean-variance relationship).

Interestingly, reducing all duplicate alignments to a single copy (a commonly utilized strategy based on the assumption that those alignments represent PCR artifacts, as described in section 2.1.6) results in under-dispersion, as shown for exp4 WT replicates in Figure 4.8 (the other replicate pairs show similar behavior). This emphasizes the importance of not discarding the data unnecessarily.

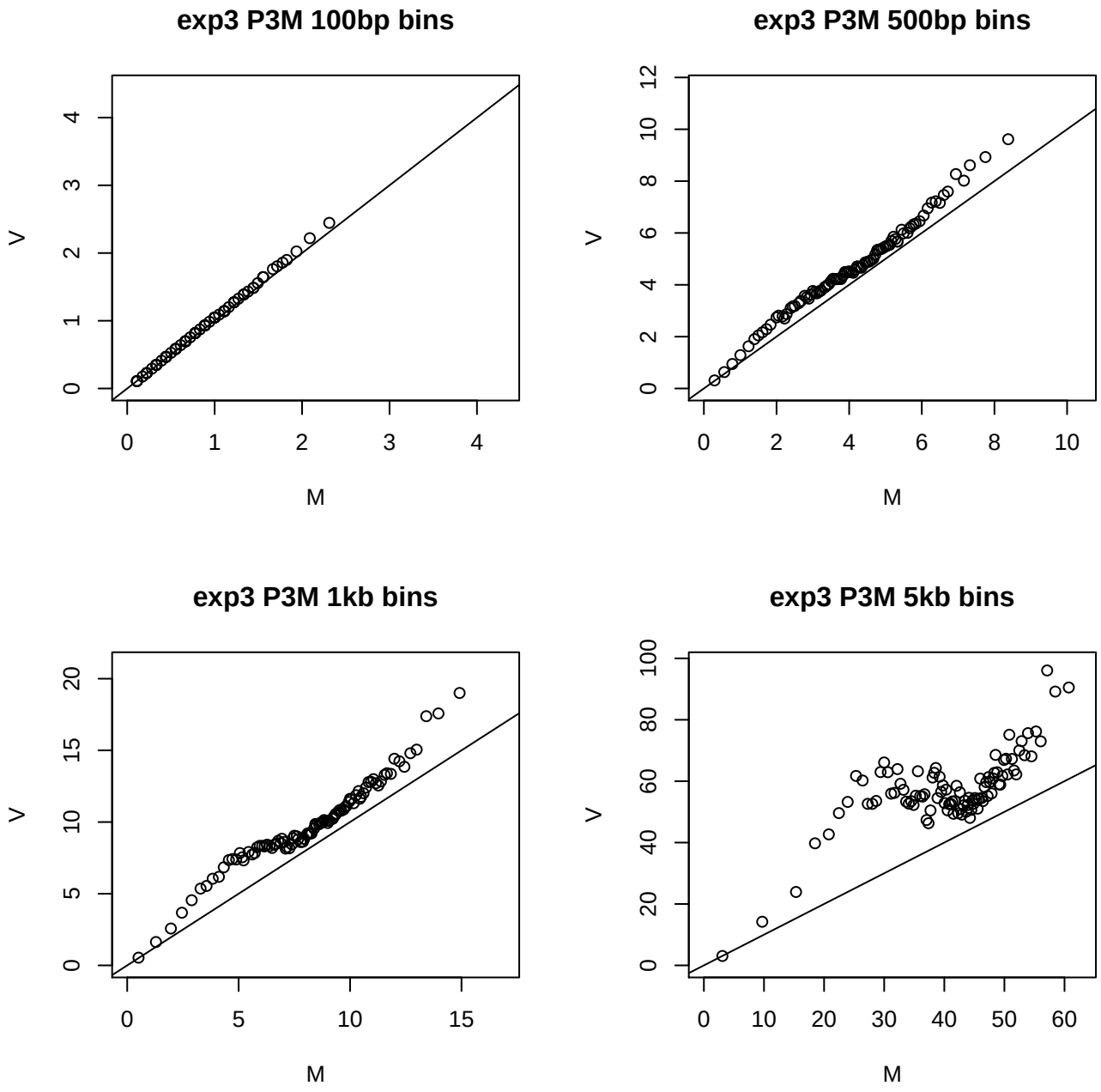

Figure 4.4: M-V plots for exp3 P3M technical replicates. Lane totals were scaled by 10^6 to make x- and y-axis easier to read.

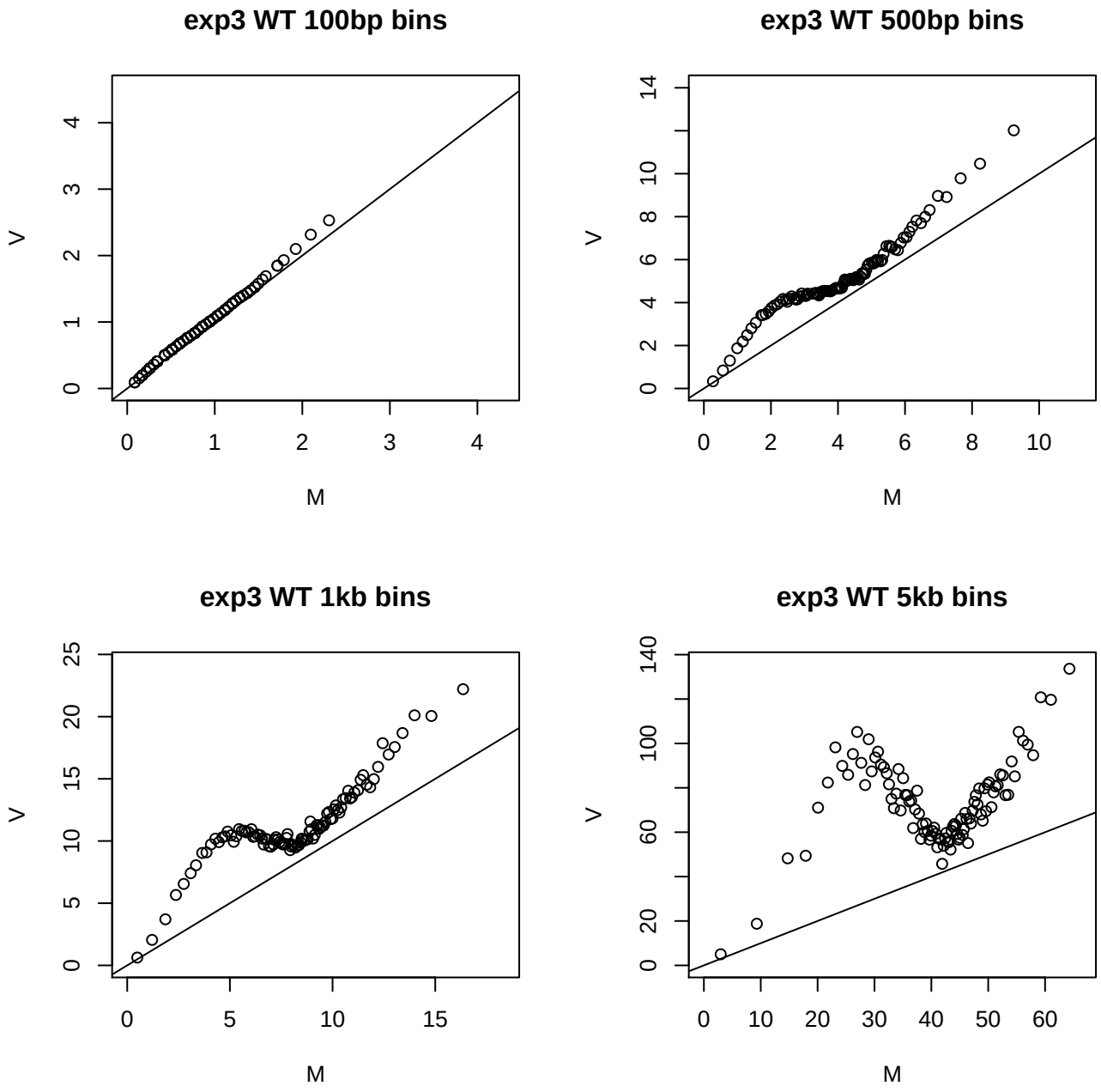

Figure 4.5: M-V plots for exp3 WT technical replicates. Lane totals were scaled by 10^6 to make x- and y-axis easier to read.

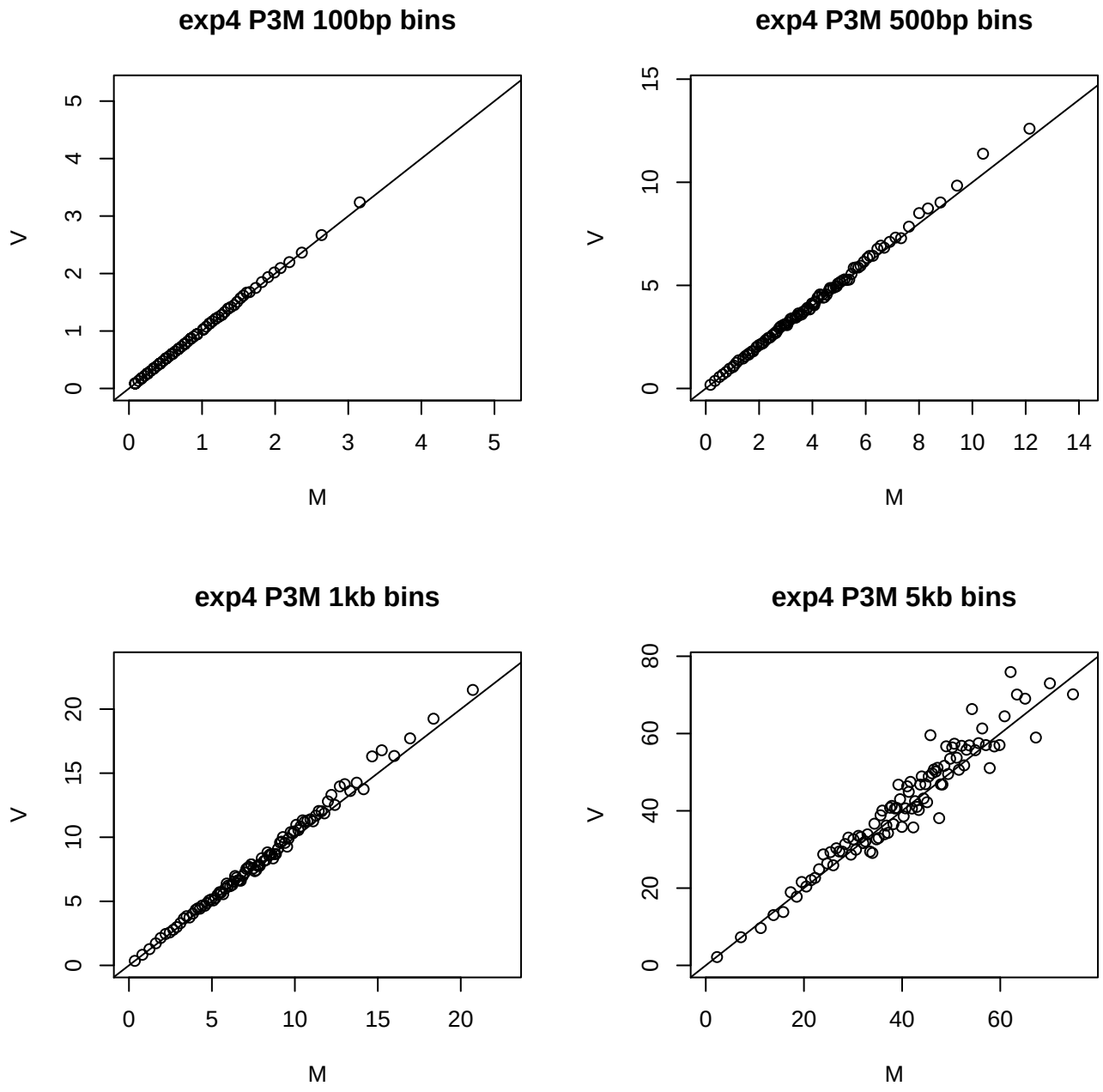

Figure 4.6: M-V plots for exp4 P3M technical replicates. Lane totals were scaled by 10^6 to make x- and y-axis easier to read.

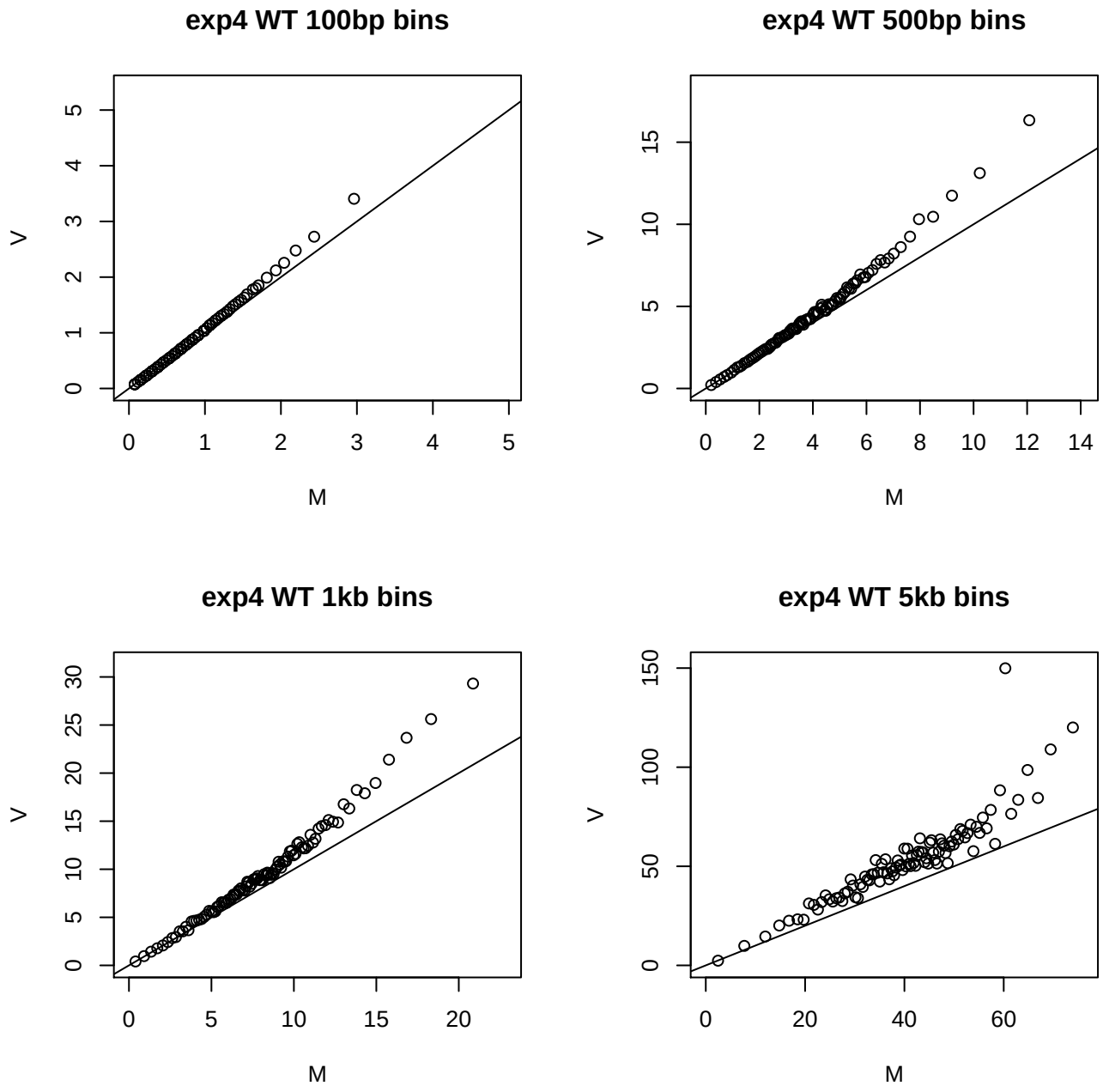

Figure 4.7: M-V plots for exp4 WT technical replicates. Lane totals were scaled by 10^6 to make x- and y-axis easier to read.

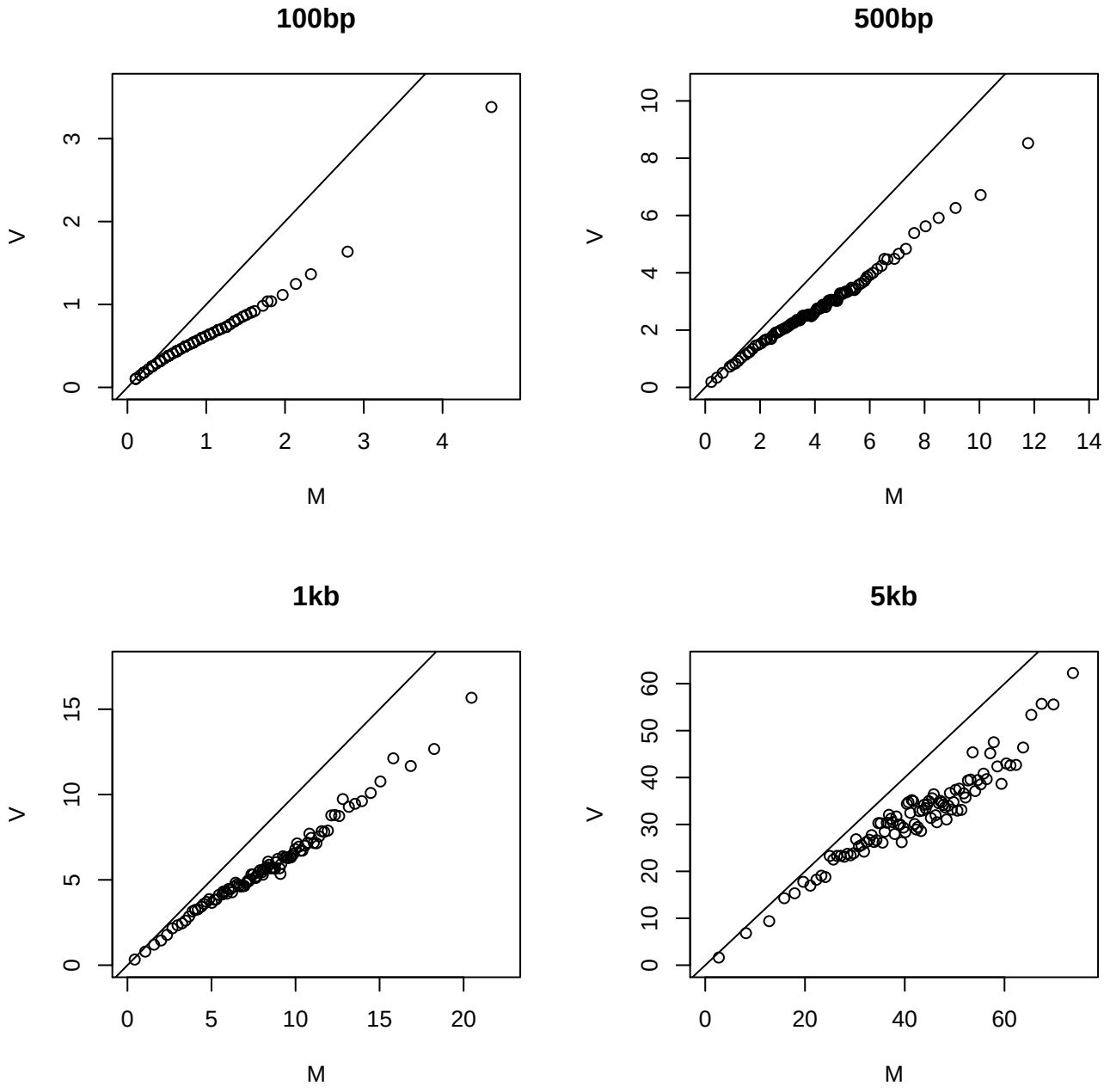

Figure 4.8: M-V plots for exp4 WT technical replicates after reducing all duplicate alignment to a single copy.

We can derive the conditions under which $V > M$ (dropping the subscript j for the moment) as follows:

$$\begin{aligned}
V &> M \\
&\Leftrightarrow \\
\frac{\left(\frac{X_1}{L_1} - \frac{X_2}{L_2}\right)^2}{\frac{1}{L_1} + \frac{1}{L_2}} &> \frac{X_1 + X_2}{L_1 + L_2} \\
&\Leftrightarrow \\
\left(\frac{X_1}{L_1} - \frac{X_2}{L_2}\right)^2 &> \frac{X_1 + X_2}{L_1 L_2} \\
&\Leftrightarrow \\
\frac{X_1^2}{L_1^2} - \frac{X_1 + 2X_1X_2}{L_1L_2} &> \frac{X_2}{L_1L_2} - \frac{X_2^2}{L_2^2} \\
&\Leftrightarrow \\
\left(\frac{X_1}{L_1} - \frac{1 + 2X_2}{2L_2}\right)^2 &> \frac{X_2}{L_1L_2} - \frac{X_2^2}{L_2^2} + \left(\frac{1 + 2X_2}{2L_2}\right)^2 \\
&\Leftrightarrow \\
\left|\frac{X_1}{L_1} - \frac{1 + 2X_2}{2L2}\right| &> \sqrt{\frac{X_2}{L_1L_2} + \frac{X_2}{L_2^2} + \frac{1}{4L_2^2}}
\end{aligned}$$

The expression under the square root sign is always positive. We can simplify further to obtain the relationship $V > M \Leftrightarrow$:

$$\begin{aligned}
X_1 &\notin \left(L_1\left(\frac{1 + 2X_2}{2L_2} - \sqrt{\frac{1}{4L_2^2} + \frac{X_2}{L_2}\left(\frac{1}{L_1} + \frac{1}{L_2}\right)}\right), L_1\left(\frac{1 + 2X_2}{2L_2} + \sqrt{\frac{1}{4L_2^2} + \frac{X_2}{L_2}\left(\frac{1}{L_1} + \frac{1}{L_2}\right)}\right)\right) \\
&\Leftrightarrow \\
X_1 &\notin \left(r\left(\frac{1 + 2X_2}{2}\right) - \sqrt{\frac{1}{4}r^2 + X_2(r + r^2)}, r\left(\frac{1 + 2X_2}{2}\right) + \sqrt{\frac{1}{4}r^2 + X_2(r + r^2)}\right)
\end{aligned}$$

where $r = L_1/L_2$.

The 'good' ($V \leq M$) interval for X_1 is not centered on rX_2 but on a slightly higher value and the width of the interval grows as $\sqrt{X_2}$. This means that as X_2 increases, the value of X_1 has to increase linearly with it to keep the ratio of read counts the same as the ratio of normalizing constants, but the effective width of the 'good' interval shrinks since it is only increasing as the square root of X_2. This explains why departures from linearity are more easily detected with larger bin sizes in M-V plots.

It will be shown that L_1/L_2 may not always be the appropriate normalization (background) ratio and that a different value of r might restore the proper $M = V$ relationship. As an example, let $L_1/L_2 = 0.75$, while the true (e.g., GC-dependent) background ratio is 0.6. If we let $X_2 = 10$ and build the intervals for X_1 corresponding to $V \leq M$ (as given

above) using the 2 different values of r, we find that the intervals overlap by about 75%. However, if we let $X_2 = 100$ and build the 2 intervals, they overlap by about 30%. This again demonstrates that at lower resolution it is easier to detect that underlying background ratio is different from the lane total ratio and also that the departures are easier to see in higher count bins than in lower count bins.

It would be beneficial to discover the underlying cause of these departures from $M = V$ relationship. This behavior and the associated 'cubic' M-V pattern are not unique to this data set and have been observed in other ChIP-Seq experiments (Davis McCarthy, personal communication).

One possible explanation is that the increased variability is due to the regions with low mappability. Figures 4.9-4.12 show the M-V plots for technical replicate pairs after discarding all genomic bins with fewer than 90% of mappable bases (depending on bin size, this results in discarding 15%-20% of all bins). The high-mappability bins retain the abnormal M-V relationship, although the 'cubic' pattern is changed to a U-shape. We conclude that the only effect of low-mappability bins on the M-V plot is to lower the variance for the bins corresponding to the lowest mean quantiles. One can also check to see if the bins with different mappability content require different normalization ratio r to recover the proper mean-variance relationship. Figure 4.13 shows the boxplot of raw read count ratios X_{1j}/X_{2j} for bins stratified by their mappability content (restricted to 5kb bins, since this lower resolution allows for the easier detection of the lack of fit), showing that the ratios are consistent across mappability strata and are centered around the lane total ratio.

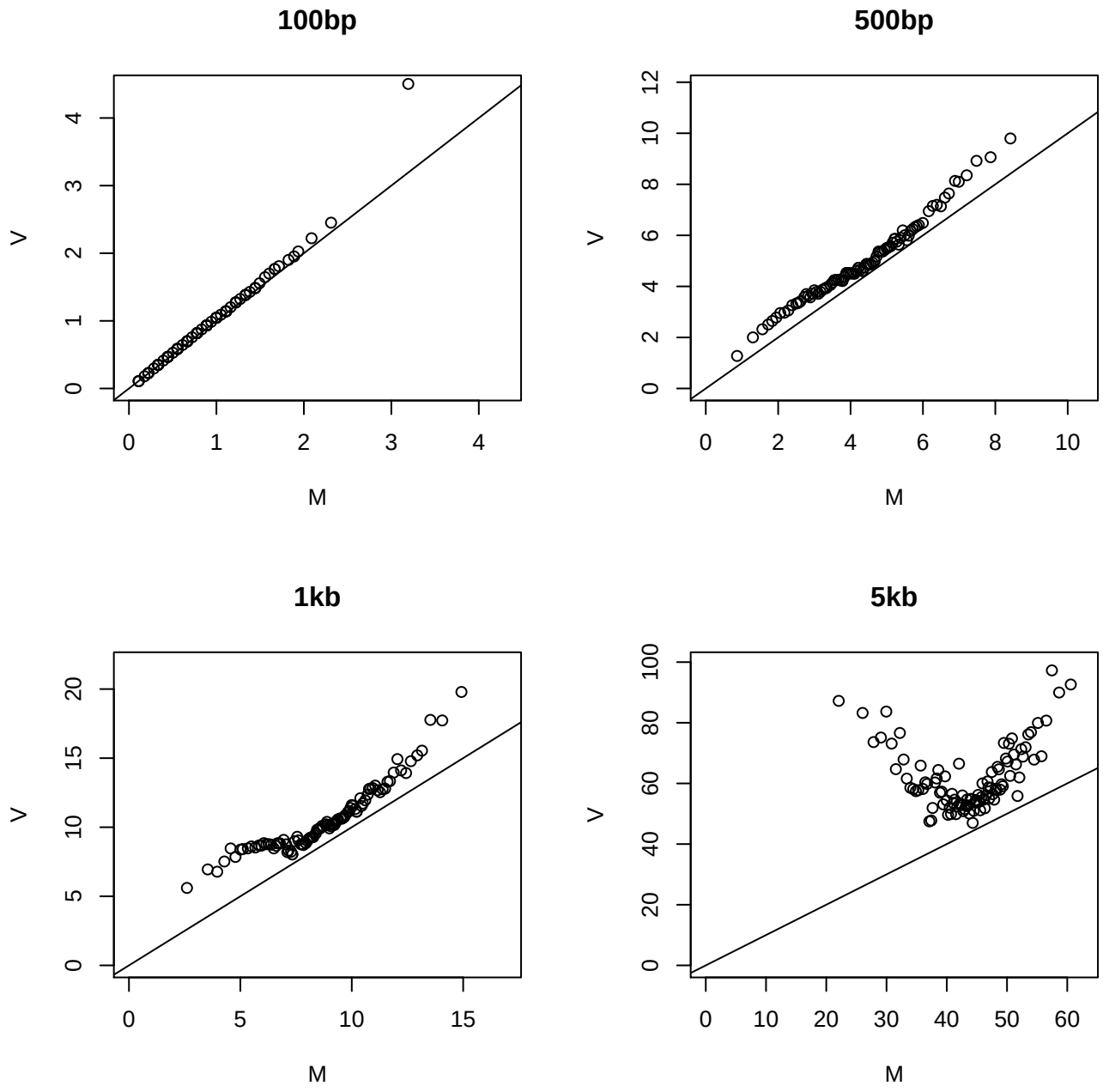

Figure 4.9: M-V plots for exp3 P3M technical replicates, after excluding all bins with mappability < 0.9. Lane totals were scaled by 10^6 to make x- and y-axis easier to read.

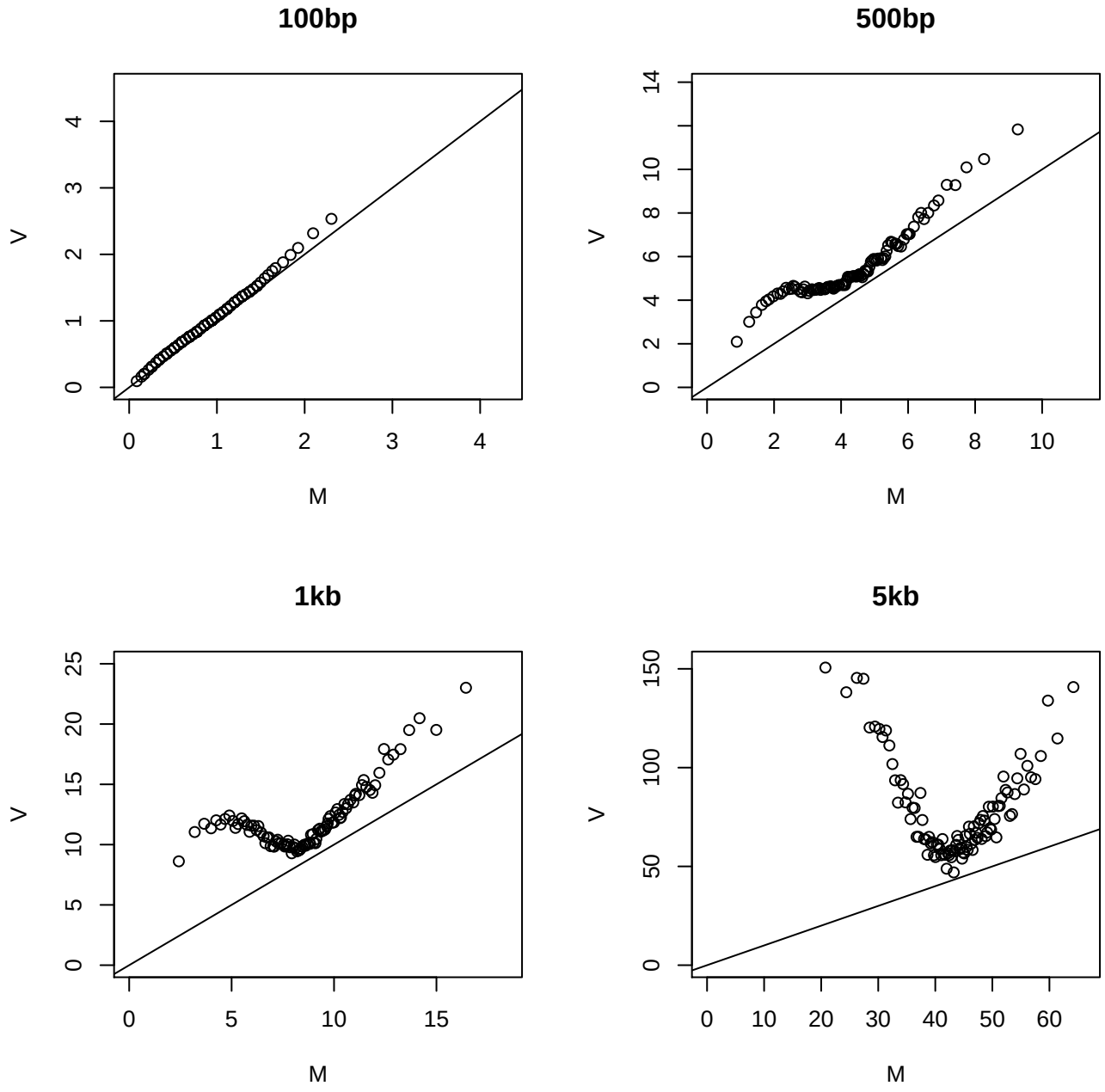

Figure 4.10: M-V plots for exp3 WT technical replicate, after excluding all bins with mappability < 0.9s. Lane totals were scaled by 10^6 to make x- and y-axis easier to read.

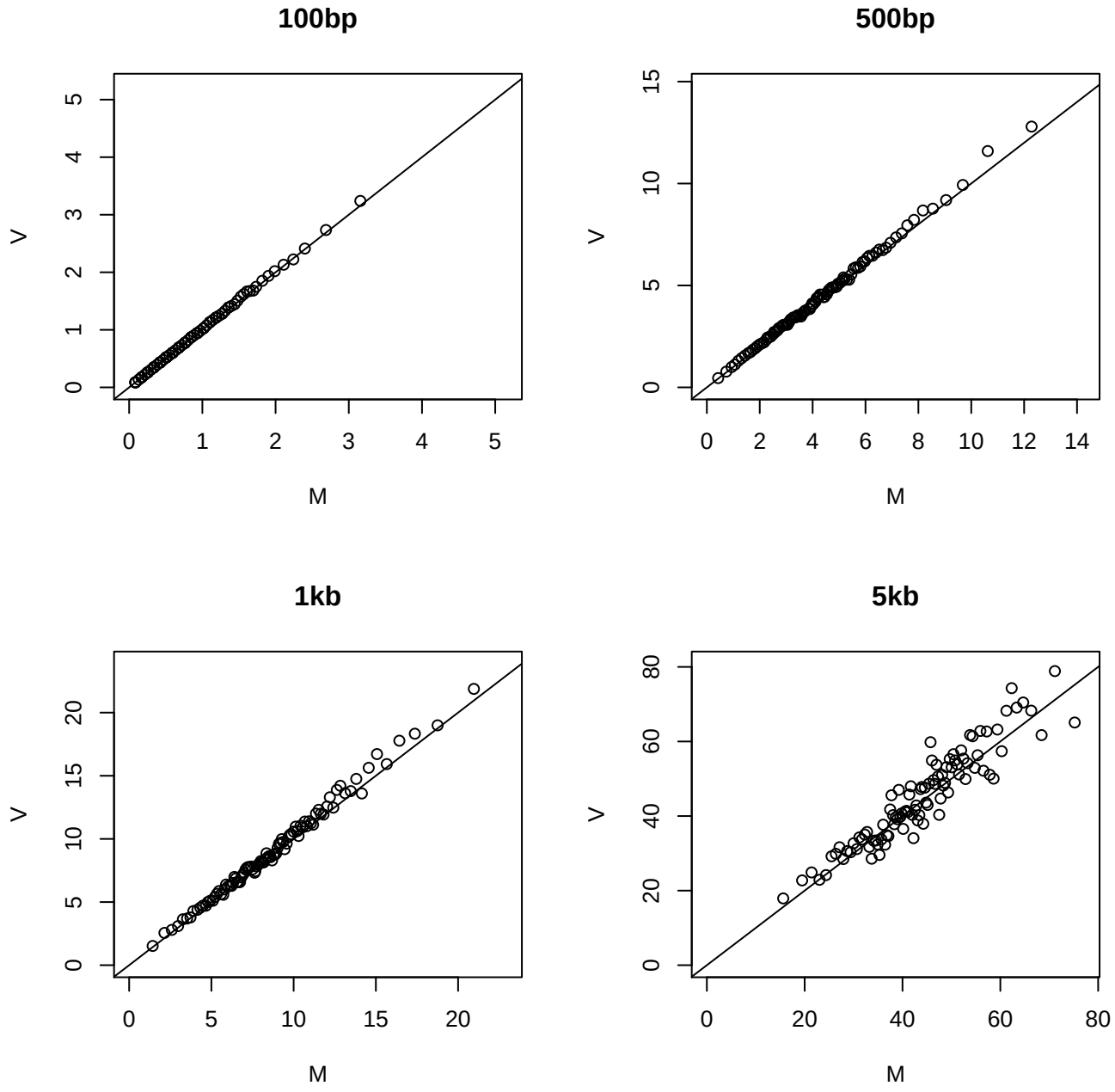

Figure 4.11: M-V plots for exp4 P3M technical replicates, after excluding all bins with mappability < 0.9. Lane totals were scaled by 10^6 to make x- and y-axis easier to read.

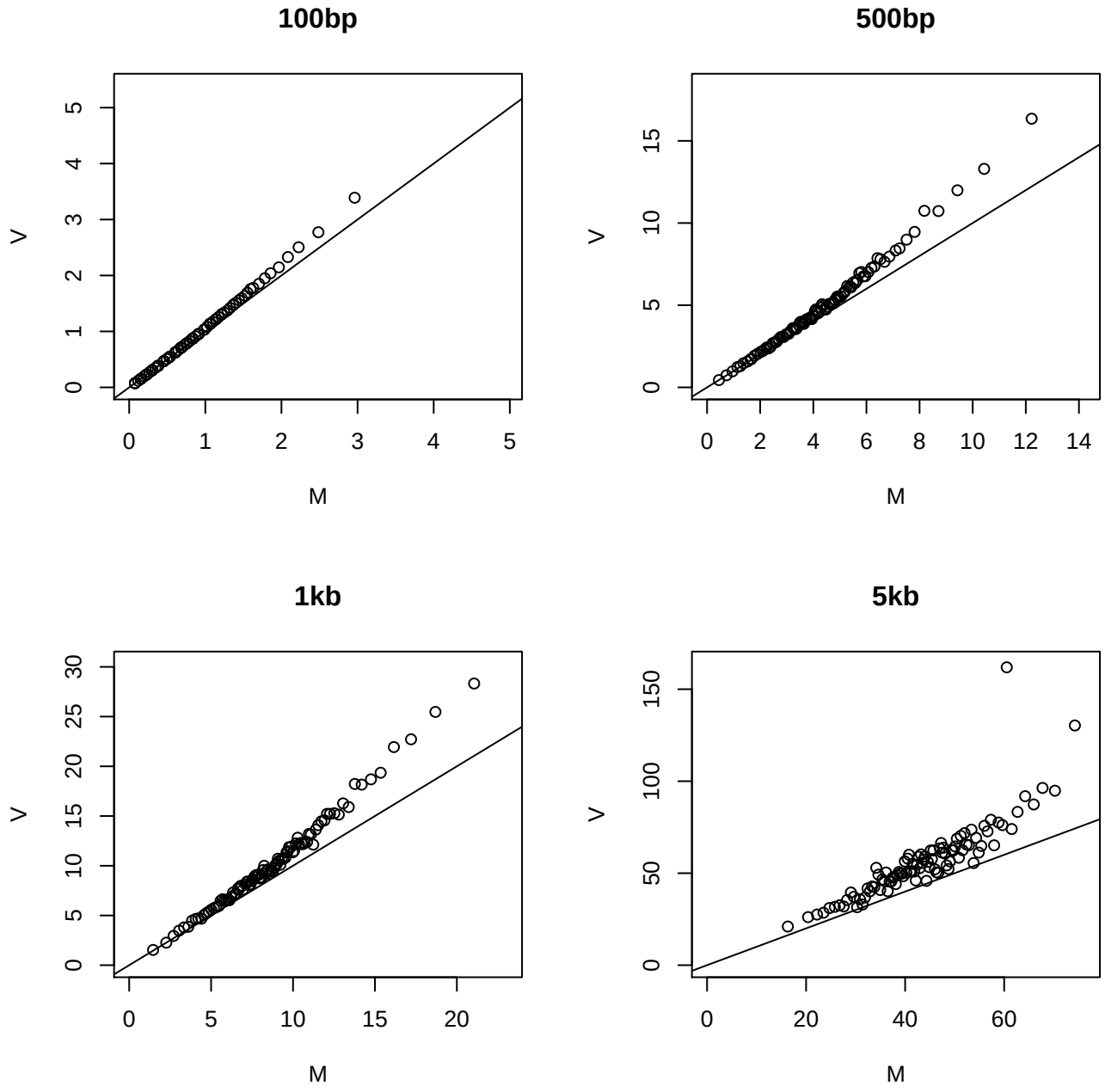

Figure 4.12: M-V plots for exp4 WT technical replicates, after excluding all bins with mappability < 0.9. Lane totals were scaled by 10^6 to make x- and y-axis easier to read.

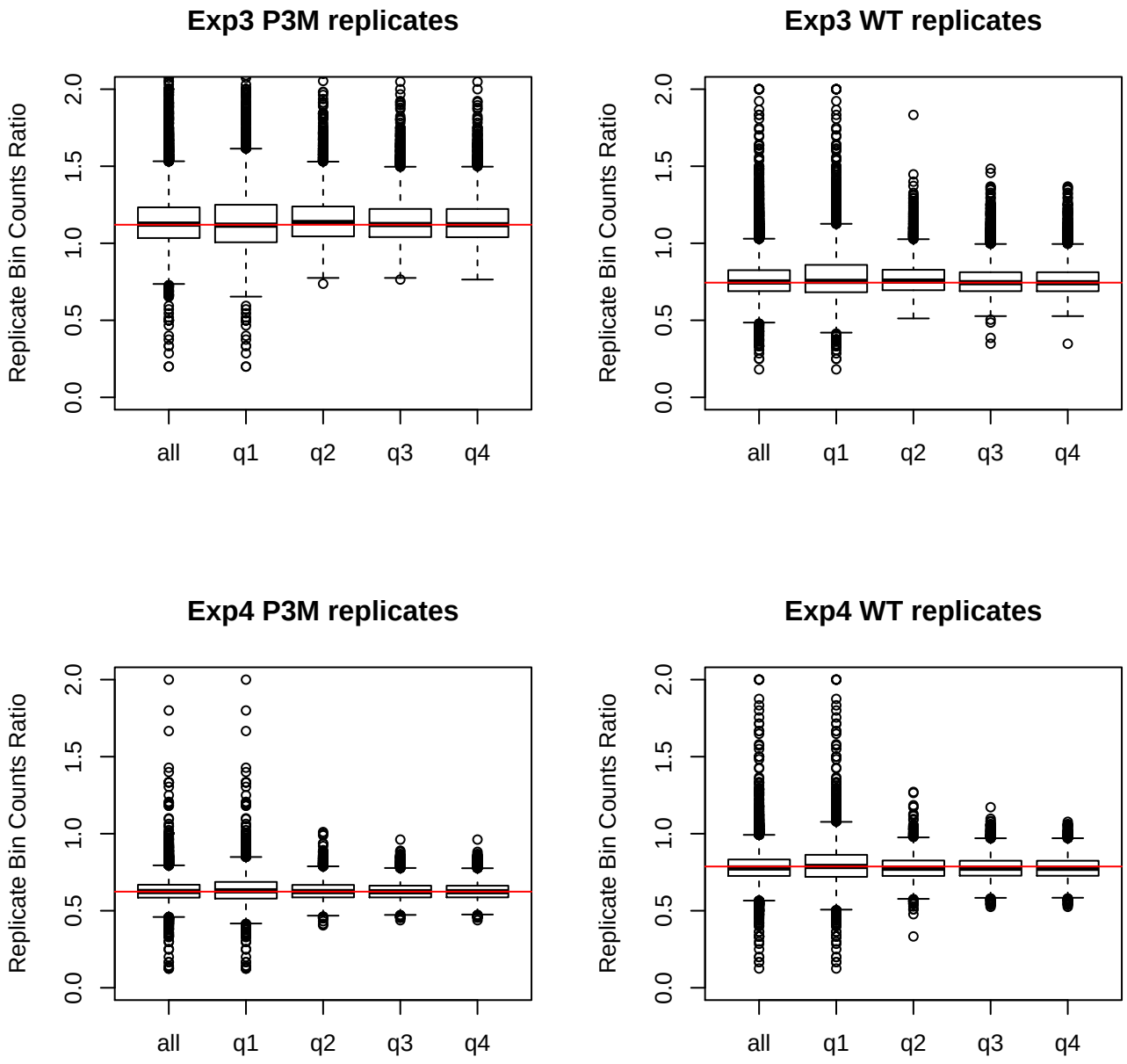

Figure 4.13: Ratios of X_{1j}/X_{2j} stratified by mappability content (5kb bins). q1-4 refer to ranges based on quartiles of the distribution of mappability. Horizontal red line is the ratio of lane totals. The ratios do not seem to be mappability-dependent.

Another explanation is that the flow cell environment somehow affects the enrichment profile of the replicate samples. This might explain why exp3 replicates show qualitatively different M-V relationship from exp4 replicates. However, this would not explain the differences between exp4 replicates (there might be some lane environment effects too). One possible manifestation of these effects could be the difference in GC bias, described in section 1.4.5. Figure 4.14 shows the boxplots of raw read count ratios stratified by GC content, which are centered on the lane total ratio for exp4 P3M replicates, but not for the other 3 pairs of technical replicates. The patterns for the pairs of exp3 technical replicates are more pronounced and show opposite trend (decreasing vs. increasing pattern of count ratios as function of GC) than exp4 WT technical replicates. These plots suggest that the proper normalization ratio r for the two replicates is a function of the bin GC content. Figure 4.15 shows the GC-stratified M-V plots for the 4 pairs of technical replicates. For each pair, all genomic 5kb bins were stratified into 4 groups based on quartiles of GC content and for each group, separate M-V calculation was performed using $r = L_1/L_2$ and the points for all 4 groups were plotted together. Low mappability (< 0.9) bins were discarded for a more clear presentation. Figure 4.16 shows the same M-V plots, but using $r_g = L'_{1g}/L'_{2g}$, where L'_{1g}, L'_{2g} are GC group-specific total read counts. It is clear that the GC content effect on proper normalization seems to be the driving force behind the strange M-V relationship observed for exp3 technical replicates. Switching to GC-specific normalization ratio helps somewhat to bring M-V relationship to its expected shape but not completely, indicating the need for a finer resolution of r.

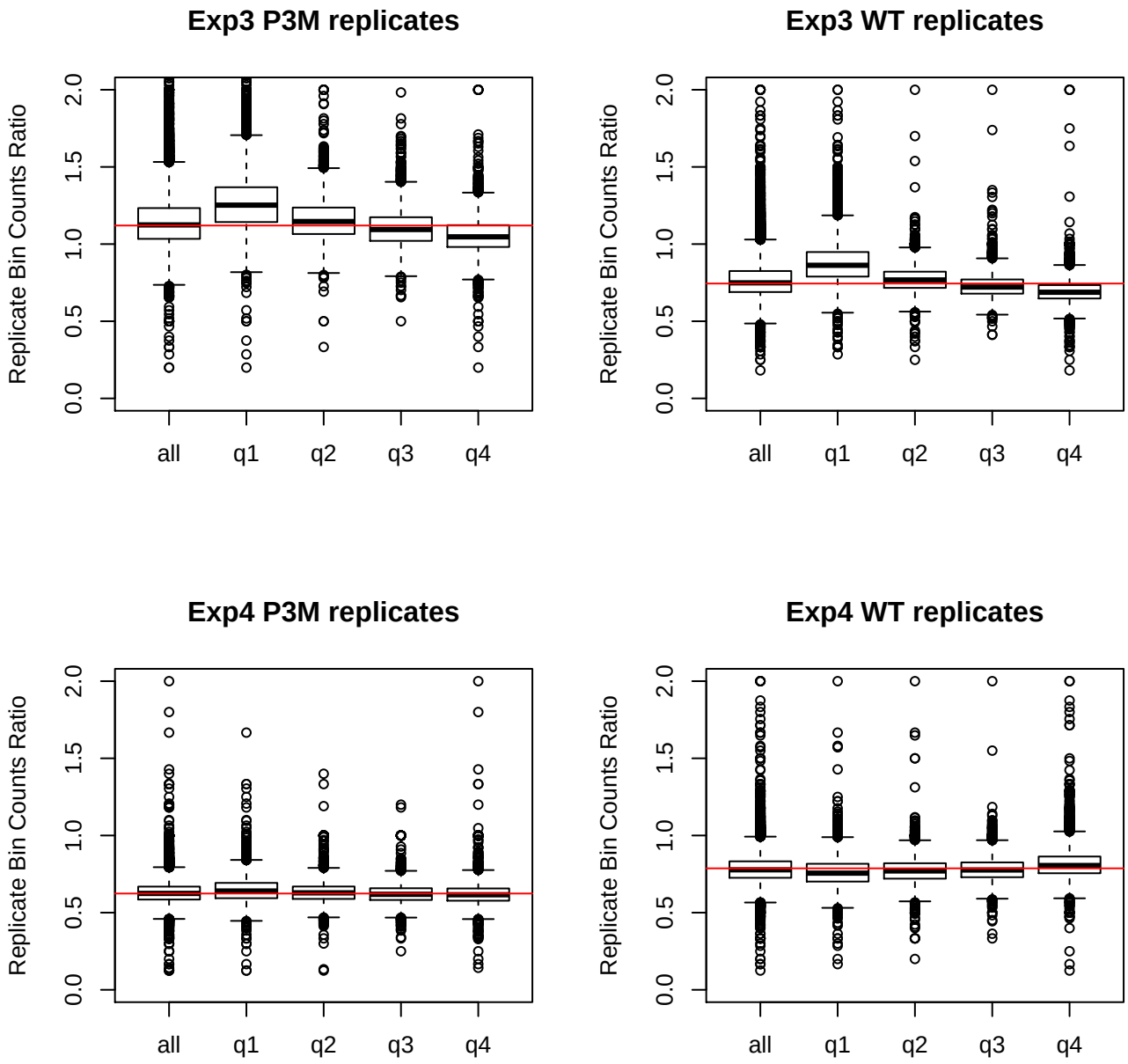

Figure 4.14: Ratios of X_{1j}/X_{2j} stratified by GC content (5kb bins). q1-4 refer to ranges based on quartiles of the distribution of GC content. Horizontal red line is the ratio of lane totals. The ratios increase as function of GC for exp4 WT technical replicates, decrease for exp3 P3M and WT technical replicates and are GC-independent for exp4 P3M technical replicates.

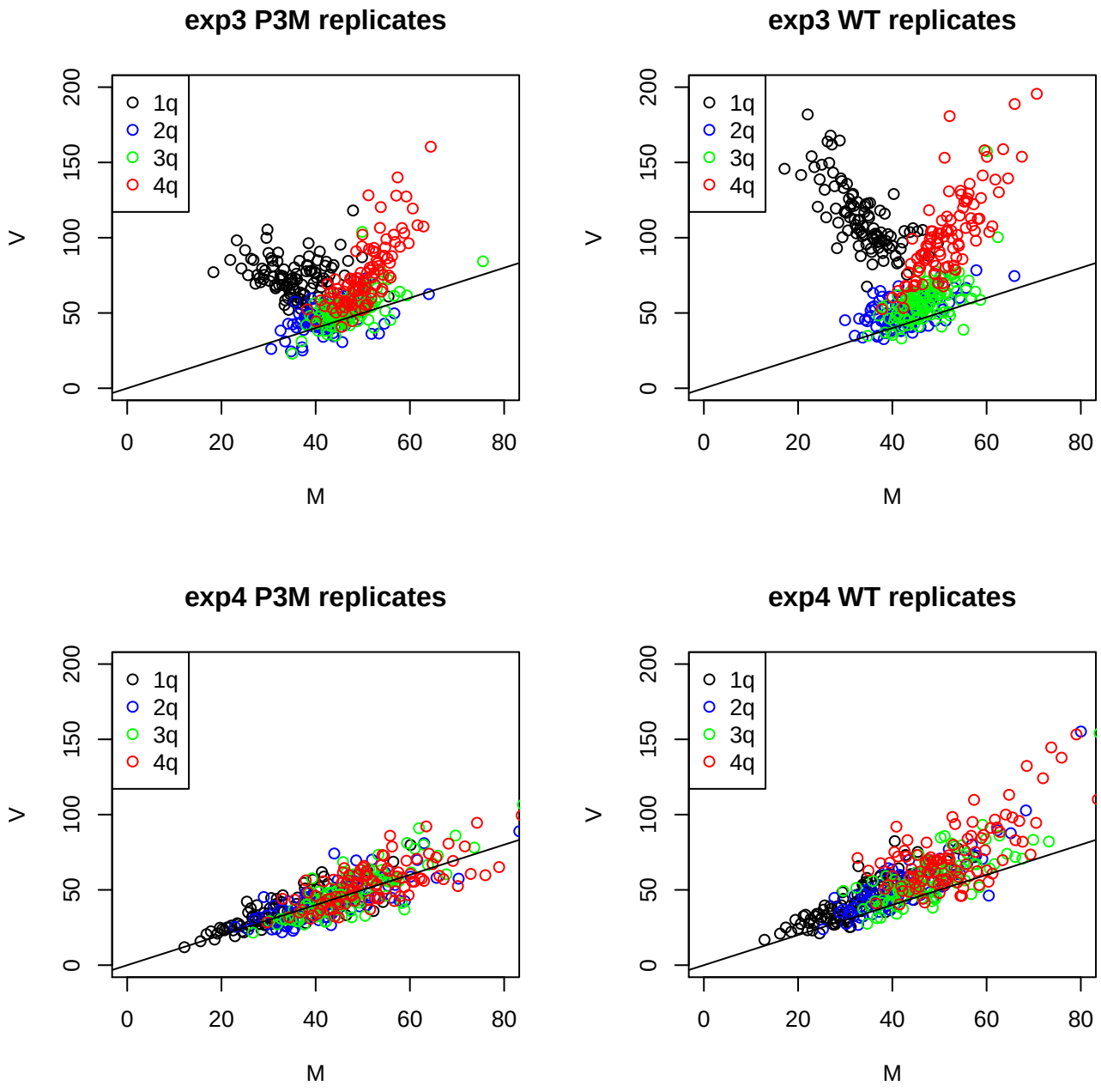

Figure 4.15: M-V plots of GC-stratified 5kb bin counts (bins with mappability < 0.9 are excluded). q1-4 refer to ranges based on quartiles of GC content. Normalization is done by overall lane totals.

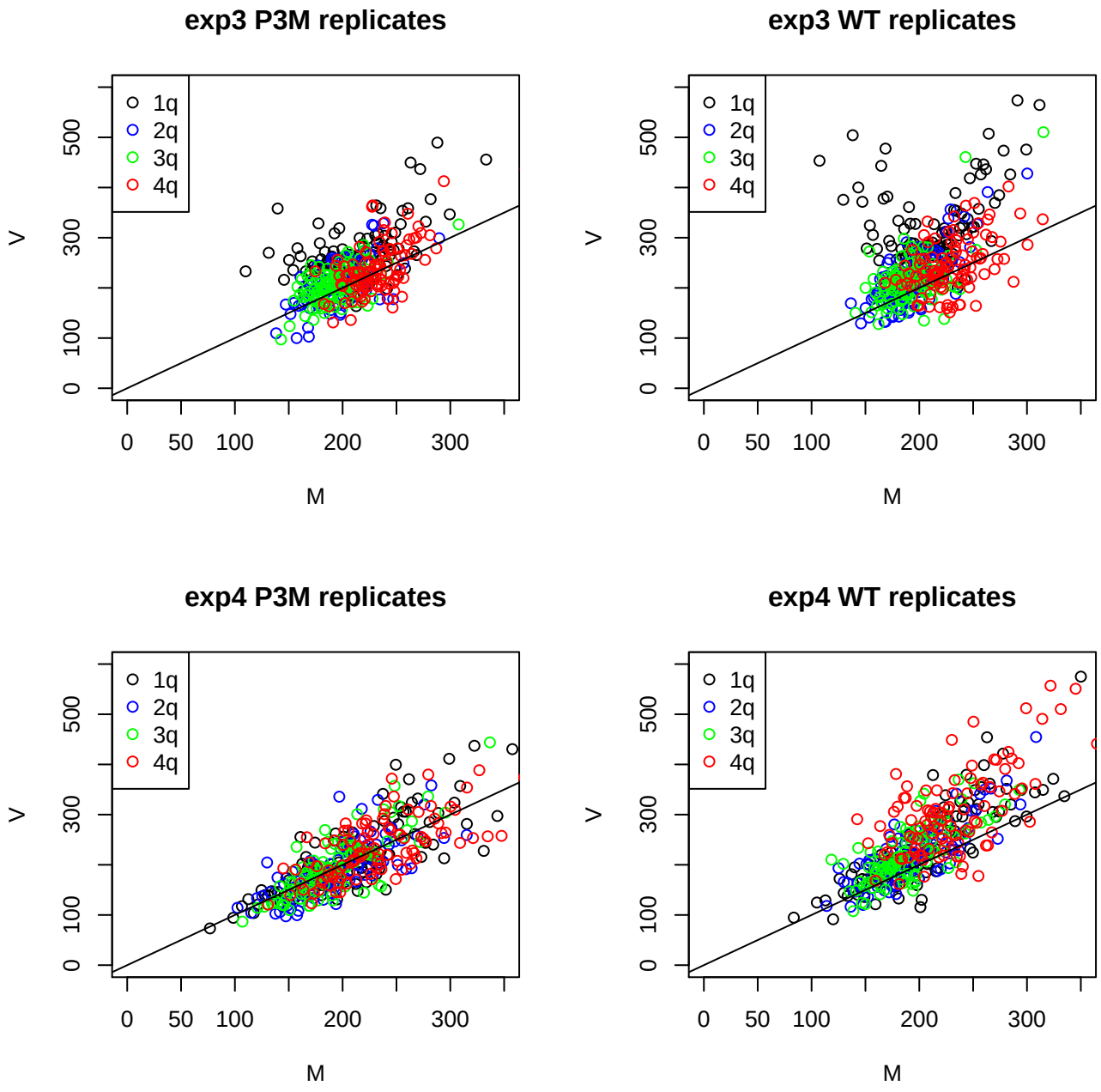

Figure 4.16: M-V plots of GC-stratified 5kb bin counts (bins with mappability < 0.9 are excluded). q1-4 refer to ranges based on quartiles of GC content. Normalization is done by GC-specific read totals.

4.4 Conclusion

In summary, it appears that sometimes the samples from the same library do not tend to behave like samples from the same fragment pool when sequenced on different flow cells. In particular, the dependence of read counts on GC might be different and give rise to the observed lack of fit behavior. We have also examined several data sets with technical replicates that were sequenced on the same flow cell and did not detect any departures from the Poisson model. This leads us to believe that there might be flow cell-specific effects that are related to the GC content of the sequenced fragments and might come about through the use of different batches of sequencing chemicals or through other means. There also seem to be some lane-specific effects as can be seen in the different relationships for the two pairs of exp4 replicates, which share the flow cell experimental structure.

Overall, we feel that the reads from technical replicates that adhere to the Poisson model can be combined without reservation, and that the diagnostic plots along the lines of what is presented here might be useful in determining whether the model holds. If the replicates show discrepancies similar to what is seen here for exp3 samples ('cubic' or U-shape M-V plots), then some caution should be used when pooling the reads. In the data presented here, the appropriate normalizing constants were GC-dependent, and thus a normalization technique adjusting for the GC content of the two replicates might be employed as a correction device. This normalization could be carried out on a bp-level basis, in which case the pooled units might be fractional reads, or on a region-level basis, in which case such a procedure would need to be direclty incorporated into a peak-finding algorithm. A recent study [43] proposes a bp-level adjustment for GC effect that seems to perform very well and might form a basis for normalization prior to combing reads from technical replicates.

The issue of combining information from biological replicates should be investigated further, especially in the light of very different results produced by the two biological replicates in table 4.1. Also, even though examination of existing data did not reveal any departures from Poisson model for technical replicates from the same flow cell, we do find that the relationships between the two exp4 P3M technical replicates and the two exp4 WT technical replicates are not the same (the WT pair shows considerably greater overdispersion) even though the flow cell structure is the same. Thus, the possible existence of lane-specific effects should be investigated as well, when multiplexed samples become available.

Concluding Remarks

In this work we present an overview of the statistical approaches to ChIP-Seq data analysis and focus on two problems in particular, those of signal identification in the absence of a control sample, and of combining data across replicate samples.

The methodology for statistical analysis of ChIP-Seq experiments is not yet mature and there remain many areas for improvement. The major thrust of methodology development has been directed towards point-like signal, e.g. transcription factor binding, and tools specific to other types of signal (histone modifications, polymerase binding) remain rare, while even for the point-like binding events, no method has emerged as unequivocally superior. We provide a conceptual framework for dissecting the statistical methodology of a peak-finding tool in terms of enrichment measures, enrichment statistics and their assumed or estimated null distributions. Using this framework, we examine a variety of existing peak-finding tools described in the literature and summarize the main approaches to different modules (e.g. normalization or statistical significance calculations) of the peak-finding procedure.

A major obstacle to conducting a proper method comparison is the lack of high-quality validation data, e.g. biological spike-in experiments or biologically verified signal and non-signal sites. Until such data becomes available, one has to rely on ad-hoc procedures to assess the performance of a peak-finder, and we attempt to deduce sets of true positive and true negative regions for data sets in this work based on motif occurrence, fold enrichment over control and reproducibility in other experiments.

It is almost universally recommended that control samples be used for proper background assessment in ChIP-Seq experiments. We observe that in some instances proper control samples are not available and propose a novel method to identifying signal sites in this one-sample scenario. We develop an approach to identifying candidate signal regions based on adjusting the observed read density for local GC and mappability content, the two features known to correlate with the background coverage. Our approach to identifying signal sites among the candidates takes advantage of our knowledge of the read density profile at the sites of punctate events of interest, and classifies the candidate regions into signal and noise based on the estimated shape parameters and the goodness-of-fit measures. The particular choice of shape we employ is driven by our observations of multiple data sets and can in practice be adjusted if a data set with a different read density profile presents itself. We show that our approach is superior to other one-sample approaches in discarding the non-

specifically enriched artifacts and often provides a superior resolution of the point binding event as well. We also show that our method's sensitivity suffers due to the existence of true signal sites with read density profile markedly different from that of most of the rest of the signal sites, and that its specificity may be affected by the existence of 'universal' peaks - genomic regions that show the read density profile characteristic of point-like transcription factor binding in multiple data sets across tissue types and transcription factors assayed.

The issue of combining information across replicate samples has not been addressed very much in the literature up to now. The usual strategies of pooling reads or intersecting peak sets are ad-hoc and lack justification, while giving vastly different results. We propose some diagnostic plots that can aid the analyst in deciding whether combining reads across technical replicate samples can be done without reservations to increase the power to detect true binding events. We show that the departures from the expected replicate behavior appear to be due to flow cell-specific sequence composition effects (GC bias) and that those need to be taken into account if the data is to be aggregated across flow cells. We do not address the important issue of combining information across biological replicates, leaving this as a subject for future work.

There are many aspects of ChIP-Seq data analysis that require a rigorous statistical treatment and the work presented here has attempted to deal with some of them. Many more are bound to present themselves as the technology develops and new sequencing platforms appear that have their own properties and biases. It is important to get a grip of these issues in order to establish a proper data analysis pipeline that would allow the highest quality results to be obtained from the ChIP-Seq data itself, and for the data to be incorporated with other types of data in the studies of regulatory cell networks. Elucidating such networks is one of the most important tasks in modern biology, and ChIP-Seq experiments are quickly becoming an indispensable part of these analyses.

Appendix A

Details of Selected Peak-Finders

A.1 A Description of Selected One-Sample Methods

ERANGE, an extended version of the peak-finder used in [1], scans the genome looking for clusters of reads of size at least N that are separated by no more than B basepairs. The minimum number of reads N is specified in units of Reads Per Million (RPM), given by formula Reads$\times 1,000,000/N_t$, thus providing a similar cutoff for samples with different numbers of total reads. The number of reads in a cluster serves as enrichment statistic T_i, and p-values are calculated based on the global Poisson model. A variety of optional filters are available that allow one to specify the minimum fraction of reads of each strandedness, as well as the minimum fraction of reads of the appropriate strandedness in the 2 directions away from the point estimate of the peak location. Additionally, ERANGE optionally retains only one copy of each duplicate alignment prior to analysis. There is also optional functionality to shift reads towards estimated binding events by either the user supplying the shift distance (1/2 of average fragment length), or estimating it from strong sites on chr1 or estimating it on a per-region basis. ERANGE reports the point estimate for the binding event and allows one to trim the reported peak regions to desired length.

FindPeaks takes a user-supplied estimate of the average fragment length and extends reads to fragments either deterministically or probabilistically, using the supplied fragment length as a center for the distribution of extensions. It then scans the genome and identifies all islands (genomic regions covered by one or more fragments) of height greater than a pre-specificed cutoff as candidate regions. The maximum height of a candidate region serves as an enrichment statistic T_i and p-values are obtained based on the global Poisson model either analytically or through simulatoins. FindPeaks optionally retains only one copy of duplicate alignments, trims peaks to a desired width, and splits peaks into sub-peaks based on the size of enrichment gap between local maxima within a peak region.

MACS employs a model-based approach to retaining duplicate alignments. The authors assume that the number of reads k mapping to any particular location follows a

$Bin(N_t, 1/G_{map})$ distribution. This model is essentially equivalent to the uniform (global Poisson) background model. Model-based p-values for various values of k are easily obtained and the pre-determined significance cutoff is used to pick maximum allowed k. MACS attempts to estimate the fragment length by studying strand-specific profiles in the most read-rich regions and taking the median of estimated profile shifts, d. The reads are then shifted towards their 3' end by $d/2$. A scanning window of the size $2d$ is employed, and enrichment statistics T_i = (number of reads in the window) are used to calculate enrichment p-values under global Poisson assumption. Windows with p-values below a specified cutoff are retained and the overlapping windows are joined. Finally, two more constant Poisson rate models are explored for each candidate region, based on two intervals of user-specified legnth (e.g. 2kb and 10kb) surrounding the candidate region, and the largest of the 3 resulting p-values is assigned to the region. These 2 additional tests are designed to guard against non-specific enrichment, e.g. open chromatin regions. The user-supplied p-value cutoff is used to identify the final set of called peaks and the point estimate of the binding event is obtained by smoothing the read density and looking for maxima points.

USeq optionally retains only 1 copy of duplicate alignments and estimates the read shift based on either a composite peak created from top read-rich regions or by taking median of estimated shift for such regions. The reads are then shifted to their 3' end. It then uses a sliding window of user supplied length to scan the genome, and calculates enrichment p-value based on global Poisson assumption, with read counts in windows serving as enrichment statistics T_i. There are a lot of other parameters a user can specify and exploit, for example a list of pre-compiled false positive regions (e.g. centromeres or telomeres) to exclude from analysis.

SISSRS estimates average fragment length F from distances between the closest reads of opposite strandedness, or uses a user-supplied value and then shifts the reads towards their 3' end by $F/2$. It then scans the genome with a sliding window looking for regions of transition from high positive-strand read density to high negative-strand read density, and those are declared to be candidate binding sites. There are adjustable constraints that such sites must satisfy, e.g., minimum number of reads of each strandedness F bps upstream and downstream. The combined number of reads of both strandedness in the $2F$ bp region centered on putative binding site serves as enrichment statistic T_i and p-values are calculated based on global Poisson model.

CisGenome uses a global Negative Binomial model instead of Poisson to account for excessive heterogeneity of background. It estimates the parameters for the background NB distribution from windows with 2 or fewer reads (which are all assumed to come from the background) and then uses a sliding window to detect enriched regions along the genome. The number of reads $n_{i,t}$ in the window serves as enrichment statistic T_i. For each possible number of reads n in the window, an estimate of the FDR is obtained by letting FDR=E_n/O_n, where E_n is the number of expected windows with n or more reads under the NB model and O_n is the number of observed windows with n or more reads. A user-supplied FDR cutoff is chosen to pick the cutoff value of n, and windows with more than n reads are identified as

binding sites. Strand-specific enrichment profiles are built, shift size is calculated from these profiles, reads are shifted towards their 3' end, and the entire preceding process is repeated to obtain greater resolution of binding sites.

BayesPeak is an HMM-based method. It uses a sliding window approach and records numbers of reads on each strand in the windows. The size of the window is chosen to be approximately 1/2 of fragment length, so there is dependence between the numbers of positive-strand reads in one window and negative-strand reads in the one directly downstream. The hidden states correspond to presence or absence of a signal in the window, and due to the dependence between consecutive windows, a 4-state model is employed. The method assumes a negative binomial distribution for observed read counts in absence of signal and a mixture of two negative binomials for windows with signal, the two components representing signal and background. The model is fit using an MCMC framework and the regions with posterior probability of enrichment > 0.5 are identified as binding events.

MOSAiCS extends reads to the expected fragment size and then records the total numbers of fragments overlapping sliding windows of pre-specified size. It also calculates the mappability and GC scores for the windows, which are essentially window-specific averages. They model the counts in background bins as having negative binomial distribution with a mean that is a function of GC and mappability content. For the signal-containing bins the counts are modeled as a mixture of two distributions: a background NB and a signal component which itself can be either a NB distribution or a mixture of 2 NB distributions. Thus, for each bin, the distribution of counts is a mixture of the background distribution and the enrichment distribution (which is itself a mixture), with mixing proportions given by unobserved fractions of non-enriched and enriched windows, respectively. They fit their model using assumption that any bins with < 3 reads do not contain any signal, and use the direct posterior probability approach to control FDR.

A.2 A Description of Selected Two-Sample Methods

ERANGE normalizes treatment and control samples by converting read counts to RPM values (lane total normalization) as described above. The user is asked to specify the minimum fold enrichment value that candidate regions must satisfy. For each candidate peak, a Poisson p-value is calculated based on normalized reads in the control sample, and a single FDR is calculated based on library swapping approach and the number of total peaks identified in the swapped comparison.

FindPeaks normalizes treatment and control by using its 1-sample analysis module to call peaks in the two samples and then identifying the sets of common peaks. The heights of these are plotted against each other and a line minimizing the perpendicular distances to the plotted points is fitted. The distances from all peaks (including non-paired ones) to the line are calculated, and a Gaussian distribution is fit to the distances (with points above the line having negative distances and points below positive). A specified p-value is then

used as a cutoff to find points significantly distant from the line, which form the final set of sample-specific peaks. For non-paired peaks, the second coordinate during the plotting is given by the maximum height of the corresponding region in the control. Alternatively, normalization is performed by identifying the common peaks in the 2 samples and fitting the line with intercept=0 and slope given by the ratio of reads in the common peaks. In this alternative approach, hyperbolic curves are fit to the peak heights, producing a set of peaks enriched in treatment over control and another set of peaks enriched in control over treatment. A pre-specified FDR cutoff is then used to pick the set of curves that result in the desired ratio of control peaks to treatment peaks. Optional refinements are available, similar to the one-sample module of this peak-finder.

MACS normalizes treatment and control by lane totals. The reads are shifted and duplicate alignments reduced in a model-based fashion similar to one-sample module. The candidate regions are first identified by calculating the p-value for the observed read counts in sliding window under global Poisson model. The local Poisson model is then employed to calculate a p-value for each candidate based on the observed read count rate in the larger surrounding region in the control sample (cf. using a larger surrounding region in the treatment sample itself in the one-sample module of this peak-finder). The regions with p-values below a pre-defined cutoff are retained. The library swap is then employed to obtain an empirical FDR estimate for each p-value.

USeq normalizes treatment and control by subsampling the larger of the two samples to ensure they have the same total number of reads. The treatment of duplicate alignments and read shift is same as for one-sample module. A sliding window approach is taken and letting $n_{i,t}, n_{i,c}$ denote the number of reads in window i in treatment and control, respectively. The distribution of $n_{i,t}$ is modeled as $\text{Bin}(n_{i,t} + n_{i,c}, 0.5)$. This produces window-specific p-values that can be used as measure of statistical significance. If only regions enriched in treatment relative to control (and not vice versa) are of interest, then an empirical FDR is also calculated for p-value cutoffs by running an additional control sample against the original control. In practice, a single control sample and a single treatment sample are subsampled to result in 2 control samples and a treatment sample, all of the same size.

QuEST uses a kernel-density smoothing approach to build a continous enrichment profile along the genome (a variation on count-based enrichment measure). It does so in a strand-specific fashion and then the top enriched (both in absolute terms and relative to a control sample) regions are used to calculate the optimal profile shift distance by looking at correlations between shifted strand-specific profiles. The genome-wide strand-specific profiles are then shifted towards each other and their contributions are combined. Candidate regions are called by identifying local maxima in the combined profile that exceed pre-defined cutoffs for absolute enrichment relative to treatment sample's background and relative enrichment to the control sample. A variety of statistical significance measures are calculated, among them region-specific p-values based on a local Poisson model (similar to MACS), and peak-specific p-values based on enrichment profile in the control sample, both adjusted for multiple hypothesis testing. The distinction arises from the possibility of more

than 1 peak in peak-containg region. A second control sample of same size as the original treatment sample is used to obtain an empirical FDR estimate, when available (in practice a single control sample is split, if large enough). QuEST employs a sophisticated formula to decide when to reduce duplicate alignments to a single copy and when to keep all copies, based on pre-defined cutoffs that seek to guard against PCR artifacts but to preserve natural duplication due to signal.

SISSRS uses a similar approach to identifying candidate regions as its one-sample module. The control sample is incorporated into peak-finding by looking for the number of peaks identified in the control sample (using a one-sample approach) at a given enrichment cutoff. This number must be less than a user-specified value. The control is further employed to calculate fold enrichment for each candidate region and the null distribution of fold enrichment values is obtained by randomly sampling windows along the genome. The fold enrichment exceeding a pre-specified p-value cutoff is then required for the region to be declared a binding event.

CisGenome uses windows with low read counts to estimate the normalization ratio p between control sample and the background portion of the treatment sample. Defining $n_{i,t}, n_{i,c}$ to be the number of treatment and control reads in window i, respectively, the distribution of $n_{i,t}$ is modeled as $\text{Bin}(n_{i,t} + n_{i,c}, p)$. The rest of the method is similar to the one-sample module, with FDR calculated for each value of $(n_{i,t}, n_{i,c})$ providing a measure of statistical significance. The total number of reads $n_{i,t} + n_{i,c}$ also has to exceed cutoff based on FDR from one-sample approach.

PeakSeq reduces the number of allowed duplicate alignments to some pre-specified number and extends reads to the user-supplied fragment length, with an enrichment measure provided by the number of fragments overlapping a genomic location. The method proceeds to divide genome into 1Mb segments and the random uniform background model is assumed for the non-mappable portion of each segment. Enrichment profiles are simulated under this model for each segment, where the mappable portion is modeled as a contiguous set of bps, and the observed numbers of local maxima exceeding various cutoffs are compared for the real and simulated data. A segment-specific threshold is chosen based on an FDR cutoff and the candidate binding sites are identified in each segment. The treatment and control samples are then normalized to each other by a linear regression of read counts in 10bk bins, after excluding a pre-specified portion of candidate binding sites from the data. The linear regression fit is done on a per-chromosome basis. After normalization,the distribution of read counts $n_{i,t}$ is assumed to be $\text{Bin}(n_{i,t} + n_{i,c}, 0.5)$ using our earlier notation, but with $n_{i,c}$ being normalized control counts in the region, rounded to nearest integer (similar to USeq approach). Per-region p-values are obtained and adjusted for multiple hypothesis testing.

GLITR reduces duplicate alignments to a single copy and extends reads to user-specified fragment length. It seems to require an inordinately large control sample size relative to treatment (at least 4X). It obtains a pseudo-ChIP sample by sampling N_t reads from a control sample, where N_t is the treatment read total. The rest of the control reads are used for background estimation. Contiguous regions of the genome with overlap enrichment

exceeding some cutoff are selected as candidate binding sites in both treatment and pseudo-ChIP samples. The remaining control sample is further sampled to yield two more samples, each of size equal to original treatment sample, one used to provide background for the treatment sample and one for pseudo-ChIP. The fold enrichment (defined as ratio of average overlap enrichment measures in a region) of treatment and pseudo-CHIP candidate regions to their background samples is calculated and the median value is taken over several rounds of background subsampling. Each candidate region in both the treatment and pseudo-ChIP samples now has 2 scores associated with it: height and median FE. The scores are normalized and a k-nearest neighbor approach is used to calculate the significance cutoff. For each candidate in both samples, the candidate is declared to be a binding site if more than n of its k neighbors in the 2-D score space come from the treatment sample. For each n, an empirical FDR is calculated as C_n/T_n, where C_n, T_n are numbers of control and treatment candidates, respectively, classified as binding sites at cutoff n, and a specified FDR cutoff is chosen to decide on n.

SPP normalizes treatment and control samples by excluding any highly enriched regions (1kb windows with read counts in excess of some pre-specified p-value under a global Poisson model) from both treatment and cotnrol, and by taking the ratio of the remaining reads. The control read counts are then normalized by this ratio and are subtracted from window-based counts of reads in the treatment sample as described further. An enrichment score is calculated at each genomic location and is given by $S = 2\sqrt{p_U n_D} - (p_D + n_U)$, where p_U is the number of positive strand reads (after background subtraction) in a window of some pre-defined size upstream of the location of interest, and the other quantities are defined similarly (with n denoting negative strand reads and D denoting downstream window). Positions of the local maxima of the score profiles are candidate binding sites. The empirical FDR estimate for each score cutoff is obtained by library swapping, or alternatively, can be obtained from simulations where reads or clusters of reads are sampled and placed along the genome. There is also another scoring algorithm implemented based on the score defined above which is omitted here.

T-PIC extends reads to the user-supplied fragment length and calculates overlap enrichment (coverage) for each bp in the genome. The authors develop a tree statistic that has high values for regions showing departures from a homogeneous Poisson fragment start site process. The method breaks up the genome into smaller pieces using the control sample as follows: for each genomic location it calculates local per-bp read start rate (not overlap) in a window of 1000bp centered on the location in the control sample, producing a location-specific local rate. Then, starting with some segment of length K at the start of the chromosome, the average location-specific local rate is calculated across locations in the segment. Downstream nucleotides are added then one by one to the growing segment until a location is encountered with local rate that is different from current average rate by a specified amount. At this point a new segment starts with a size of at least K, and the segmentation proceeds until the entire genome is partitioned. For each segment, the average per-bp rate is rounded to nearest integer and the segments with same rounded average rate

are grouped into regions (consisting of disjoint segments) with same underlying background rate. The candidate binding sites are obtained as contigous stretches of genome with coverage exceeding pre-specified cutoff in the treatment sample. A null distribution for the tree statistic is then obtained through simulations for locations in each candidate binding site, under the assumption of constant Poisson fragment start site process with region-specific rate, and the p-values are obtained and corrected for multiple hypothesis testing.

BayesPeak's two-sample approach is virtually identical to its one-sample module, with the window-specific control read counts incorporated into the window-specific negative binomial model of read counts.

MOSAiCS two-sample approach is also essentially the same as its one-sample approach, with control window counts incorporated into the window-specific negative binomial model, alongside the mappability and GC content. More specifically, depending on the number of reads in the window in the control sample, the negative binomial distribution has the mean that is a function of either just the control reads themselves (when number of reads is large) or the control reads, mappability and GC content (when the number of of reads is small).

CCAT shifts reads by 1/2 of the estimated fragment length and then subsamples treatment or control sample so that the the total number of reads in control sample is equal to the fraction of treatment reads that come from background. This fraction is estimated iteratively by looking for 1kb bins along the genome that have $n^+_{i,t} < \hat{\alpha} N_t n^+_{i,c}/N_c$, where $n^+_{i,t}, n^+_{i,c}$ are bin-specific counts of positive strand reads in treatment and control, respectively, N_t, N_c are treatment and control sample totals and α is the current estimate of fraction of BG reads. The bins satisfying this inequality are declared to not contain signal. Then new estimate $\hat{\alpha}$ is obtained as $\hat{\alpha} = N^-_{n,t} N_c/(N^-_{n,c} N_t)$, where $N^-_{n,t}, N^-_{n,c}$ are total negative strand read counts in the non-signal containing bins in treatment and contorl samples, respectively. The starting estimate $\hat{\alpha} = 1$ is used and the procedure is said to converge rapidly. After subsampling, a sliding window is used to identify regions in treatment sample with 2-fold or greater enrichment over control and empirical FDR is obtained by library swapping. Other statistical-significance measures are produced as well, including p-values based on a Binomial model (similar to PeakSeq) and on a local Poisson model (similar to MACS).

PICS identifies candidate regions using a sliding window and retaining candidates with a minimum number of positive strand reads in left half-window and negative strand reads in right half-window. It then models the positive strand reads positions in the candidate region i as $f_i \sim \sum_{k=1}^{K} w_k t_4(\mu - \delta_k/2, \sigma^2_{fk})$ and the negative strand reads positions as $r_i \sim \sum_{k=1}^{K} w_k t_4(\mu + \delta_k/2), \sigma^2_{rk})$, with K mixture components corresponding to K binding events (peaks) in a region (K varies between candidate regions and will be 1 for single binding events). Here t_4 denotes the t distribution with 4 degrees of freedom, μ is the point location of the binding event, δ is the separation between strand enrichment profiles (assumed to be equal to average fragment length for this binding event) and $\sigma^2_{fk}, \sigma^2_{rk}$ are strand-specific

variances of the read position distributions. Priors are assumed for $\delta, \sigma_f^2, \sigma_r^2$ parameters and the missing data approach is taken with respect to non-mappable bases. BIC is used to estimate number of binding events K for each region. After model fitting, some peaks are merged, standard errors are obtained for parameter estimates and the regions with reasonable values (as defined by pre-specified cutoffs) of parameter estimates and small standard errors are retained. Fold enrichment of read counts in treatment relative to control (adjusted for lane totals) is used as an enrichment score for candidate regions and empirical FDR is obtained by library swap.

Sole-Search attempts to adjust for genomic copy number variations by identifying long regions that are either devoid of reads or substantially enriched relative to average genome read coverage. The identified duplications relative to reference genome are flagged and the counts in windows coming from these regions are later adjusted by the fold enrichment over the genomic average. They use sliding window to identify all 30bp windows that contain reads, and produce a simulation-based FDR for peak height cutoff by randomly assigning reads to the collection of these 30bp windows, thus trying to account for local mappability effects. Regions passing this FDR cutoff are considered candidates, the reads from candidate regions are removed and the process is repeated to obtain a better FDR estimate. Finally, a control sample with same number of reads as the treatment is used and the reads in slidng windows are compared between the two, using a t-test.

CSAR extends reads to user-supplied fragment length and builds overlap enrichment profile along the genome for both treatment and control samples (after further normalizing by lane totals). The distribution of treatment enrichment measures is normalized with respect to the mean and variance of control enrichment measures. For each genomic location i the distribution of enrichment measure is assumed to be Poisson with rate given by $\max(\lambda_G, \lambda_{i,c})$, where λ_G is the rate under the global Poisson model and $\lambda_{i,c}$ is the control enrichment score at location i. A p-value is calculated at each bp and locations within some window size of each other are merged to provide candidate peaks. Each peak is then assigned a score which is the smallest p-value for bps in that peak. FDR estimates for each p-value cutoff are obtained by several rounds of simulations, where the control and treatment reads are merged together, a pseudo-treatment sample is sampled from this set and the analysis is repeated.

Bibliography

[1] D. Johnson, A. Mortazavi, R. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," *Science*, vol. 316, no. 5830, p. 1497, 2007.

[2] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, *et al.*, "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing," *Nature methods*, vol. 4, no. 8, pp. 651–657, 2007.

[3] A. Barski and K. Zhao, "Genomic location analysis by ChIP-Seq," *Journal of cellular biochemistry*, vol. 107, no. 1, pp. 11–18, 2009.

[4] Y. Zhang, T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, B. Bernstein, C. Nussbaum, R. Myers, M. Brown, W. Li, *et al.*, "Model-based analysis of ChIP-Seq (MACS)," *Genome biology*, vol. 9, no. 9, p. R137, 2008.

[5] H. Ji, H. Jiang, W. Ma, D. Johnson, R. Myers, and W. Wong, "An integrated software system for analyzing ChIP-chip and ChIP-seq data," *Nature Biotechnology*, vol. 26, no. 11, pp. 1293–1300, 2008.

[6] P. Kharchenko, M. Tolstorukov, and P. Park, "Design and analysis of ChIP-seq experiments for DNA-binding proteins," *Nature biotechnology*, vol. 26, no. 12, pp. 1351–1359, 2008.

[7] J. Rozowsky, G. Euskirchen, R. Auerbach, Z. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. Gerstein, "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls," *Nature biotechnology*, vol. 27, no. 1, pp. 66–75, 2009.

[8] B. Hoffman and S. Jones, "Genome-wide identification of DNA-protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing," *Journal of Endocrinology*, vol. 201, no. 1, p. 1, 2009.

[9] C. Zang, D. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng, "A clustering approach for identification of enriched domains from histone modification ChIP-Seq data," *Bioinformatics*, vol. 25, no. 15, p. 1952, 2009.

[10] K. Blahnik, L. Dou, H. O'Geen, T. McPhillips, X. Xu, A. Cao, S. Iyengar, C. Nicolet, B. Ludascher, I. Korf, *et al.*, "Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data," *Nucleic acids research*, vol. 38, no. 3, p. e13, 2010.

[11] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome research*, vol. 18, no. 11, p. 1851, 2008.

[12] R. Li, C. Yu, Y. Li, T. Lam, S. Yiu, K. Kristiansen, and J. Wang, "SOAP2: an improved ultrafast tool for short read alignment," *Bioinformatics*, vol. 25, no. 15, p. 1966, 2009.

[13] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," *Bioinformatics*, vol. 25, no. 14, p. 1754, 2009.

[14] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol*, vol. 10, no. 3, p. R25, 2009.

[15] A. Boyle, S. Davis, H. Shulha, P. Meltzer, E. Margulies, Z. Weng, T. Furey, and G. Crawford, "High-resolution mapping and characterization of open chromatin across the genome," *Cell*, vol. 132, no. 2, pp. 311–322, 2008.

[16] D. Nix, S. Courdy, and K. Boucher, "Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks," *BMC bioinformatics*, vol. 9, no. 1, p. 523, 2008.

[17] V. Vega, E. Cheung, N. Palanisamy, and W. Sung, "Inherent signals in sequencing-based chromatin-immunoprecipitation control libraries," *PLoS One*, vol. 4, no. 4, p. 5241, 2009.

[18] L. Teytelman, B. Ozaydin, O. Zill, P. Lefrançois, M. Snyder, J. Rine, and M. Eisen, "Impact of chromatin structures on DNA processing for genomic analyses," *PLoS One*, vol. 4, no. 8, p. e6700, 2009.

[19] M. Quail, I. Kozarewa, F. Smith, A. Scally, P. Stephens, R. Durbin, H. Swerdlow, and D. Turner, "A large genome center's improvements to the Illumina sequencing system," *Nature methods*, vol. 5, no. 12, pp. 1005–1010, 2008.

[20] S. Pepke, B. Wold, and A. Mortazavi, "Computation for ChIP-seq and RNA-seq studies," *Nature methods*, vol. 6, pp. S22–S32, 2009.

[21] G. Tuteja, P. White, J. Schug, and K. Kaestner, "Extracting transcription factor targets from ChIP-Seq data," *Nucleic acids research*, vol. 37, no. 17, p. e113, 2009.

[22] P. Kuan, G. Pan, J. Thomson, and K. Stewart Ra, "A statistical framework for the analysis of ChIP-Seq data," tech. rep., University of Wisconsin, Department of Statistics, 2009.

[23] J. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing," *Nucleic acids research*, 2008.

[24] I. Kozarewa, Z. Ning, M. Quail, M. Sanders, M. Berriman, and D. Turner, "Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+ C)-biased genomes," *Nature methods*, vol. 6, no. 4, pp. 291–295, 2009.

[25] Z. Zhang, J. Rozowsky, M. Snyder, J. Chang, M. Gerstein, *et al.*, "Modeling ChIP sequencing in silico with applications," *PLoS Comput Biol*, vol. 4, no. 8, p. e1000158, 2008.

[26] A. Fejes, G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, and S. Jones, "FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology," *Bioinformatics*, vol. 24, no. 15, p. 1729, 2008.

[27] R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao, "Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data," *Nucleic acids research*, vol. 36, no. 16, p. 5221, 2008.

[28] C. Spyrou, R. Stark, A. Lynch, and S. Tavaré, "BayesPeak: Bayesian analysis of ChIP-seq data," *BMC bioinformatics*, vol. 10, no. 1, p. 299, 2009.

[29] A. Valouev, D. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. Myers, and A. Sidow, "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data," *Nature methods*, vol. 5, no. 9, pp. 829–834, 2008.

[30] V. Hower, S. Evans, and L. Pachter, "Shape-based peak identification for ChIP-Seq," *BMC bioinformatics*, vol. 12, p. 15, 2011.

[31] H. Xu, L. Handoko, X. Wei, C. Ye, J. Sheng, C. Wei, F. Lin, and W. Sung, "A signal–noise model for significance analysis of ChIP-seq with negative control," *Bioinformatics*, vol. 26, no. 9, p. 1199, 2010.

[32] X. Zhang, G. Robertson, M. Krzywinski, K. Ning, A. Droit, S. Jones, and R. Gottardo, "PICS: Probabilistic Inference for ChIP-seq," *Biometrics*, 2010.

[33] K. Kaufmann, J. Muiño, R. Jauregui, C. Airoldi, C. Smaczniak, P. Krajewski, and G. Angenent, "Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower," *PLoS Biol*, vol. 7, no. 4, p. e1000090, 2009.

[34] H. Xu, C. Wei, F. Lin, and W. Sung, "An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data," *Bioinformatics*, vol. 24, no. 20, p. 2344, 2008.

[35] W. Feng, Y. Liu, J. Wu, K. Nephew, T. Huang, and L. Li, "A Poisson mixture model to identify changes in RNA polymerase II binding quantity using high-throughput sequencing technology," *BMC genomics*, vol. 9, no. Suppl 2, p. S23, 2008.

[36] A. Barski, S. Cuddapah, K. Cui, T. Roh, D. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, "High-resolution profiling of histone methylations in the human genome," *Cell*, vol. 129, no. 4, pp. 823–837, 2007.

[37] T. Laajala, S. Raghav, S. Tuomela, R. Lahesmaa, T. Aittokallio, and L. Elo, "A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments," *BMC genomics*, vol. 10, no. 1, p. 618, 2009.

[38] E. Wilbanks and M. Facciotti, "Evaluation of algorithm performance in ChIP-seq peak detection," *PloS one*, vol. 5, no. 7, p. e11471, 2010.

[39] O. Vivar, X. Zhao, E. Saunier, C. Griffin, O. Mayba, M. Tagliaferri, I. Cohen, T. Speed, and D. Leitman, "Estrogen receptor β binds to and regulates three distinct classes of target genes," *Journal of Biological Chemistry*, vol. 285, no. 29, p. 22059, 2010.

[40] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[41] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, p. R106, 2010.

[42] J. Bullard, E. Purdom, K. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC bioinformatics*, vol. 11, no. 1, p. 94, 2010.

[43] Y. Benjamini and T. Speed, "Estimation and correction for GC-content bias in high throughput sequencing," tech. rep., University of California - Berkeley Department of Statistics, 2011.