Chromatin immunoprecipitation and high-density tiling microarrays: a generative model, methods for analysis, and methodology assessment in the absence of a "gold standard"

by

Richard Walter Bourgon

B.A. (Brown University, Providence) 1992
B.S. (Brown University, Providence) 1992
M.A. (University of California, Berkeley) 2003

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

in

Statistics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Terence P. Speed, Chair
Professor Michael B. Eisen
Professor Mark van der Laan

Fall 2006

Chromatin immunoprecipitation and high-density tiling microarrays: a generative model, methods for analysis, and methodology assessment in the absence of a “gold standard”

Abstract

Chromatin immunoprecipitation and high-density tiling microarrays:
a generative model, methods for analysis, and methodology
assessment in the absence of a "gold standard"

by

Richard Walter Bourgon
Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Terence P. Speed, Chair

The combination of chromatin immunoprecipitation (ChIP) and high-density tiling microarrays permits precise localization of protein-DNA interaction—sites of transcription factor binding, for example, or regions exhibiting chromatin modifications associated with the regulation of gene expression. Early ChIP-chip studies used low density spotted arrays, for which the assumption of statistical independence across features was justified; modern, *in situ* synthesized arrays, on the other hand, achieve such dense coverage that this assumption is no longer justified. In this document we explore the nature of the dependence which arises, and present a generative model for the data. This model predicts behavior of probe-level statistics which is shown to be largely consistent with observation; it also provides a basis for estimating covariance among the test statistics associated with neighboring genomic positions and thereby correctly assessing the statistical significance of apparent enrichment.

Assessing the effectiveness of the proposed procedure relative to existing alternatives is still challenging: for most transcription factors, for example, only a small number of real binding sites are known, and even these are not likely to all be biologically active in a given sample. Performance assessments based on simulated or artificial data, while useful, are unsatisfying and leave generalizability in question. To address this issue, we build on the existing literature for sensitivity and specificity estimation in the absence of "gold standard" test set data, and propose a new variant on receiver operating characteristic (ROC) analysis. The relationship between true ROC curves and the proposed "pseudo-ROC" curves is described, and sufficient conditions under which the latter lead researchers to select the correct procedure as superior are given. While informal application of the intuition underlying pseudo-ROC comparisons is common in the computational biology literature, authors have rarely asked why and, more importantly, when such comparisons are valid; here we provide a clear framework for addressing these questions.

Finally, ROC and pseudo-ROC curves, based on both artificial spike-in and real ChIP-chip experimental data, are used to compare the performance of the enrichment detection method detailed here to that of other, recently proposed methods. Particular attention is given to the behavior of probe-level statistics—an issue which has already been carefully explored in the literature on analysis of gene expression microarray data—and to how different methods approach this behavior. Our enrichment detection method is shown to perform as well as, or better than, more complicated methods. This strong performance, coupled with the fact that it alone correctly incorporates correlation when assessing statistical significance, argues for its general application.

Professor Terence P. Speed
Dissertation Committee Chair

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank the Statistics Department at UCB for their outstanding instruction and guidance. In particular, I would like to thank Terry Speed: his unflagging energy, critical yet constructive eye, vast store of statistical knowledge, and insistence that anything worth doing is worth doing right have set a very high standard, indeed, for how our work *ought* to be done. I hope to measure up...

Affymetrix, Inc. and numerous individuals working there—Tom Gingeras, Antonio Piccolboni, Stefan Bekiranov, Srinka Ghosh, and David Nix—have been instrumental is shaping much of this document's contents. I especially wish to thank Simon Cawley for his mentorship, and for providing internship opportunities.

Members of the Berkeley *Drosophila* Transcription Network Project—in particular, Mark Biggin, Mike Eisen, Xiaoyong Li, and Stewart MacArthur—have been very generous with their experimental data and their wealth of experience with ChIP-chip.

Finally, I would like to thank my parents for their unconditional support for the educational process. (I can promise you that this is the end of it, though!) Y, la última en la lista pero la primera en mi corazón: un fuerte abrazo pa' mi querida Julie, que tuvo que aguantar tantas desveledas durante esta larga carrera. ¡No se hubiera podido sin ti!

Chapter 1

Introduction

1.1 Chromatin immunoprecipitation and microarrays

In the six years since the first major applications in yeast (Ren et al., 2000), the combination of chromatin immunoprecipitation (ChIP) and microarrays has flourished. While traditional microarray experiments seek to quantify the level of expression for a large set of genes simultaneously, "ChIP-chip," as this newer procedure has been dubbed in the literature, focuses on one protein at a time, and seeks to identify the locations at which this protein interacts with the DNA. The typical proteins of interest thus far have been transcription factors (DNA-binding proteins involved in the regulation of gene expression) and histones exhibiting one of several possible modifications associated with epigenetic regulation of gene expression. These choices are natural: the cells within a multicellular organism exhibit a wide range of morphological and functional characteristics in spite of the fact that they all contain essentially the same DNA, and thus the same genes; the same can be said of a single cell, or a single-celled organism, over the course of its life. What varies from cell type to cell type (or stage to stage), though, is the way in which these genes are switched on and off, or up and down. ChIP-chip provides a means of studying some of the core mechanisms of this regulation, in action in living cells and on a scale that had not previously been possible.

The ChIP-chip assay consists of two parts.[1] First, chromatin immunoprecipitation is used to select for fragments of the genome which are in more-or-less direct contact, *in vivo*, with the protein of interest—by binding them to it and then pulling down the assembly via an antibody with specific affinity for the protein. Chromatin immunoprecipitation was then traditionally followed by PCR—to check for the presence

[1] Here we give only a sketch; a more formal description is provided in Chapter 2.

of specific suspected sequences among the precipitated fragments—or with cloning and sequencing—to identify novel regions of the genome (Weinmann and Farnham, 2002). The former, however, only permits one to check for sites which are already known; the latter is time consuming, expensive, and does not scale well.

Microarrays provide a more powerful and cost effective alternative for mapping ChIP-enriched fragments back to genomic coordinates. Microarrays contain large numbers of single-stranded DNA probes with sequence derived from known genomic positions, each of which responds in a way that is roughly proportional to the amount complementary target sequence found in the ChIP precipitate. Like PCR, microarrays can only report on fragments containing known, pre-selected sequence; fragments with no complementary probes on the array will be overlooked (and probes whose complementary sequence appears in multiple locations throughout the genome will provide ambiguous information, at best, about the source of their target fragments). This limitation, however, has become less and less restrictive as the technology has matured. Modern arrays interrogate a very large fraction of the genome, even in higher organisms, and have already lead to the discovery of large numbers of novel regions of protein-DNA association.

The range of applications to which ChIP-chip has been brought to bear has grown rapidly. Specifically, the method has been used for *in vivo* localization of transcription factors (e.g., Cawley et al., 2004; Harbison et al., 2004; Odom et al., 2004; Carroll et al., 2005; Lee et al., 2006; Schwartz et al., 2006), the Pol II and Pol III transcriptional machinery (Moqtaderi and Struhl, 2004; Odom et al., 2004; Brodsky et al., 2005; Kim et al., 2005; Lee et al., 2006), nucleosomes (Yuan et al., 2005), post-transcriptionally modified histone proteins (Bernstein et al., 2005; Pokholok et al., 2005; Lee et al., 2006; Schwartz et al., 2006), and origin recognition complexes (MacAlpine et al., 2004), among others.

The array platforms used for ChIP-chip have also evolved considerably. Early studies used “spotted” arrays, with PCR products for probes. Such arrays achieved good coverage in simple organisms, interrogating essentially all intergenic sequence (Ren et al., 2000), or all coding and non-coding sequence (Iyer et al., 2001) in *Saccharomyces cerevisiae*. For higher eukaryotes with more complex genomes, however, spotted arrays can only interrogate a small portion of the genome. Initially, researchers restricted focus to the most interesting subsets of these genomes: putative promoter regions, for instance, or CpG islands (Ren et al., 2002; Mao et al., 2003; Odom et al., 2004). Biased arrays of this latter type obviously can only identify sites of interest within the interrogated range.

In the case of transcription factor binding, biased arrays will therefore overlook binding sites which are located at larger than expected distances from annotated transcription start sites, which are found in non-canonical positions—within introns or exons, or 3' to the transcribed sequence—or which are associated with unannotated genes. There is a growing body of evidence, however, that such interaction occurs with appreciable frequency (Martone et al., 2003; Cawley et al., 2004; Euskirchen et al., 2004; Kirmizis and Farnham, 2004; Bertone et al., 2005; Sikder and Kodadek, 2005).

Recently, unbiased arrays which "tile" probes across a higher eukaryotic genome—over single chromosomes at first, and now across essentially all non-repetitive sequence—have been used. (As mentioned above, probes which target repetitive sequence are difficult to use because their observed intensities cannot be associated with a specific genomic locus.) Some such studies have continued to rely on spotted array technology: MacAlpine et al. (2004), for example, tiled 90% of the non-repetitive euchromatic sequence from the left arm of *Drosophila* chromosome 2 with ≈11,000 PCR products; Martone et al. (2003) and Euskirchen et al. (2004) mapped NF-κB and CREB binding, respectively, with a spotted array based on a library of ≈21,000 PCR products interrogating all non-repetitive sequence on human chromosome 22. Other studies—in *Arabidopsis* (Yamada et al., 2003), *Drosophila* (Schwartz et al., 2006), and human (Cawley et al., 2004; Kim et al., 2005; Lee et al., 2006)—have used high-density *in situ* synthesized oligonucleotide arrays, which contain short probes synthesized directly on the array substrate using photo-lithography or ink-jet printing. Due to the difficulty in producing and maintaining large PCR product libraries, as well as the higher feature density and improved reproducibility achievable with oligonucleotide arrays, the latter provide the platform of choice for future whole-genome tiling applications in these organisms (Mockler et al., 2005).

1.2 Issues for analysis of ChIP-chip data

The main topics making up the body of this dissertation arose from an examination of the extensive ChIP-chip data set presented in Cawley et al. (2004)—the first study to illustrate the potential of oligonucleotide tiling arrays for high-resolution transcription factor binding site identification in humans. To analyze their data, Cawley et al. applied the classic Wilcoxon rank sum test in moving windows: for each queried genomic position, all control array scores associated with probes within ±500 base pairs (bp) were gathered into one set, and all treatment (ChIP) array scores from the same set of probes were gathered into another. Ranks were assigned based on the union of the two sets, the ranks for the treatment set were summed to produce a test static, and a p-value was assigned based on the normal approximation to the rank sum's null distribution. This same approach is still implemented in the Tiling Analysis Software (TAS) that Affymetrix makes available to its ChIP-chip users (Affymetrix, Inc., 2006).

The use of the non-parametric Wilcoxon rank sum was largely motivated by the authors' choice of probe-level score. Affymetrix arrays have traditionally included both perfect match (PM) probes, which are exactly complementary to their target, and mismatch (MM) probes, which differ from the PM probe only at the central base. Short oligonucleotide probes of the type found on Affymetrix arrays are known to hybridize non-specifically in many cases, returning undesired signal when presented with targets which are partially, but not exactly complementary. The MM probes, it was thought, can quantify this effect: neither the PM nor the MM probe is exactly complementary to such unintended targets, so one might hope that the two would respond in the same way. In this case, $PM - MM$ should provide a measure of specific signal. In a typical ChIP-chip experiment using such arrays, however, a problem arises: it is common to find $MM > PM$ for anywhere between 20% and 50% of the probe pairs. Since negative specific hybridization is not physically possible, and negative estimates preclude a log transformation (which is convenient for reducing skew and stabilizing variance), Cawley et al. selected $\log(PM - MM \vee 1)$ as their probe-level score, and proposed a non-parametric test statistic to accommodate its odd distribution.

Unfortunately, as will be discussed in Chapter 2, ChIP-chip data violate several assumptions required for validity of the asymptotic Wilcoxon rank sum p-values. The work of Chapter 2, then, was motivated by a desire to produce an alternative procedure yielding marginally valid p-values, and, if possible, to improve power at the same time. Doing so, of course, first requires an understanding of the process which produces the probe-level intensity scores. Accordingly, in Chapter 2 we present a generative model for the probe-level signal obtained from ChIP-chip using high-density tiling arrays, show that this model's characteristics are consistent with observed data, and then suggest a straightforward testing procedure which is appropriate for data so-generated.

To convincingly demonstrate the superiority of the proposed testing procedure, one would ideally use knowledge of the real locations at which the protein under study interacts with the DNA, and then evaluate sensitivity and specificity. As is often the case, however, such knowledge is only available in a limited supply, or is artificial in nature, leaving the generalizability of results in question. In Chapter 3, we present a partial solution to this problem: a variant on traditional receiver operating characteristic (ROC) curves which permits, under certain assumptions, comparisons of competing methods in the absence of gold standard test set data.

Finally, Chapter 4 contains a detailed review of analysis methods recently proposed for ChIP-chip data. These are considered in light of some of the issues raised in Chapter 2, and compared via the ROC and pseudo-ROC metrics of Chapter 3.

Chapter 2

A generative model

2.1 Oligonucleotide tiling arrays

As discussed in the Introduction, tiling arrays—which cover all non-repetitive sequence with a more-or-less evenly spaced grid of probes—feature one of two possible probe types: spotted PCR products, or *in situ* synthesized oligonucleotides. Analysis of data obtained from spotted tiling arrays is largely analogous to analysis of data obtained from spotted biased arrays, since the resolution of the probe grid on such arrays is typically low enough that correlation between probes with neighboring genomic targets may be safely ignored. Oligonucleotide tiling arrays, on the other hand, achieve a very high probe density. The *D. melanogaster* tiling arrays discussed below feature gaps of only 11 bp, on average, between the 25-mer probes; tiling arrays for yeast feature probes which actually overlap, with each sharing 21 of its 25 bases with the next probe in the tiling (David et al., 2006). Under typical protocols, the average fragment size for the labeled DNA hybridized to the microarray in a ChIP experiment ranges from several hundred to over one thousand bases, and is thus much larger than the typical inter-probe interval (Schwartz et al., 2005; Sikder and Kodadek, 2005). Several authors have speculated about how this relationship between fragment size and probe spacing impacts the statistical properties of estimators and the precision with which binding sites can be specified (Buck and Lieb, 2004; Keleş et al., 2006). In this chapter, we develop these ideas is greater detail with a simple statistical model. We focus on the search for transcription factor (TF) binding sites, although the model also applies to the Pol II and III or ORC localization examples cited in the Introduction—or to any other point-like phenomenon. (It does not directly apply to the interval-like mapping of histone modifications, although the adaptations necessary for handling such phenomena are straightforward.)

Although the model only approximates reality, we show that its general characteristics agree with observation. Further, its mathematical tractability permits description of two important structural aspects of the data: (i) the size and shape of signal response in regions of transcription factor binding, and (ii) the nature of the correlation we expect to find among probe-level statistics—both in the neighborhood of a binding site and also in background noise. The first item is relevant to the design of powerful statistical tests for binding site detection; the second is crucial for selecting appropriate methods for estimation of error rates, and for avoiding statistical traps that result from incorrect assumptions about background noise.

More specifically, many of the currently proposed binding site detection algorithms look for increased signal on the treatment arrays in multiple, consecutive probes in order to make their positive calls. The presence of appreciable spatial correlation in the background signal—i.e., background signal with a naturally occurring wave and trough character—means that spurious, spatially coordinated "enrichment" will appear throughout the data, favoring the treatment arrays in some cases, and the control arrays in others. Failure to take this behavior into account can lead to incorrect p-values (or posterior probabilities, for hidden Markov models) and excessive false positives. In Figure 2.1, for example, we have applied the non-parametric test statistic proposed by Cawley et al. (2004) to the *D. melanogaster* Zeste data set described in the next section. This procedure (i) ignores differences in probe hybridization efficiency, and (ii) assumes independence in fluorescence intensity signal from one probe to the next. Ignoring differences in probe hybridization efficiency should, in fact, lead to overestimation of test statistic variance and a null p-value distribution which is compressed away from the extremes of the unit interval and towards .5. (Varying probe hybridization efficiency causes each probe's observed intensities to cluster across replicates: efficient probes repeatedly produce higher intensities, and less efficient probes, lower intensities. As a consequence, there is less variability in the rank sum than would have been the case were all probe intensities equally distributed.) Instead, we often see the exact opposite, with computed p-values skewed towards the extremes of the unit interval. One possible explanation for this is the presence of coordinated movement in signal, arising from spatial correlation, even in regions of the genome where no ChIP-based enrichment has occurred. TAS's windowed Wilcoxon rank sum procedure will assign extreme p-values—values close to 0 for a wave favoring the treatment arrays, and values close to 1 for a trough favoring the control arrays—to such regions. In practice, of course, we would

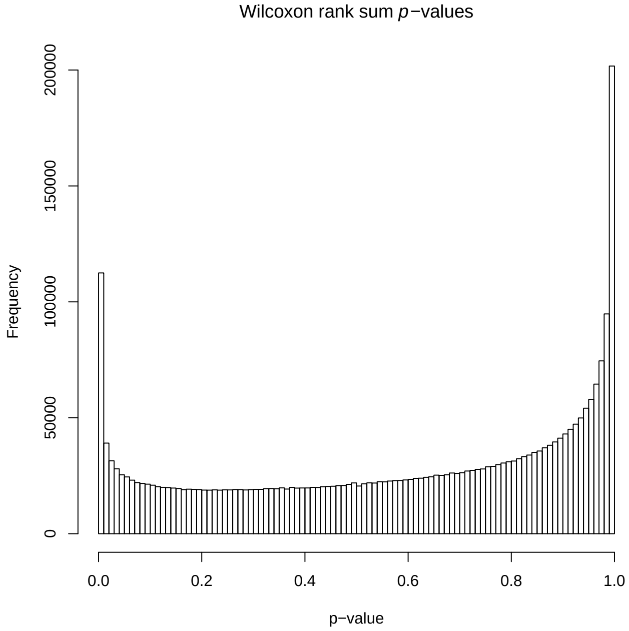

Figure 2.1: Distribution of p-values produced by the non-parametric test statistic proposed by Cawley et al. (2004). This procedure ignores differences in probe hybridization efficiency; as a consequence it should overestimate test statistic variance, and produce null p-values which are compressed from both ends towards .5. In fact, the impact of spatial correlation in the background signal is strong enough to overcome this effect, producing an excess of "significant" p-values at both ends of the distribution.

only be interested in the extreme p-values near 0; but extreme p-values near 1 define regions we would have identified as significant had we reversed the roles of treatment and control. It is troubling to observe that this method finds so many "binding sites" in the control experiment.

Later in this chapter, we will describe a simple statistical procedure which automatically incorporates the wave-and-trough behavior of the background signal into its estimated null distribution. Thus, significant p-values are only assigned to regions of coordinated enrichment in the treatment experiments whose magnitude exceeds that of the background behavior. Figure 2.2 shows the distribution of p-values obtained when this method is applied to the same *D. melanogaster* Zeste data set: the number of extreme p-values near 0 is reduced, and the distribution is uniform throughout the rest of the range.

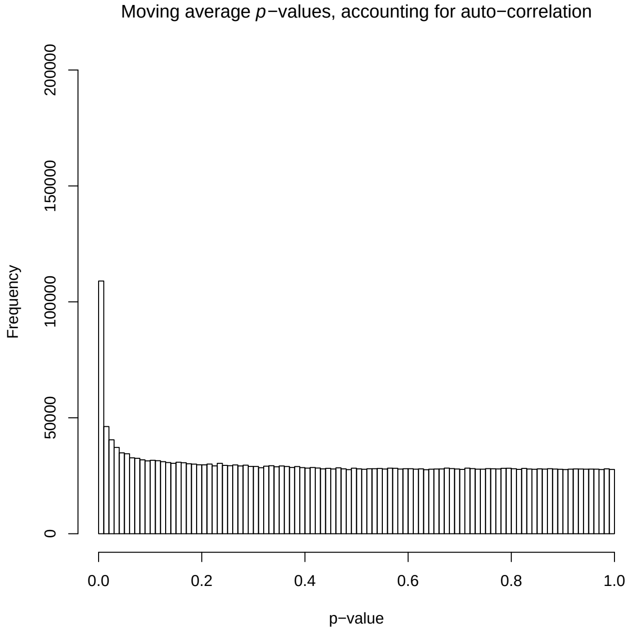

Figure 2.2: Distribution of p-values produced by a statistical method which takes spatial correlation in the background signal into account. Regions are only assigned significant p-values when (i) probes in that region exhibit coordinated enrichment in the treatment experiments, and (ii) when the magnitude of such enrichment exceeds that of the wave-and-trough background behavior. The slow rise on the left is related to the use of a moving window.

2.2 *D. melanogaster* data used in examples

Examples in this chapter are derived from a set of nine microarrays produced by the Berkeley *Drosophila* Transcription Network Project (BDTNP), which kindly made the data available. BDTNP is currently preparing manuscripts with a complete description of experimental methods and analysis of implications for transcriptional regulation in the *Drosophila* embryo. For our purposes, however, an brief overview of the experimental structure is sufficient:

For treatment arrays, two independent immunoprecipitations were performed on formaldehyde-crosslinked *D. melanogaster* embryos using an anti-Zeste antibody. Each of the two immunoprecipitated DNA samples was then amplified by random-primed PCR, labeled, and sequentially hybridized to three Affymetrix microarrays, yielding six treatment arrays in total. The microarrays contained 25-mer, *in situ* synthesized probes mapping to almost 3 million positions in non-repeat, euchromatic DNA on Drosophila chromosomes 2, 3, 4 and X. Median spacing between probe starts was 36 bases. Three control arrays were prepared using amplified and labeled input DNA (i.e., the immunoprecipitation step was omitted).

2.3 Steps in the ChIP-chip assay

Excellent descriptions of the ChIP-chip experimental procedures have been published elsewhere (Buck and Lieb, 2004; Bertone et al., 2005; Mockler et al., 2005; Sikder and Kodadek, 2005). Nonetheless, we quickly review the main steps, in more detail than was given in the Introduction, to fix terminology:

1. Protein-DNA and protein-protein crosslinks are created by exposing cells to formaldehyde.
2. Chromatin is extracted from cells and fragmented by sonication.
3. DNA fragments crosslinked to the transcription factor of interest are preferentially selected by immunoprecipitation using a protein-specific antibody.
4. Control DNA may be obtained directly from input DNA, or from a "mock IP" in which the antibody is omitted, an alternative non-specific antibody is used, or in which chromatin is obtained from cells not expressing the target protein.
5. Crosslinks are reversed and DNA is purified.
6. DNA is amplified by random-primed or ligation-mediated PCR. Alternative protocols—e.g., *in vitro* transcription (IVT)—may be used, or amplification may be omitted if multiple IP reactions are carried out and their results are pooled.
7. Fluorescent labels may be incorporated during PCR; alternately, amplicons or source DNA (if PCR is omitted) may be labeled separately.
8. Standard hybridization, wash, stain, and scan protocols are followed.

2.4 A statistical model and its implications

2.4.1 Abundance and fluorescence intensity

In situ synthesized oligonucleotide arrays provide a much higher probe density than is possible with spotted PCR products, thereby permitting more precise localization of binding sites. A major drawback of the short oligonucleotide probes, however, is

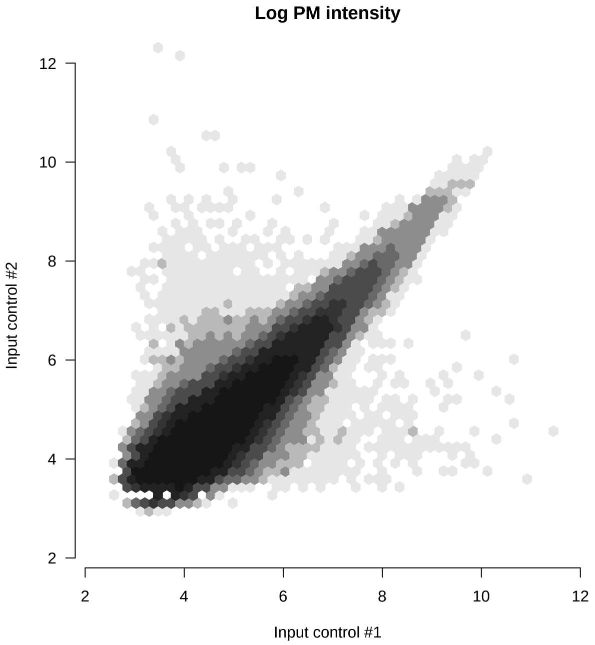

Figure 2.3: Scatter plot of log-scale PM intensities for two input DNA control experiments. Darker bins indicate a higher density of points. Strong correlation ($r = .87$) is largely due to the wide range of target affinities.

the sequence-dependent variability in hybridization efficiency, or "target affinity." This variability arises in part from differences in GC content (paired G and C bases are joined by three hydrogen bonds instead of two) and higher-order effects of sequence on melting temperature, and in part, under certain labeling protocols, from the biotinylation of some nucleotides but not others (Naef and Magnasco, 2003; Zhang et al., 2003; Buck and Lieb, 2004; Wu et al., 2004).

Whatever be the source of the variability, Figure 2.3 demonstrates the relevance of this issue to ChIP and short oligonucleotide tiling arrays. Here, we compare all log-scale perfect match[1] probe intensities from two Affymetrix *D. melanogaster* tiling arrays to which two different input DNA control samples were hybridized. In theory, the same amount of DNA should be available for hybridization to every probe, and observed variability in fluorescence intensity across the probes on a single array should arise from noise only; as such, for a given probe the observed intensity from one experiment to the

[1] Although mismatch probes were included on the arrays used to generate data shown in this chapter, we have chosen not to use them. For gene expression analysis, the use of MM probes is controversial. For ChIP experiments, where both treatment and control arrays are available, it is our opinion that the value of MM probes is even more limited, and does not justify sacrificing 50% of the space on the arrays—particularly when high-density, unbiased coverage is desired. In Chapter 4, a comparison of PM-only analyses with methods using both PM and MM probes supports this claim.

next should vary independently. A small proportion of probes, of course, are expected to be subject to systematic effects—excessive cross hybridization, poor synthesis efficiency, etc.—and to exhibit consistently high, or low, intensities in both experiments in a pair; in fact, observed intensity for *most* probes is strongly correlated. Although bias in the random-primed PCR amplification step may contribute to some extent, the observed correlation is largely due to target affinity effects: "sticky" probes tend to produce high fluorescence intensities on all arrays, whereas probes with poor hybridization properties tend to produce low fluorescence intensities on all arrays. Further the range of intensities shown in Figure 2.3 spans the dynamic range of the arrays: the magnitude of this nuisance variability is comparable to that of the enrichment we are hoping to detect.

Most methods for the analysis of data from gene expression experiments using Affymetrix arrays model the effect of target affinity in a multiplicative fashion, with a different multiplier for each probe to capture the sequence-based variation in hybridization efficiency (Li and Wong, 2001; Irizarry et al., 2003; Zhang et al., 2003; Wu et al., 2004). Further, the more successful methods combine this multiplicative target affinity with a "multiplicative error, additive background" model, and such error models are empirically well-supported (Durbin et al., 2002; Irizarry et al., 2003; Wu et al., 2004). The multiplicative error reflects intensity-dependent imprecision in the relationship between target abundance and measured fluorescence; the additive background is assumed to arise from optical noise and cross hybridization. Under such a model, if it were possible to eliminate the additive background (including the additive contribution made by non-specific hybridization), then the observed intensity for probe i on array j would be

$$I_{ij} = \alpha_i A_{ij} \epsilon_{ij}, \tag{2.1}$$

where α_i is the target affinity multiplier, A_{ij} is the actual abundance of target DNA available for hybridization to probe i in sample j, and ϵ_{ij} is the non-negative multiplicative error. Again in the absence of additive background, the use of control arrays permits us to define a log-ratio statistic—the difference of average log-scale intensities—for which the nuisance probe effect terms cancel out. Given n^T treatment arrays and n^C control arrays, define

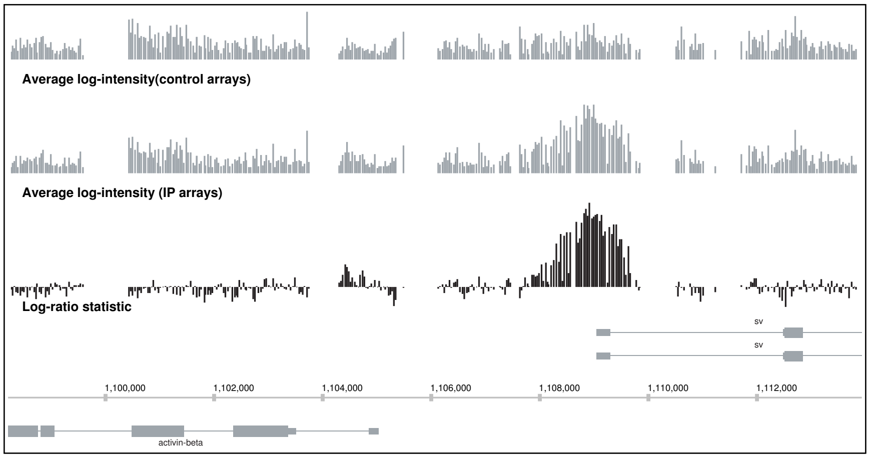

Figure 2.4: An example of average log-scale PM intensities (in gray) and the associated log-ratio statistics (in black) for a region of *D. melanogaster* chromosome 4 ($n^T = n^C = 3$). Correlation due to varying target affinities can be seen by comparing the two intensity graphs. The log-ratio statistic largely cancels the effect of varying probe affinities, causing a region of apparent transcription factor binding to more clearly stand out above the noise.

$$\begin{aligned}
LR_i &= \frac{1}{n^T}\sum_{j=1}^{n^T} \log I_{ij}^T - \frac{1}{n^C}\sum_{j=1}^{n^C} \log I_{ij}^C \\
&= \frac{1}{n^T}\sum_{j=1}^{n^T} \log A_{ij}^T - \frac{1}{n^C}\sum_{j=1}^{n^C} \log A_{ij}^C + \left(\frac{1}{n^T}\sum_{j=1}^{n^T} \log \epsilon_{ij}^T - \frac{1}{n^C}\sum_{j=1}^{n^C} \log \epsilon_{ij}^C \right) \qquad (2.2) \\
&\equiv \frac{1}{n^T}\sum_{j=1}^{n^T} \log A_{ij}^T - \frac{1}{n^C}\sum_{j=1}^{n^C} \log A_{ij}^C + \delta_i.
\end{aligned}$$

Thus, the difference in *observable* average log-scale, background-corrected intensities is just a noisy version of the difference in *unobservable* average log-scale abundances. Variants on this log-ratio statistic divide it by an estimated standard deviation for the difference (e.g., Ji and Wong, 2005; Keleş et al., 2006), and such t-statistics (examined in more detail in Chapter 4) also cancel out the target affinity terms.

Unfortunately, the elimination of the additive background is not a simple matter. Affymetrix originally included mismatch probes on their arrays for precisely this purpose, but the use of mismatch probes has been shown to be problematic in several respects. More recent PM-only approaches attempt to correct for the additive background through statistical estimation procedures (Naef et al., 2002; Irizarry et al., 2003; Wu et al., 2004). (We will also revisit the topics of mismatch probes and additive

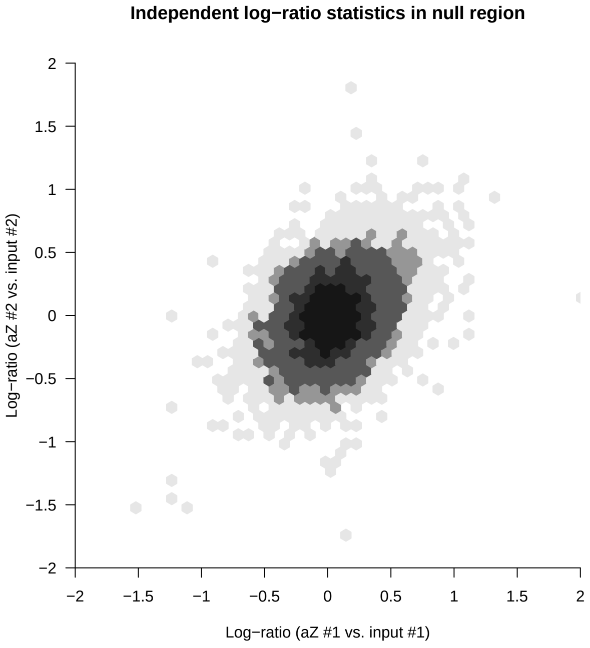

Figure 2.5: Scatter plot of the log-ratio statistics from two independent anti-Zeste/input pairs. A region with no apparent TF binding (including approximately 11.2K probes) was manually selected. Darker bins indicate a higher density of points. Taking ratios has dramatically (but not completely) reduced the impact of probe-to-probe variation in target affinity: some residual correlation ($r = .31$) remains.

background in Chapter 4.) Nonetheless, Figure 2.4 demonstrates that significant improvements in signal-to-noise ratio can be achieved with a ratio-based statistic, even if additive background is ignored altogether: visual comparison of the first two graphs suggests an enriched region just upstream of the positive strand annotation, and also confirms a high degree of similarity in observed fluorescence intensity throughout the region shown. In the third graph, where the log-ratio statistic of (2.2) is plotted, the enriched area now stands out much more clearly above the noise. Thus, by using a ratio to (at least partially) eliminate the nuisance variability caused by probe-to-probe differences in hybridization efficiency, we can more easily see the interesting variability created by transcription factor binding.

Figure 2.5 suggests that this reduction in nuisance variability is significant, but is not complete. Here, two independent treatment-to-control log-ratio statistics were computed for a manually selected region which exhibited no apparent enrichment. Were the multiplicative model in (2.1) exactly correct (and sequence-specific differences in PCR efficiency negligible), we would expect zero correlation between the two log-ratio statistics. In fact, some residual correlation remains—most likely because non-zero additive background prevents exact cancellation of the α_i terms, or because the log-scale linearity implied by (2.1) is only approximate. The residual correlation is, however, far less than the correlation seen in Figure 2.3.

For the remainder of this chapter we will assume that an appropriate statistical estimation procedure has corrected for additive background, or that the magnitude of such additive background is small enough to safely ignore.

2.4.2 Assay model

Assuming that the intensity model (2.1) is approximately correct, we next present a statistical model for the assay, which permits a description of the relationship between the A_{ij} (and thus the LR_i) at neighboring probe positions, as well as the expected behavior of these quantities near a transcription factor binding site.

Ideally, chromatin immunoprecipitation would only permit DNA fragments which are crosslinked to an antibody-bound TF protein to pass; in fact, DNA from all regions of the genome passes to some extent.[2] To represent the process of IP passage and subsequent amplification, we propose the following assay model (represented graphically in Figure 2.6):

1. N input copies of the full DNA strand begin the process, each consisting of L bases.

2. Sonication leads to uniform, random fragmentation of the chromatin, so that the probability of a break at any given base of an input strand is some small constant, θ. Further, there is no interference: breakage at a given position is independent of whether or not other breaks have occurred nearby or on other input strands.

3. Fragments with no TF binding site pass IP anyway with some small positive probability ϕ; fragments with a binding site pass with probability ϕ', where $\phi' \gg \phi$. (The parameter ϕ' may be thought of as reflecting the binding site occupancy rate in the sample, the probability of antibody-antigen binding, and the probability of an antibody-tagged fragment being pulled down by IP.)

[2] There is also, presumably, some binding between the antibody and unintended protein targets. This issue is an inherent weakness in the ChIP method, and cannot be addressed by statistical techniques. One possible solution is to conduct multiple experiments using different antibodies which target distinct epitopes, and to focus on regions identified by both antibodies (Cawley et al., 2004; Kim et al., 2005).

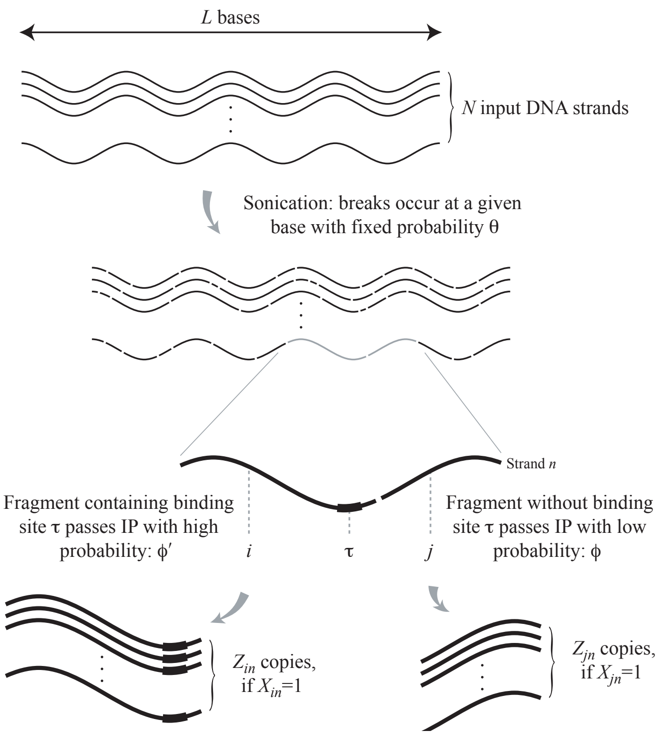
L bases
N input DNA strands
Sonication: breaks occur at a given base with fixed probability θ
Strand n
Fragment containing binding site τ passes IP with high probability: ϕ′
i
τ
j
Fragment without binding site τ passes IP with low probability: ϕ
Z_{in} copies, if X_{in}=1
Z_{jn} copies, if X_{jn}=1

Figure 2.6: A model for the ChIP assay. Sonication fragments input strands uniformly at random. Fragments pass IP with a probability which depends on whether or not they contain a binding site. ($X_{in} = 1$ implies that fragment i on input strand n passes IP.) Fragments passing IP are amplified to Z_{in} copies, where Z_{in} is independent of Z_{jm} if $n \neq m$, or if $n = m$ but i and j end up on distinct fragments as shown above.

4. All fragments passing IP are subject to a common amplification process, yielding a random number of copies of each fragment. The random variable Z will be used for amplification. Z may arise from a branching process for PCR, a linear birth process for IVT, or simply be the unit constant if multiple IP reactions are pooled as an alternative to amplification (Shaw, 2002). The exact nature of Z's distribution is unimportant, provided that its variance is finite.

More formally, for a single sample, the abundance of fragments available for specific hybridization to probe i is given by

$$A_i = \sum_{n=1}^{N} X_{in} Z_{in}, \tag{2.3}$$

where the X_{in} are 0/1 indicator variables which describe whether the fragment of DNA on strand n containing probe i passed IP, and for which $\mathbb{P}(X_{in} = 1)$ will depend on ϕ, ϕ', and the proximity of i to a binding site. The Z_{in} are amplification variables, whose value is only relevant when $X_{in} = 1$.

Clearly, such a model is only an approximation. The assumption of no interference cannot be taken literally, for example, and there is evidence that chromatin shearing is not uniform, but rather varies with chromatin density, causing regions which associate with high molecular weight protein complexes to be more sensitive to shearing (Schwartz et al., 2005). There is also known bias in PCR amplification, with respect to both base composition and fragment size (Liu et al., 2003). Despite such shortcomings, the model provides insight, and permits us to makes several important predictions about the statistical properties of the probe intensities and log-ratio statistics. As shown below, these predictions are in large part born out by actual data.

2.4.3 Fragment size

The assay model implies an intuitively obvious inverse relationship between the breakage probability parameter θ and the average fragment size after sonication. Because fragmentation for each input strand is driven by the same mechanical process, the overall average fragment size is the same as the average fragment size for a single strand. Under the assay model, the number of breaks produced in a single input strand is a random variable M with the Binomial(L, θ) distribution. Conditional on M, the average fragment size is just $L/(M+1)$, because M breaks imply $M+1$ fragments, and the total size of all fragments must equal L. So using the binomial distribution density,

$$\begin{aligned} \mathbb{E}F &= \sum_{m=0}^{L} \frac{L}{m+1} \binom{L}{m} \theta^m (1-\theta)^{L-m} \\ &\approx \frac{1}{\theta}, \end{aligned} \tag{2.4}$$

provided that θL is large. Since θL is the expected number of breaks in each input strand, this will in fact be large under standard ChIP protocols, and the approximation will be very good. A justification for the approximation of (2.4), and for other results to follow, is given in the Derivations section at the end of this chapter.

2.4.4 Expected signal size and shape

Consider a probe i at some distance Δ from a transcription factor binding site τ, and assume for the moment that i is sufficiently far from any other binding sites that their combined effects are negligible. Buck and Lieb (2004) suggested, informally, that detected enrichment should correlate inversely with the distance of the binding site from the arrayed element; and Kim et al. (2005) attempted to more formally relate the log-ratio statistic to Δ (although their derivation of a linear decay model is incorrect).

In fact, an inverse relationship between target abundance and distance follows directly from the assay model. As illustrated in Figure 2.6, for a DNA fragment containing sequence complementary to probe i, the probability of passing IP is either ϕ' or ϕ, depending on whether or not the fragment also contains τ. Further, i and τ are joined if and only if no breaks occur between the two loci—an event which occurs with probability $(1-\theta)^\Delta$. Therefore,

$$\begin{aligned}\mathbb{P}(X_i = 1) &= (1-\theta)^\Delta \phi' + \big(1 - (1-\theta)^\Delta\big)\phi \\ &\equiv \pi(\Delta).\end{aligned} \tag{2.5}$$

Given the definition in (2.3), it now follows that after IP and amplification, the expected abundance of fragments possessing sequence complementary to probe i is

$$\mathbb{E}A_i = N\pi(\Delta)\mathbb{E}Z. \tag{2.6}$$

Observe that the fragment passage probability, $\pi(\Delta)$, decays exponentially from ϕ' to ϕ as the distance from the probe to the binding site increases.

Were all probe affinities—the α_i in (2.1)—equal, then equation (2.6) would also imply that the expected values of background-corrected fluorescence intensities should decay exponentially in Δ, from $N\phi'\mathbb{E}Z$ to $N\phi\mathbb{E}Z$ (up to a proportionality constant). Unfortunately, we have seen this is not the case; the relationship in (2.6) can, however, be extended to the log-ratio statistic, for which the impact of varying probe affinities is largely eliminated. Assuming that there is no antibody-directed specific enrichment in the control experiments, it is shown below that

$$\begin{aligned}\mathbb{E}LR_i &\approx \log\left(\frac{\pi(\Delta)}{\phi}\right) + K \\ &= \log\big((1-\theta)^\Delta(\phi'/\phi - 1) + 1\big) + K,\end{aligned} \tag{2.7}$$

where K is a constant that is typically zeroed out by normalization, and the ϕ and ϕ' parameters apply to the treatment experiments, i.e., those in which the full IP procedure is used. (There is no ϕ' for the control experiments since there is no antibody-directed specific enrichment. The baseline fragment passage rate, ϕ, is relevant to the control experiments, and in practice need not be the same as its counterpart in the treatment experiments. Similarly, the amplification variable Z will typically not have the same distribution in the treatment experiments as it has in the control experiments because

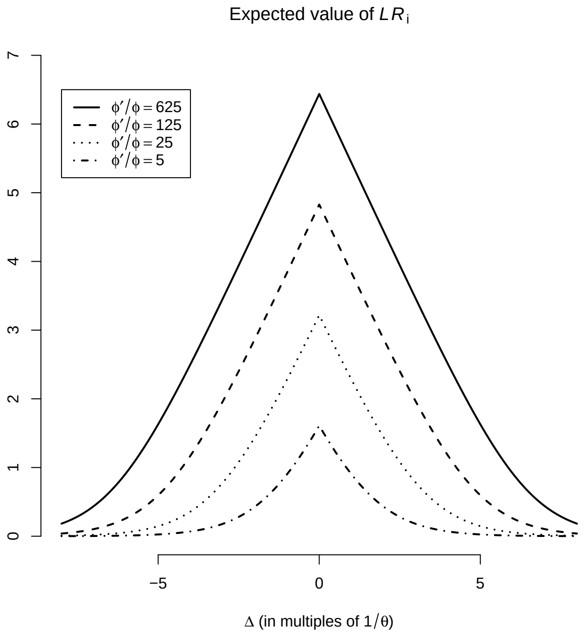

Figure 2.7: Expected value of the log-ratio statistic under the assay model of Section 2.4.2. The width of the peak is given in units of $1/\theta$, which in (2.4) is shown to be approximately equal to the average fragment size after sonication. The dimensions of the peak, as well as the speed with which it transitions from a linear to an exponential regime, are a function of the ratio ϕ'/ϕ, i.e., of the stringency with which the IP step filters out unwanted fragments.

different numbers of amplification cycles are often required. These differences can, however, all be pushed into the constant K, and most normalization methods effectively center the log-ratio process, i.e., force the value of K to 0.)

Equation 2.7 no longer represents simple exponential decay. As shown in Figure 2.7, the decay is approximately linear near the binding site, and then becomes more exponential further away. The scale on the horizontal axis in Figure 2.7 is given in units of $1/\theta$, which we have seen to be the expected fragment size after sonication. The horizontal and vertical dimensions of the log-ratio peak near a TF binding site are also seen to be a function of the ratio ϕ'/ϕ. This ratio relates the probability, in the treatment experiments, that an antibody-tagged fragment passes IP to the probability that an untagged fragment sneaks through. Figure 2.7 demonstrates, naturally, that more stringent and effective IP filtering leads to a more easily detectable signal at a binding site.

Recall that the parameter ϕ' is binding site specific, reflecting the occupancy rate of the binding site and the antibody-TF affinity, in addition to other aspects of the IP procedure. (So we should really write ϕ'_τ to reflect this dependency.) For another binding site σ with a lower occupancy rate in the sample than the site at τ, or at which the epitope is less accessible to the antibody, we would have $\phi'_\sigma < \phi'_\tau$. As a consequence, the peak in the expected log-ratio statistic at σ would be both shorter and narrower than the peak at τ. (See, for example, the secondary peaks in several of the examples of Figure 2.8.)

Do we in fact see a peak-shaped response in the log-ratio statistics in the neighborhood of a transcription factor binding site? Figure 2.8 shows five examples which are typical of the putative Zeste binding sites identified using the log-ratio statistic. In each example, two log-ratio processes are shown: one for each of the two independent IP and PCR amplification reactions. (Only one set of three input arrays was run, so these log-ratio processes are not fully independent. Similar graphs generated from other experiments with multiple, independent treatment and control arrays, however, show equivalent results.)

It is important to recognize that (2.7) gives only the expected value of the log-ratio statistic in the neighborhood of a binding site. The degree to which the functional form of (2.7) is actually visible in the data depends on the magnitude of the variance of the LR statistic relative to the width and height of the peak, as well as the spacing of probes relative to these same dimensions (and, of course, on the degree to which our statistical model has actually captured the character of the underlying process). From Figure 2.8 it is clear that a peak-shaped signal does appear, even if some discrepancies exist between what the model predicts and what is observed in the data. In Section 2.5, we address the implications of a peak-shaped signal for the design of statistical procedures and for downstream analysis.

2.4.5 Covariance away from binding sites

Current statistical approaches to binding site detection using short oligonucleotide tiling arrays have, explicitly or implicitly, required multiple, consecutive probes to report positive signal before making a positive call (Cawley et al., 2004; Ji and Wong, 2005; Keleş, 2005; Kim et al., 2005; Li et al., 2005; Johnson et al., 2006; Keleş et al., 2006). Requiring such spatial corroboration helps to reduce the impact of single, errant probes which are very bright due to cross-hybridization or to specks and streaks on the slide. And, indeed, the results of Section 2.4.4 confirm that multiple probes should respond when real binding occurs: if, for example, average fragment size is 500 bp and

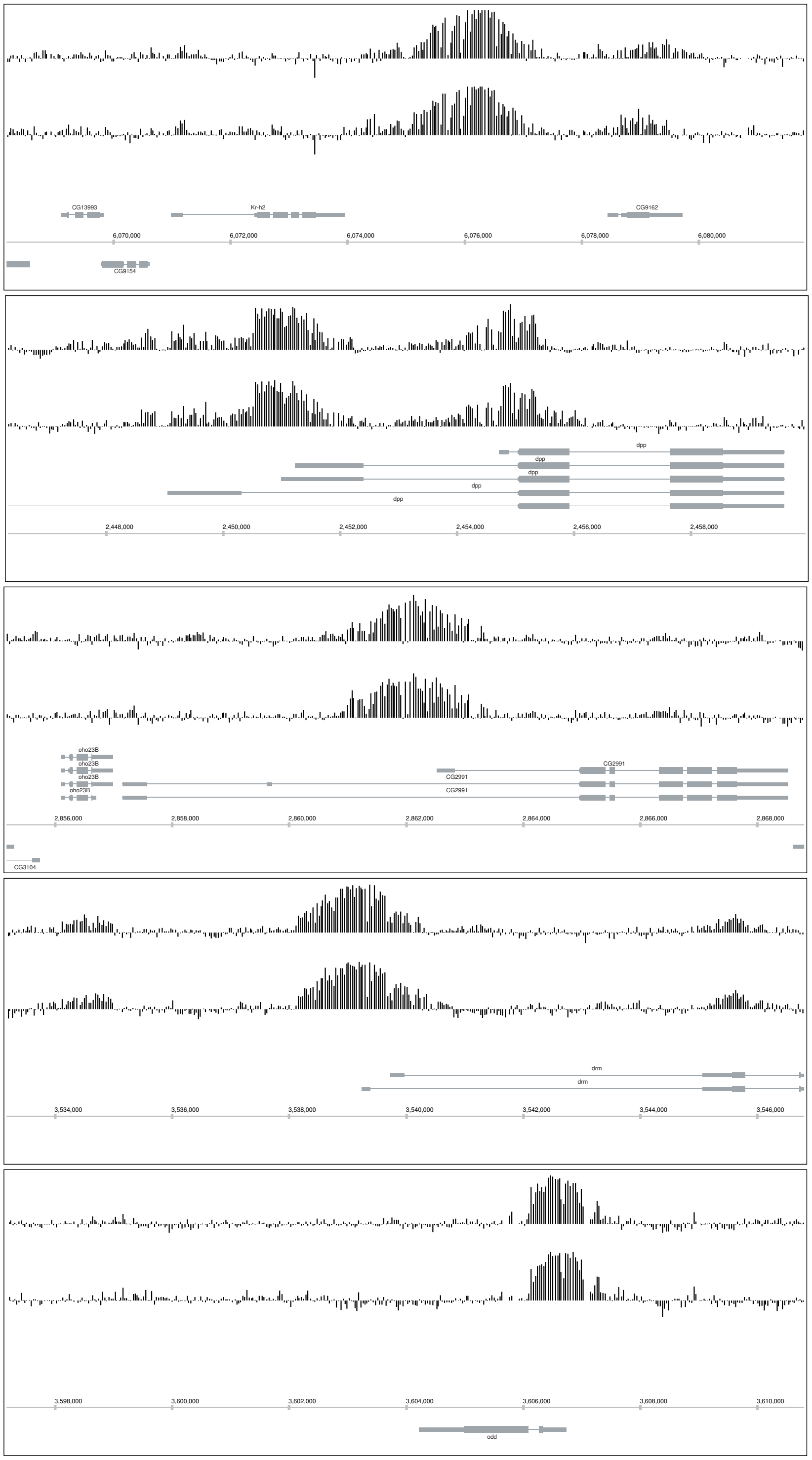

Figure 2.8: 3-chip vs. 3-chip log-ratio statistics (anti-Zeste vs. input DNA) for two independent IP and amplification reactions. Regions of enrichment exhibit a peak-shaped form.

the ϕ'/ϕ ratio is 25, the expected log-ratio statistic would be elevated above baseline over a range of several thousand bases. In practice we expect the tails of the peak to disappear into the noise to some extent, but with probes spaced every 36 bp, numerous consecutive positions should still detect appreciable signal.

A further consequence of our assay model, however, suggests that an increase in intensity for a run of neighboring probes should be a necessary condition for making a positive call, but it is *not sufficient.* Consider two probes i and j which are relatively far from any binding site, and are separated from one another by d bases. Taking Δ to be large in (2.5) gives $\pi(\Delta_i) \approx \pi(\Delta_j) \approx \phi$, so the expected abundance of target DNA for each position is $N\phi\mathbb{E}Z$. Further, it is shown in the Derivations section of this chapter that

$$\operatorname{Corr}(A_i, A_j) = (1-\theta)^d, \tag{2.8}$$

i.e., that there is correlation in target fragment abundance which arises from the proximity of i to j and the nature of the fragmentation procedure—even when no binding site is present. Figure 2.9 shows the rate of decay of this spatial correlation when the average fragment size is 500 bases (so that $\theta \approx .002$). Positive correlation is still appreciable at a distance of 1000 bases, i.e., over a large number of probes.

Abundance is not directly observable in ChIP experiments, but the spatial correlation present in the abundance variable passes through—in a modified form—to the observable log-ratio statistics as well. If we focus again on probes i and j which are distant from any binding site, then

$$\operatorname{Corr}(LR_i, LR_j) \approx (1-\theta)^d \left(\frac{\operatorname{Var}(\log A^T)/n^T + \operatorname{Var}(\log A^C)/n^C}{\operatorname{Var}(\log A^T)/n^T + \operatorname{Var}(\log A^C)/n^C + \operatorname{Var}\delta} \right). \tag{2.9}$$

Observe that this correlation is inversely related to the distance d as before, but does not tend to 1 as $d \to 0$ unless $\operatorname{Var}\delta = 0$ (which is, of course, never the case in practice). Further, the limit at short distances is a function of the relationship between the two sources of variability in the experiment: (i) $\operatorname{Var}\delta$, which relates to the array hybridization and scanning components of the experiment, and (ii) the $\operatorname{Var}(\log A)$ terms, which relate to variation in actual target abundance arising from the ChIP and amplification components. When the array side is very noisy (i.e., $\operatorname{Var}\delta$ is large), the correlation between the log-ratio statistics at i and j may be negligible, even for small d. When the array side is less noisy, however, the correlation between the two log ratios may be more clearly perceived. Thus two different researchers, each running their own arrays but beginning with common, post-IP/amplification samples, could create different levels

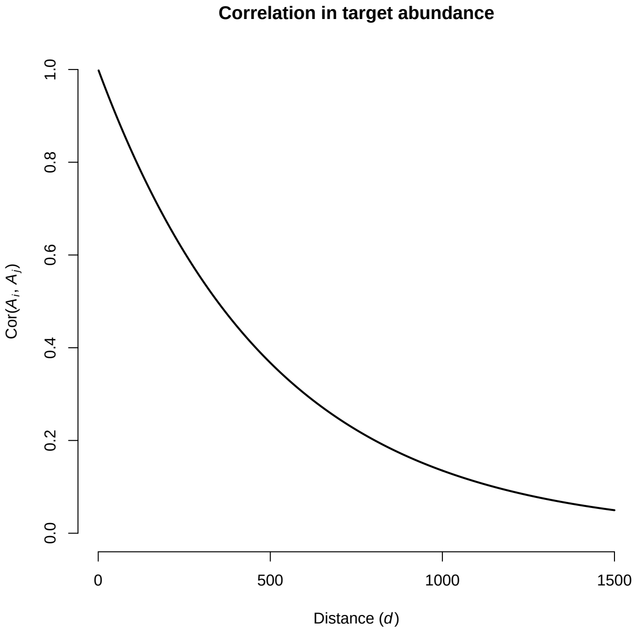

Figure 2.9: Correlation between target DNA abundance for two probes i and j, spaced d bases apart. Both are assumed to lie in a region distant from transcription factor binding sites, and the average fragment size (and thus $1/\theta$) is assumed to be 500 bases.

of variability in their corresponding δ terms, and might therefore see different levels of spatial correlation in the log-ratio statistics computed from their array data.

To see that appreciable spatial correlation does, in fact, exist in the log-ratio process, even in regions far from apparent binding sites, we again hand select a null region. Repeat masking and synthesis-efficiency filtering introduce gaps of varying sizes into to the regularly spaced probe tiling. To compute auto-correlation estimates, however, these gaps were ignored for simplicity: probe index number rather than exact position was used to compute the lag m auto-correlation estimate,

$$AC_m = \frac{\sum_{i=1}^{n-m}(LR_{i+m} - \overline{LR})(LR_i - \overline{LR})}{\sum_{i=1}^{n}(LR_i - \overline{LR})^2},$$

where n is the number of positions included in the null region, and $\overline{LR}$ represents the average log-ratio statistic over this region. Use of index number means that some probe pairs nominally separated by m steps are in fact separated by a much greater distance; as a consequence, downward bias is introduced, and actual autocorrelation estimates are most likely *higher* than those reported. Figure 2.10 shows that even in null regions, statistically significant spatial correlation is evident at lags of up to 20 or 30 positions,

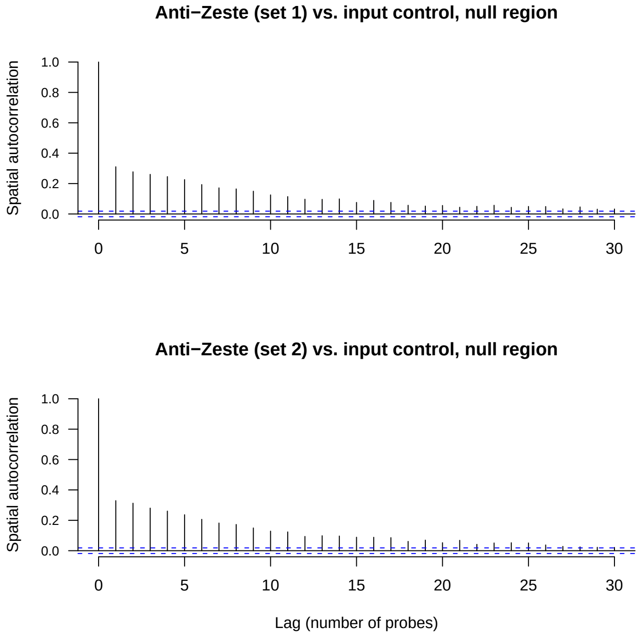

Figure 2.10: Observed spatial correlation in the log-ratio statistic, in a manually selected region (including approximately 11.2K probes) containing no apparent binding sites. Irregularity of probe spacing was ignored when the autocorrelation estimates were computed—producing downward bias. Thus true spatial correlation may be slightly higher than what is shown. (Dotted lines represent the level at which observed autocorrelation differs significantly from 0.)

i.e., 720 to 1080 bp.

2.5 Implications of the generative model

Before introducing a statistical enrichment detection procedure which can address data of this type, we first summarize the results of the preceding section and discuss their implications:

We have shown that for ChIP experiments which use *in situ* synthesized tiling microarrays, nuisance variability is created by the wide range of target affinities that result from differences in the probes' base composition. Figure 2.3 and Figure 2.4 show that the magnitude of the target affinity effects is quite large; as a consequence, statistical methods which fail to address target affinity—through cancellation or even, perhaps, direct estimation[3]—will most likely pay a price in sensitivity. When control

[3] For expression experiments using *in situ* synthesized arrays, the most successful methods work on multiple samples simultaneously, permitting direct estimation of the target affinity parameters. Such

experiments are available, however, the use of a ratio-based statistic can significantly reduce the impact of target affinity.

We next observed that our statistical model predicts—and actual data exhibit—peak-shaped signal in the vicinity of a binding site. Further, both the width and height of the peak are related to the binding site occupancy rate, the antibody's affinity for its target, and the stringency of IP filtering. (With some transcription factors, multiple binding sites may be found in close proximity. We omit details, but for a probe i located between two closely spaced binding sites, calculations similar to those used to derive $\pi(\Delta)$ show that the expected abundance has a catenary form between the binding sites, and that a superposition of the individual peaks still reasonably approximates the log-ratio signal.) The existence of peak-shaped signal has several important implications:

- Although multiple probes in the vicinity of a binding site may detect enrichment in the treatment experiments, those nearest the binding site are expected to detect more than those further away. Some authors have simply reported intervals over which enrichment has been detected, but this throws away valuable information: up to noise, the binding site is in fact most likely to be found in the center of the region, where the magnitude of enrichment is largest. Thus an understanding of the shape of expected signal permits more precise localization of binding sites than is possible using binary enriched/non-enriched calls.

- Often, a quantitative estimate of enrichment at a detected binding site is desired, and the average (or trimmed mean, median, etc.) signal over a range of probes near the site is computed. A second consequence of peak-shaped signal is that such estimates will be downwardly biased: although we are really interested in an estimate of signal at the peak exactly, averaging over multiple neighboring positions—for which signal trails off rapidly—will substantially dilute this quantity. An enrichment estimate based on the single position nearest to the apparent center of a peak would be too noisy to be of practical use; an estimator based on a parametric model for the peak (e.g., Kim et al. (2005)), on the other hand, would reduce variance and susceptibility to outliers by incorporating data from multiple probes, but would also avoid the watering-down of simple averaging. Taking an

methods focus, however, on small sets of probes believed to lie within a single transcriptional unit, so that for a given sample the expected underlying abundance can be assumed to be the same for each probe in the probe set. In the ChIP context, this common expected abundance assumption would only be appropriate for probe sets which fell entirely within a null region. As shown in Section 2.4.4, it would not hold for a probe set with members near a binding site.

average over an interval of apparent enrichment would, however, provide an unbiased estimate if the phenomenon under consideration were interval-like rather than point-like (e.g., histone modifications).

- Two-state hidden Markov models have recently been proposed for the detection of transcription factor binding sites in ChIP experiments (Li et al., 2005). While such methods appear to work reasonably well, they ignore the difference between point-like and interval-like phenomena. Presumably, improved sensitivity could be achieved by modeling signal response in regions of TF binding in a manner more consistent with expectation.

- A last consequence of peak-shaped signal relates to the use of exogenous or artificial sequence for estimation of sensitivity, specificity, and false discovery rates. In Chapter 4, for example, we will consider an experiment in which cloned DNA complementary to relatively long runs of probes was spiked into an input DNA background, to provide a positive control. Spiked-in clones bypass sonication and amplification, and as a consequence generate an interval-like signal. Sensitivity and specificity estimates based on the ability to detect this type of artificial enrichment may therefore overestimate a method's actual effectiveness at detection of the point-like response created by TF binding sites.

Finally, we have shown that the unobservable abundance processes, as well as statistics derived from the observable intensity measures, should—and do—exhibit spatial correlation. This correlation arises from the simple fact that, for two probes i and j whose targets are in close proximity, the corresponding target sequence tends to end up on the same fragment after sonication. Because both IP filtering and amplification take place at the fragment level, we see similar levels of post-amplification abundance.

2.6 An enrichment detection procedure

In Section 2.1 we showed how failure to account for the nature of the data can lead to misspecification of the distribution of the test statistic, and, as a consequence, to an excess of false positives. Cawley et al. (2004) ameliorated the problem to some extent by choosing a very stringent cutoff for the nominal "p-values." In general, however, one would like to select cutoffs based on the properties of the test statistic rather than on an *ad hoc* basis, and to have a given cutoff mean more or less the same thing across different experiments carried out under a range of conditions. No method presented thus far has satisfactorily addressed this thresholding issue: correlation makes analytic specification of the null distribution tricky, and is therefore typically ignored by both parametric and resampling- or permutation-based procedures.

The p-values shown in Figure 2.2, however, are derived from a ratio-based statistic—which, as we have seen, largely corrects for varying target affinity—and were computed with a semi-parametric method which can explicitly incorporate spatial correlation into the null model:

2.6.1 Probe- and window-level statistics

We begin with the probe-level log ratio statistic of (2.2):

$$LR_i = \frac{1}{n^T}\sum_{j=1}^{n^T} \log I_{ij}^T - \frac{1}{n^C}\sum_{j=1}^{n^C} \log I_{ij}^C.$$

We then smooth with a moving window, assigning values to the position of the probe at each window's center:

$$W_i = \frac{1}{|\{j : d(i,j) \leq w\}|} \sum_{\{j:d(i,j)\leq w\}} LR_j. \tag{2.10}$$

This smoothing provides increased detection power when, as is usually the case, peaks span multiple consecutive positions. In fact, assigning the W_i of (2.10) to each position i is equivalent to using a rectangular kernel to locally fit the RMA two-way linear model (Irizarry et al., 2003), with enrichment assumed to be constant in the neighborhood spanned by the kernel. This model, obtained by taking the log of both sides of (2.1), requires one linear constraint for identifiability. In the analysis of gene expression, it is customary to constrain the $\log \alpha_i$ parameters to sum to 0; here, it is more natural to assume unit abundance for the control data, so that the estimated abundance for the treatment data corresponds to a fold change. In this case, letting $\mathbf{Y}_j^k = (\log I_{1j}^k, \ldots, \log I_{mj}^k)'$ for a window centered on some position i_0 and containing a total of m probes, we may express the model as

$$\begin{pmatrix} \mathbf{Y}_1^C \\ \vdots \\ \mathbf{Y}_{n^C}^C \\ \mathbf{Y}_1^T \\ \vdots \\ \mathbf{Y}_{n^T}^T \end{pmatrix} = \begin{pmatrix} \mathrm{Id}_{m\times m} & \mathbf{0}_{m\times 1} \\ \vdots & \vdots \\ \mathrm{Id}_{m\times m} & \mathbf{0}_{m\times 1} \\ \mathrm{Id}_{m\times m} & \mathbf{1}_{m\times 1} \\ \vdots & \vdots \\ \mathrm{Id}_{m\times m} & \mathbf{1}_{m\times 1} \end{pmatrix} \begin{pmatrix} \log \alpha_1 \\ \vdots \\ \log \alpha_m \\ \beta \end{pmatrix} + \epsilon_{mn\times 1}. \tag{2.11}$$

We show in Section 2.8 that the ordinary least squares estimate of β in (2.11) is just

$$\hat{\beta} = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{n^T} \sum_{j=1}^{n^T} \log I_{ij}^T - \frac{1}{n^C} \sum_{j=1}^{n^C} \log I_{ij}^C. \right), \tag{2.12}$$

i.e., the smoothed log ratio statistic, W_{i_0}.

2.6.2 Estimating $\operatorname{Var} W_{i_0}$

Under the generative model presented in this chapter, the log ratio statistics for probes far from binding sites are identically distributed. The window-level W_i statistics, however, would only be identically distributed in null regions if the tiling were perfectly regular, such that $|\{j : d(i,j) \leq w\}| = |\{j : d(i',j) \leq w\}|$ for all i and i', and such that the spacing of probes within each window were identical as well. In fact, tiling arrays typically do not achieve such regular spacing because certain classes of probes are systematically omitted: probes targeting repetitive sequence which cannot be unambiguously mapped back to the genome, probes which would be difficult to synthesize (long runs of a single nucleotide, for example, create problems for the synthesis chemistry), and probes which are expected to have poor hybridization properties.

Given the resulting heterogeneity in the distributions of the window-level statistics, correct window-level p-values may be computed by one of two approaches:

1. Estimate the null distribution for each W_i separately, based on the spacing and number of probes falling in the corresponding window; or
2. Scale each W_i so that a single, common null distribution may be assumed, and then estimate this distribution.

For both approaches, it is useful to first look more closely at the expression for the serial correlation of null log ratio statistics shown in (2.9). Given a white noise process $\{Z_t\}_{t=-\infty}^{\infty}$, we may represent a first-order autoregressive, first-order moving average process, denoted ARMA(1,1), as

$$X_t = \kappa X_{t-1} + (Z_t + \lambda Z_{t-1}), \tag{2.13}$$

i.e., as autoregressive with parameter κ, but with first-order moving average noise $Z_t + \lambda Z_{t-1}$ taking the place of the white noise associated with a standard autoregressive (AR) process. (A stationary process of this type exists iff $\kappa \neq \pm 1$ and $\kappa + \lambda \neq 0$; we may further assume $|\kappa| < 1$ without loss of generality. For details, see, for example, Brockwell and Davis (2002).) For such a process, it is straightforward to show that for $i \neq j$,

$$\operatorname{Corr}(X_i, X_j) = \kappa^{d(i,j)} \left(1 + \frac{1-\kappa^2}{\kappa}(\lambda^{-1} + \lambda + 2\kappa)^{-1}\right). \tag{2.14}$$

Can κ and λ be selected to provide the autocorrelation function of (2.9)? The answer is "yes": we may simply set $\kappa = 1 - \theta$; further, we show in Section 2.8 below that given such a choice for κ, one may always select an appropriate λ so that

$$1 + \frac{1-\kappa^2}{\kappa}(\lambda^{-1} + \lambda + 2\kappa)^{-1} = \left(\frac{\operatorname{Var}(\log A^T)/n^T + \operatorname{Var}(\log A^C)/n^C}{\operatorname{Var}(\log A^T)/n^T + \operatorname{Var}(\log A^C)/n^C + \operatorname{Var}\delta}\right). \tag{2.15}$$

In other words, for any breakage parameter θ and ratio of array-side to ChIP/amplification-side variances, there is an ARMA(1,1) process consistent with the (approximate) log ratio autocorrelation function implied by the generative model presented in this chapter.

As an example, we applied a standard ARMA maximum likelihood estimation procedure (implemented in the `arima` function in R) to the same manually selected null region used to estimate the log ratio autocorrelation functions in Figure 2.10. The resulting parameter estimates were $(\hat{\kappa}, \hat{\lambda}) = (.91, -.72)$ for data set 1, and $(\hat{\kappa}, \hat{\lambda}) = (.91, -.71)$ for data set 2. Figure 2.11 shows the remaining spatial correlation in the residuals from these fits: only 2 estimates in 60 exceed the significance boundaries, a number consistent with a size of $\alpha = .05$ and a full null hypothesis of no remaining autocorrelation at any lag. Given $\hat{\kappa} = .91$, we may also estimate θ for these data. (Recall that $1/\theta$ gives the expected fragment size under the generative model.) Observing that single "steps" in our autocorrelation function typically correspond to steps of 36 bases for these data, we equate $(1 - \hat{\theta})^{36}$ with $\hat{\kappa}$. Solving, we obtain $1/\hat{\theta} \approx 382$, which is more or less consistent with measured values for the average fragment size found after sonication (e.g., Qi et al., 2006).

Assuming that an ARMA(1,1) model provides a reasonable fit to the log ratio process in regions far from binding sites, and assuming that we have $\hat{\kappa}$ and $\hat{\lambda}$ in hand, it is now possible to compute a variance estimate for W_i which accounts for the serial correlation of the log ratios. First, for simplicity and computational efficiency, "grid" the probes by mapping gaps to the nearest integer multiple of the typical inter-probe spacing. Then, let $\gamma(k)$ be the lag k autocovariance function for the so-modified process—so that k corresponds to the number of steps on the grid, rather than bases

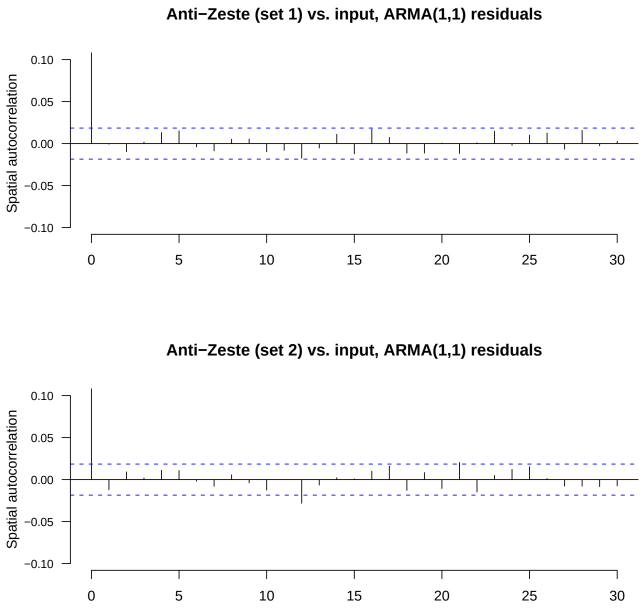

Figure 2.11: Fitting an ARMA(1,1) model to the same data presented in Figure 2.10 produces residuals with statistically significant autocorrelation: the model is sufficient to describe the spatial correlation present in the log ratio data. (Dotted lines represent the level at which observed autocorrelation differs significantly from 0.)

along the genome. It is simple to show (again, see Brockwell and Davis (2002)) that

$$\gamma(0) = \sigma^2 \left(1 + \frac{(\lambda + \kappa)^2}{1 - \kappa^2} \right) \equiv \sigma^2 \gamma_0$$

and for $h > 0$,

$$\gamma(h) = \sigma^2 (\lambda/\kappa + \gamma_0) \kappa^h \equiv \sigma^2 \gamma_1 \kappa^h.$$

If W_{i_0} corresponds to a window with m probes, each separated by a single step on the grid, then the variance of W_{i_0} can be computed by summing the entries of the covariance matrix $\left(\gamma(|i-j|)\right)_{i,j=1}^{m}$ and dividing by m^2:

$$
\begin{aligned}
\operatorname{Var} W_{i_0} &= \operatorname{Var}\left(\frac{1}{m}\sum_{i=1}^{m} LR_i\right) \\
&= \frac{\sigma^2}{m^2}\left(m\gamma_0 + \sum_{i\neq j}\gamma_1\kappa^{|i-j|}\right) \\
&= \frac{\sigma^2}{m^2}\left(m\gamma_0 + 2\gamma_1\sum_{j=1}^{m-1}(m-j)\kappa^j\right) \\
&= \frac{\sigma^2}{m^2}\left(m\gamma_0 + 2\gamma_1\kappa\frac{m - m\kappa - 1 + \kappa^m}{(1-\kappa)^2}\right) \\
&\equiv \frac{\sigma^2}{m^2}\gamma_2(m).
\end{aligned} \tag{2.16}
$$

(Here and in (2.17) below, we use standard expressions for $\sum_{j=1}^{m-1}\kappa^j$ and $\sum_{j=1}^{m-1} j\kappa^j$, but omit the algebra.) If a small number of gridded positions have been dropped, as is often the case, this can be easily accommodated by removing the corresponding rows and columns from $\big(\gamma(|i-j|)\big)_{i,j=1}^{m}$ before summing. To do this efficiently, observe that the sum of the i^{th} row (or column) of the covariance matrix is given by

$$
\begin{aligned}
\sum_{j=1}^{m}\gamma(|i-j|) &= \gamma(0) + \sum_{j=1}^{i-1}\gamma(j) + \sum_{j=1}^{m-i}\gamma(j) \\
&= \sigma^2\left(\gamma_0 + \gamma_1\kappa\frac{2-\kappa^{i-1}-\kappa^{m-i}}{1-\kappa}\right) \\
&\equiv \sigma^2\gamma_3(m,i).
\end{aligned} \tag{2.17}
$$

When gaps in the grid are sparse, it is simple to subtract a small number of row and column sums from (2.16), taking care to add back intersection points, which are subtracted off twice. When filled-in positions are sparse, on the other hand, $\operatorname{Var} W_{i_0}$ may be computed by adding up a small number of row and column sums, then subtracting the double-counted intersection points. Specifically, if the grid positions within a given window are locally indexed by $i = 1, \ldots, m$ and if $\bar{I}$ denotes the set of indices corresponding to gaps in the tiling,

$$
\begin{aligned}
\operatorname{Var} W_{i_0} &= \frac{\sigma^2}{(m-|\bar{I}|)^2}\left(\gamma_2(m) - 2\sum_{i\in\bar{I}}\gamma_3(m,i) + |\bar{I}|\gamma_0 + 2\sum_{\substack{i,j\in\bar{I}\\ i<j}}\gamma_1\kappa^{|i-j|}\right) \\
&\equiv \frac{\sigma^2}{(m-|\bar{I}|)^2}\gamma_4(m,\bar{I}).
\end{aligned} \tag{2.18}
$$

2.6.3 Assessing statistical significance

In the preceding section, we suggest two possible approaches to assessing the statistical significance of a window-level averaged log ratio which appears to differ substantially from 0. If one is willing to assume that the δ_i in (2.2) are normally distributed, then under a null hypothesis of no nearby enrichment, the W_i are normally distributed with a mean of 0 (see Equation (2.7), with $\phi' = \phi$) and a variance as described above. This variance may be estimated by computing $\hat{\kappa}$, $\hat{\lambda}$, and $\hat{\sigma}$ from a manually selected region which exhibits no apparent enrichment, or, alternatively, in an iterative fashion: by initializing with crude $\hat{\kappa}_0$, $\hat{\lambda}_0$, and $\hat{\sigma}_0$; identifying significantly enriched regions; and then repeating with new estimates $\hat{\kappa}_i$, $\hat{\lambda}_i$, and $\hat{\sigma}_i$ ($i = 1$, 2, etc.) based on regions with no called enrichment. The normal CDF may then be used to assign p-values.

Even if the the distribution of the ϵ_i in (2.1) deviates from log-normal to some degree, averaging—across replicates and across positions within the window—is still likely to make the null distribution of W_i approximately normal. In cases where the observed distribution of the window-level scores appears to deviate from normality substantially, however, we propose a simple alternative, which makes the following less restrictive assumptions:

1. For non-enriched regions, the marginal distribution of the LR_i is symmetric.
2. When position i is in a null region but j is near a binding site, LR_i is stochastically smaller than LR_j.
3. Only a small fraction of the genome exhibits ChIP-induced enrichment.

Under these assumptions, and given $\hat{\kappa}$ and $\hat{\lambda}$ as before, we scale each W_i by $m_i/\sqrt{\gamma_4(m_i, \bar{I}_i)}$, and call the results $\tilde{W}_i$. Null $\tilde{W}_i$ will, up to estimation error, have a common variance of σ^2, but they will not, strictly speaking, have a common distribution: as a result of averaging, they will tend towards normal to a greater or lesser degree depending on the number of positions, and the configuration of these positions, within the corresponding windows. Assuming a common unspecified distribution is still, however, likely to be better than assuming a common *normal* distribution in such cases, so we proceed under this assumption. To obtain a non-parametric estimate of the null distribution of the $\tilde{W}_i$, we estimate the mode ($\hat{M}$) of their distribution (using, for example, Bickel (2002) and Hedges and Shah (2003); or by smoothing to obtain a unimodal continuous distribution, and then identifying the maximum numerically), and then reflect the empirical distribution of all $\tilde{W}_i \leq \hat{M}$ over $\hat{M}$. Specifically, we estimate

$$\hat{F}_0(t;\hat{M}) = \frac{1}{2|\{i : \tilde{W}_i \leq \hat{M}\}|} \sum_{\{i:\tilde{W}_i \leq \hat{M}\}} 1\{\tilde{W}_i \leq t\} + 1\{2\hat{M} - \tilde{W}_i \leq t\}, \tag{2.19}$$

and then assign p-values in the obvious way: $P_i \equiv 1 - \hat{F}_0(\tilde{W}_i;\hat{M})$. Both Gibbons et al. (2005) and Buck et al. (2005) suggest similar procedures for computing p-values, for probe-level scores obtained from ChIP-chip using spotted arrays in the former case, and for window-level scores in the latter. They, however, assume that the common null distribution for their test statistics is normal, and simply estimate its variance from a truncated data set. (Buck et al. (2005) also assume that log ratios within the same window are independent.) Efron (2004) takes a related approach to parameter estimates in a logistic regression model, assuming that null distribution parameters can be obtained from the local behavior of the observed data in a region to which false null hypotheses are not expected to contribute.

Finally, to address massive multiple testing (with the example data presented here, over 3 million null hypotheses are tested), we suggest adjusting p-values to control the False Discovery Rate (FDR), using the method of Storey et al. (2004). There, the authors provide asymptotic results which are valid under fairly general "weak dependence" structures. Although a rigorous demonstration is beyond the present scope, both the spatial correlation in the log ratio process and the additional correlation created by use of a moving average are very local relative to the genome scale, so high-density tiling array data should easily fit within this framework. In Figure 2.12, we check this empirically, by computing adjusted p-values for simulated data which replicate the spatial correlation and true positive signal shape observed in practice. In each of 50 simulated data sets, windows were deemed to be true positives if they partially overlapped any enriched region, and the observed FDR for a significance level α was computed as the ratio of false positives to all positives, when all windows with adjusted p-value below α were called positive. Consistent with the theory of FDR adjustment, individual data sets exhibited oFDR levels both higher and lower than the nominal level; average oFDR over the 50 simulations, however, was nearly indistinguishable from nominal levels.

Software, implemented as a package in R, is available for out-of-memory quantile normalization of the large data files associated with high-density tiling arrays, for computation of the probe- and window-level statistics and the variance adjustment terms required for the $\tilde{W}_i$, and for non-parametric estimation of p-values via a symmetric null assumption.

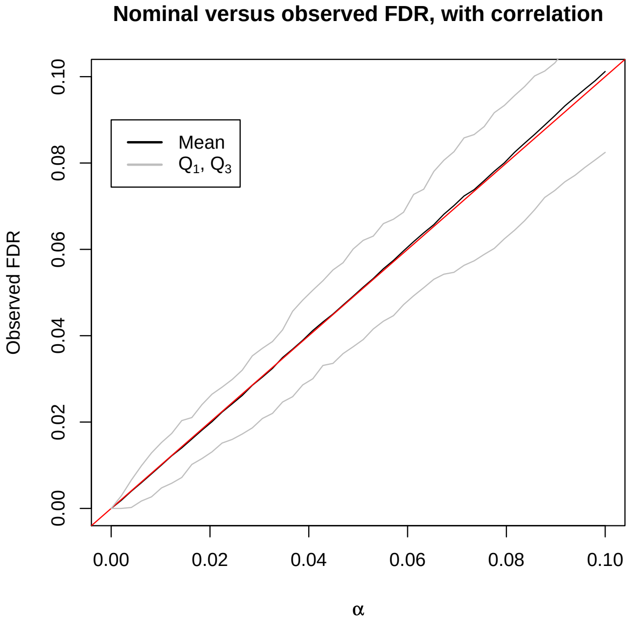

Figure 2.12: To assess the impact of serial correlation on FDR-based p-value adjustment, 50,000 consecutive positions were simulated from an ARMA(1,1) model with parameters $\kappa = .9$ and $\lambda = -.7$, similar to those found for the Zeste data. 125 peaks were added at random, with width of 20 positions and height 5σ. The data were smoothed by a moving window ($w = 5$), window-level variances were adjusted using (2.16) with κ and λ assumed known, p-values were computed using the normal CDF, and finally, adjusted p-values were computed as in Storey et al. (2004). We plot the window-level observed FDR vs. the nominal level, averaged over 50 simulations. The first and third quartiles of the oFDR values associated with each nominal level are also shown. On average, the nominal cutoff for q-values accurately reflected the observed FDR.

2.7 Summary

In this chapter, we have presented two principal results: in Section 2.4, a generative model for the ChIP-chip assay when high density tiling arrays are used, and in Section 2.6, a window-level test statistic with tractable statistical properties. The generative model formalizes the intuition of numerous other authors with respect to signal shape in the neighborhood of a point-like binding event. Moreover, it points to the existence of serial correlation in the probe-level statistics—even away from binding events. While this serial correlation is readily detectable in many data sets, it has not, to the best of our knowledge, been previously pointed out. An understanding of this serial correlation, however, permits us to estimate the variance of the window-level statistics, and to thereby assign correct p-values and establish statistically justified detection thresholds.

Regrettably, the ability to detect windows associated with statistically significant enrichment does not completely solve the problem of binding site detection. To make this more concrete, it is useful to conceptualize the data at four distinct levels: *probe*, *window*, *interval*, and *event*. The first two levels have already been addressed in detail: the LR_i and W_i presented in this chapter are probe- and interval-level statistics, respectively. In practice, though, one typically finds extended intervals over which all window-level statistics exhibit statistically significant enrichment—in part due to probe-level spatial correlation, and in part due to the very nature of a moving window. In the first or second panels of Figure 2.8, for example, windows centered on any position within the first or second peak, or the trough between them, produce significant adjusted p-values if a relatively large bandwidth is used. In these example, the detected intervals actually each contain two events; Schwartz et al. (2006) provide more examples of large intervals which, upon visual inspection, clearly contain multiple events.

The method of Section 2.6 provides a basis for error rate estimation at the window level; the biologist, however, is typically interested in error rate estimation at the event level. Assessing statistical significance at the event level, however, is not straightforward: the events themselves have random boundaries, and are not even defined until a threshold has already been selected. Further, their number and nature typically interact with the detection threshold in an awkward way: raising the threshold creates drop-outs in what had previously been a continuous interval of significant window-level statistics—thereby producing *more* called intervals rather than less. How are we to decide if the resulting sub-intervals are associated with a single event, or with multiple events? We postpone further discussion, and speculation on possible solutions, until the Conclusion.

2.8 Derivations

2.8.1 Equation 2.4

Expansion and term collection shows that

$$\begin{aligned}\mathbb{E}F &= \sum_{m=0}^{L} \frac{L}{m+1}\binom{L}{m}\theta^m(1-\theta)^{L-m} \\ &= \frac{1}{\theta}\frac{L}{L+1}\big(1-(1-\theta)^{L+1}\big).\end{aligned}$$

Since L is the length of a DNA strand in bases, the second ratio in this expression is essentially 1. The final term in parenthesis is also essentially 1: since $\log(1-\theta) < -\theta$ for all θ,

$$\begin{aligned}(1-\theta)^L &< e^{-\theta L} \\ &\to 0\end{aligned}$$

exponentially fast as θL becomes large.

2.8.2 Equation 2.7

From (2.2),

$$\mathbb{E}LR_i = \mathbb{E}\log A_{i1}^T - \mathbb{E}\log A_{i1}^C + \mathbb{E}\delta_i.$$

Consider the first term on the right, and suppress subscripts and superscripts.

$$\begin{aligned}\mathbb{E}\log A &= \mathbb{E}\left(\log\frac{A}{\mathbb{E}A} + \log\mathbb{E}A\right) \\ &= \mathbb{E}\left(\frac{A}{\mathbb{E}A} - 1 + o_P\left(\left|\frac{A}{\mathbb{E}A}-1\right|\right) + \log\mathbb{E}A\right) \qquad (2.20) \\ &= \mathbb{E}o_P\left(\left|\frac{A}{\mathbb{E}A}-1\right|\right) + \log\mathbb{E}A,\end{aligned}$$

by first-order expansion of $\log x$ around $x = 1$. Note that

$$\begin{aligned}\frac{A}{\mathbb{E}A} &= \frac{\frac{1}{N}\sum_{n=1}^{N} X_{in}Z_{in}}{\pi(\Delta)\mathbb{E}Z} \\ &< \frac{z^*}{\pi(\Delta)\mathbb{E}Z}\end{aligned} \qquad (2.21)$$

where z^* is the maximum amplification multiplier for any one fragment: 2^k for k cycles of PCR, or a smaller constant for other amplification schemes. Thus if $A/\mathbb{E}A$ converges in probability to 1 as the number of input DNA strands grows, the Bounded Convergence

Theorem implies that the first term of (2.20) vanishes as N grows. Such convergence in probability follows from application of the Weak Law of Large Numbers to the first line of (2.21).

Thus

$$\begin{aligned}\mathbb{E}LR_i &\approx \log \mathbb{E}A_{i1}^T - \log \mathbb{E}A_{i1}^C + \mathbb{E}\delta_i \\ &= \log\big(N^T \pi^T(\Delta_i)\mathbb{E}Z^T\big) - \log\big(N^C \phi^C \mathbb{E}Z^C\big) + \mathbb{E}\delta_i \\ &= \log\big((1-\theta)^{\Delta_i}(\phi'^T/\phi^T - 1) + 1\big) + K,\end{aligned}$$

where the constant K does not depend on Δ_i or ϕ'^T. (These are the only two parameters that vary from binding site to binding site; all others—including ϕ^T—are experiment-wide constants under the assay model.)

2.8.3 Equation 2.8

First observe that far from any binding site (i.e., with $\pi(\Delta)$ and thus $\mathbb{E}X$ effectively equal to ϕ), Equation 2.3 gives

$$\begin{aligned}\operatorname{Var} A &= N \operatorname{Var} XZ \\ &= N\big(\mathbb{E}(XZ)^2 - (\mathbb{E}X)^2(\mathbb{E}Z)^2\big) \\ &= N\big(\mathbb{E}X(\operatorname{Var} Z + (\mathbb{E}Z)^2) - (\mathbb{E}X)^2(\mathbb{E}Z)^2\big) \\ &= N\big(\phi \operatorname{Var} Z + \phi(1-\phi)(\mathbb{E}Z)^2\big).\end{aligned}$$

Now, also using (2.3)

$$\begin{aligned}\operatorname{Cov}(A_i, A_j) &= \sum_{m,n} \operatorname{Cov}(X_{im}Z_{im}, X_{jn}Z_{jn}) \\ &= \sum_{n} \operatorname{Cov}(X_{in}Z_{in}, X_{jn}Z_{jn}) \qquad (2.22) \\ &= N\big(\mathbb{E}X_iZ_iX_jZ_j - \phi^2(\mathbb{E}Z)^2\big).\end{aligned}$$

Next consider $X_iZ_iX_jZ_j$. For a given input strand, positions i and j end up on the same fragment with probability $(1-\theta)^d$ and in this case, X_i is exactly X_j—either both positions pass or neither does—and $Z_i = Z_j$ as well. On the other hand, i and j are separated by a break with probability $1-(1-\theta)^d$, in which case the two members of each variable pair are distinct and independent. Thus

$$\mathbb{E}X_iZ_iX_jZ_j = \phi\mathbb{E}Z^2(1-\theta)^d + \phi^2(\mathbb{E}Z)^2\big(1-(1-\theta)^d\big).$$

Substituting this expression back into (2.22) and simplifying, we obtain $\operatorname{Cov}(A_i, A_j) = (1-\theta)^d \operatorname{Var} A$, and so $\operatorname{Corr}(A_i, A_j) = (1-\theta)^d$.

2.8.4 Equation 2.9

We rely on the same approximation to $\log A$ given in (2.20). Using the expansion given in (2.2), the various independence relationships implied by the model, and our results for $\mathrm{Cov}(A_i, A_j)$ in the previous section,

$$\begin{aligned}
\mathrm{Cov}(LR_i, LR_j) &= \frac{1}{n^T}\,\mathrm{Cov}(\log A_i^T, \log A_j^T) + \frac{1}{n^C}\,\mathrm{Cov}(\log A_i^C, \log A_j^C) \\
&\approx \frac{1}{n^T}\,\mathrm{Cov}\left(\frac{A_i^T}{\mathbb{E}A_i^T}, \frac{A_j^T}{\mathbb{E}A_j^T}\right) + \frac{1}{n^C}\,\mathrm{Cov}\left(\frac{A_i^C}{\mathbb{E}A_i^C}, \frac{A_j^C}{\mathbb{E}A_j^C}\right) \\
&= \frac{1}{n^T}\frac{(1-\theta)^d\,\mathrm{Var}\,A^T}{\mathbb{E}A_i^T\mathbb{E}A_j^T} + \frac{1}{n^C}\frac{(1-\theta)^d\,\mathrm{Var}\,A^C}{\mathbb{E}A_i^C\mathbb{E}A_j^C} \\
&\approx (1-\theta)^d\left(\frac{\mathrm{Var}(\log A^T)}{n^T} + \frac{\mathrm{Var}(\log A^C)}{n^C}\right).
\end{aligned} \tag{2.23}$$

It is easy to see from the definitions that

$$\mathrm{Var}\,LR_i = \mathrm{Var}\,LR_j = \frac{\mathrm{Var}(\log A^T)}{n^T} + \frac{\mathrm{Var}(\log A^C)}{n^C} + \mathrm{Var}\,\delta,$$

and combining this with (2.23) gives the desired result.

2.8.5 Equation 2.12

Denoting the stacked identity matrices of (2.11) as $\mathbf{X}_0$, and the final column of the design matrix as $\mathbf{X}_1$, we may obtain $\hat{\beta}$ in two steps: (i) regress $\mathbf{X}_1$ on $\mathbf{X}_0$ to obtain a residual vector $\tilde{\mathbf{X}}_1$, then (ii) regress $\mathbf{Y}$ on $\tilde{\mathbf{X}}_1$. First,

$$\begin{aligned}
\tilde{\mathbf{X}}_1 &= \mathbf{X}_1 - \mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{X}_1 \\
&= \mathbf{X}_1 - n^{-1}\mathbf{X}_0\mathbf{X}_0'\mathbf{X}_1 \\
&= \mathbf{X}_1 - (n^T/n)\mathbf{1}_{mn\times 1}.
\end{aligned}$$

For the second step,

$$\begin{aligned}
\hat{\beta} &= \frac{\tilde{\mathbf{X}}_1'\mathbf{Y}}{\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1} \\
&= \frac{\sum_i\sum_j Y_{ij}^T - (n^T/n)(\sum_i\sum_j Y_{ij}^C + \sum_i\sum_j Y_{ij}^T)}{(n^T m) - 2(n^T/n)(n^T m) + (n^T/n)^2(nm)} \\
&= \frac{(n^C/n)(\sum_i\sum_j Y_{ij}^T) - (n^T/n)(\sum_i\sum_j Y_{ij}^C)}{(m/n)n^T n^C} \\
&= \frac{1}{m}\sum_i\left(\frac{1}{n_T}\sum_j Y_{ij}^T - \frac{1}{n_C}\sum_j Y_{ij}^C\right).
\end{aligned}$$

2.8.6 Equation 2.15

Observe that

$$\left(1+\frac{\operatorname{Var}\delta}{\operatorname{Var}(\log A^T)/n^T+\operatorname{Var}(\log A^C)/n^C}\right)^{-1}\in(0,1).$$

If κ is selected to equal $1-\theta$, then $\kappa\in(0,1)$ as well, so to achieve equality in (2.15), we must select $\lambda<0$, so that $(\lambda^{-1}+\lambda+2\kappa)^{-1}$ is negative. Defining

$$f(\lambda;\kappa)=\frac{1-\kappa^2}{\kappa}(\lambda^{-1}+\lambda+2\kappa)^{-1},$$

observe that for fixed $\kappa\in(0,1)$, $f(\lambda;\kappa)\to 0$ as $\lambda\to 0$, and that

$$\begin{aligned} f(-1;\kappa) &= -\frac{1+\kappa}{2\kappa} \\ &< -1. \end{aligned}$$

Because f is continuous in λ away from 0, there must therefore exist a $\lambda\in(-1,0)$ which makes $1+f(\lambda;\kappa)$ equal to any value we please in $(0,1)$.

Chapter 3

Pseudo-ROC

3.1 The receiver operating characteristic curve

3.1.1 Introduction and definitions

In the simplest possible setting for making decisions, a system under consideration may be in one of only two states (*positive/negative*, or *present/absent*, etc.), and we seek to determine this true state based on observable evidence. Our decision-making procedure may correctly identify the system's state, or it may—depending on this true state—produce one of two types of error. A *false positive* occurs when the system's true state is negative but our procedure deems it to be positive; a *false negative* occurs when the true state is positive but we deem it negative. The terminology is general and applies to non-probabilistic settings, but is especially familiar in statistics: in the Neyman-Pearson framework for statistical hypothesis testing, false positives and false negatives correspond to Type I and Type II errors, respectively, and one seeks to minimize the probability of the later while keeping the probability of the former below some known, acceptable level. In this more formal, probabilistic context, let X be a random indicator variable which represents our decision and whose distribution depends on the system's state, D; we may then further define the *sensitivity* and *specificity* of our procedure:

$$\mathrm{Sn}_X \equiv \mathbb{P}(X = 1|D = 1) \qquad\qquad \mathrm{Sp}_X \equiv \mathbb{P}(X = 0|D = 0). \tag{3.1}$$

Thus, $1 - \mathrm{Sp}_X$ is the false positive rate; sensitivity, on the other hand, is equivalent to *power* in the Neyman-Pearson framework, and $1 - \mathrm{Sn}_X$ is the false negative rate.[1]

[1] Conceptually, D need not always be a random variable: the system's state may be fixed and invariant, although unknown, and X may simply follow one of two possible candidate distributions.

If we have two competing procedures, with indicators $X^{(1)}$ and $X^{(2)}$, it is natural to ask which is superior. If one dominates the other in both sensitivity and specificity, this procedure is clearly the winner (assuming, of course, that the two are equally expensive, invasive, time-consuming to carry out, etc.). If, however, one is more specific but the other is more sensitive, the question is ill-posed: its answer depends on a subjective quantification of the harm done by the two types of errors (i.e., a loss function) and the probability that one rather than the other occurs.

When the dichotomous procedures are based on underlying, continuous-valued scores, this ambiguity may sometimes be resolved without a loss function. Suppose that our $X^{(1)}$ and $X^{(2)}$ are in fact based on a common, continuous random variable W to which two different thresholds, $t^{(1)} < t^{(2)}$, have been applied. (We assume, without loss of generality, that large values of W are associated with $D = 1$.) If $W|_{D=0} \sim F$, $W|_{D=1} \sim G$, and both F and G are strictly increasing, the situation described in the preceding paragraph must arise: $X^{(1)}$ is more sensitive, but $X^{(2)}$ is more specific. Suppose, on the other hand, that $X^{(1)}$ and $X^{(2)}$ arise by thresholding distinct $W^{(1)}$ and $W^{(2)}$, with $W^{(1)}|_{D=0}$ and $W^{(2)}|_{D=0}$ both distributed as F, but $W^{(1)}|_{D=1} \sim G^{(1)}$ whereas $W^{(2)}|_{D=1} \sim G^{(2)}$. Suppose further that $W^{(1)}$ is stochastically larger than $W^{(2)}$, i.e., for all t, $G^{(1)}(t) < G^{(2)}(t)$. If one uses $t^{(1)} < t^{(2)}$ as before, a comparison between procedures based on $X^{(1)}$ and $X^{(2)}$ is still ambiguous: again, $X^{(1)}$ is more sensitive, but $X^{(2)}$ is more specific. The use of different thresholds, however, is unnatural here; it seems “fairer” to use thresholds which yield comparable specificity in both cases. Using $t^{(1)} = t^{(2)} \equiv t$ yields common specificity, $F(t)$, but greater sensitivity for the procedure based on $W^{(1)}$—regardless of the choice of t. If we define ℓ_{rs} as the loss associated with calling “s” when $D = r$, and define $\pi \equiv \mathbb{P}(D = 1)$, it is easy to show that the risk of a procedure based on X is given by

$$(1-\pi)\big(\ell_{01} + \mathrm{Sp}_X \times (\ell_{00} - \ell_{01})\big) + \pi\big(\ell_{10} + \mathrm{Sn}_X \times (\ell_{11} - \ell_{10})\big). \tag{3.2}$$

Any sensible loss function penalizes errors relative to correct answers for a given value of D, so both $\ell_{00} - \ell_{01}$ and $\ell_{11} - \ell_{10}$ may be assumed to be negative. Thus, increasing either sensitivity or specificity reduces risk, and in the present case, $W^{(1)}$ improves on $W^{(2)}$ for *any* such loss function.

In such cases, it is still notationally convenient to write $\mathbb{P}(X = 1|D = 1)$ instead of, for example, $\mathbb{P}_{D=1}(X = 1)$.

3.1.2 Calibrated classification points

A fair comparison might also be achieved by selecting thresholds which yield comparable sensitivity and then comparing the procedures based on their specificity. Venkatraman and Begg (1996) suggest a third approach based on "calibrated" classification points which removes the asymmetry.

The marginal distribution of $W^{(j)}$ is given by

$$M^{(j)} = (1-\pi)F^{(j)} + \pi G^{(j)}. \tag{3.3}$$

A calibrated classification point for any $q \in (0,1)$ is then defined by setting $t^{(j)}$ to a q^{th} quantile of $M^{(j)}$. By definition, $M^{(1)}(t^{(1)}) = M^{(2)}(t^{(2)}) = q$, so combining the right-hand side of (3.3) for $i = 1$ and $i = 2$ gives

$$(1-\pi)\big(F^{(1)}(t^{(1)}) - F^{(2)}(t^{(2)})\big) = \pi\big(G^{(2)}(t^{(2)}) - G^{(1)}(t^{(1)})\big),$$

or equivalently,

$$(1-\pi)(\mathrm{Sp}_{X^{(1)}} - \mathrm{Sp}_{X^{(2)}}) = \pi(\mathrm{Sn}_{X^{(1)}} - \mathrm{Sn}_{X^{(2)}}). \tag{3.4}$$

In other words, when using any such calibrated threshold pair, the procedure based on $X^{(1)}$ is more specific if and only if it is also more sensitive. Further, only calibrated threshold pairs can produce procedures with equal sensitivity and specificity: if $F^{(1)}(t^{(1)}) = F^{(2)}(t^{(2)})$ and $G^{(1)}(t^{(1)}) = G^{(2)}(t^{(2)})$ as must be the case for such a pair, then (3.3) implies that $M^{(1)}(t^{(1)}) = M^{(2)}(t^{(2)})$, i.e., $t^{(1)}$ and $t^{(2)}$ form a calibrated pair.

The calibrated classification points, like the sensitivities and specificities of the preceding section, are theoretical quantities whose calculation requires information which is not typically available in practice: the prevalence rate, π, and the conditional distribution functions of W. We shall see in Section 3.4.1, however, that they provide a convenient basis for testing the equivalence of two procedures, and that empirical estimates are easily obtained.

3.1.3 Properties of the ROC curve

These three approaches to selecting threshold pairs for "fair" comparisons between $W^{(1)}$ and $W^{(2)}$ are easily visualized with receiver operating characteristic (ROC) curves. As described above, a single continuous variable W induces a family of dichotomous procedures $\{X_t : t \in [-\infty, \infty]\}$, where X_t denotes the indicator for $\{W > t\}$. The ROC curve is then the parametric plot given by

$$\begin{aligned} \mathrm{ROC}_W &\equiv \big\{(1-\mathrm{Sp}_{X_t}, \mathrm{Sn}_{X_t}) : t \in [-\infty,\infty]\big\} \\ &= \big\{\big(1-F(t), 1-G(t)\big) : t \in [-\infty,\infty]\big\}. \end{aligned} \tag{3.5}$$

If we define $F^{-1}(p) \equiv \inf\{t : F(t) > p\}$ as usual, the ROC curve may also be expressed as

$$\mathrm{ROC}_W = \big\{\big(p, 1 - G \circ F^{-1}(1-p)\big) : p \in [0,1]\big\}. \tag{3.6}$$

If there are intervals over which $F(t)$ is constant but $G(t)$ is not, the parametric form will produce a graph with vertical segments, and thus not represent a function. In this case, the functional form of (3.6) will have jumps instead, and be right-continuous. Choosing the uppermost point to represent such vertical segments makes sense: this point corresponds the the greatest sensitivity attainable at that specificity level.

It follows from (3.6) that the ROC curve is non-decreasing, and is strictly increasing if G is strictly increasing. (F^{-1} is always strictly increasing for continuous variables.) Further, the ROC curve is invariant to monotone-increasing transformations of the data: if $\tilde{W} \equiv \Phi(W)$ for some such transformation Φ, then $\tilde{F} = F \circ \Phi^{-1}$ and $\tilde{G} = G \circ \Phi^{-1}$, and the ordinate in (3.6) becomes $1 - G \circ \Phi^{-1}\big(\Phi \circ F^{-1}(1-p)\big)$. With "inverse" defined as above, $\Phi^{-1} \circ \Phi$ is the identity (although, if Φ is discontinuous, $\Phi \circ \Phi^{-1}$ will not be).

Figure 3.1 provides examples of ROC curves for continuous $W^{(1)}$ and $W^{(2)}$, and illustrates the three methods of selecting threshold pairs—specificity-matched, sensitivity-matched, or calibrated. Obviously, the specificity-matched and sensitivity-matched pairs correspond to points on the curves joined by vertical and horizontal segments, respectively. The calibrated pairs, on the other hand, correspond to points joined by a diagonal segment instead. For any procedure,

$$\begin{aligned} M(t_q) = q \qquad &\Leftrightarrow \qquad (1-\pi)F(t_q) + \pi G(t_q) = q \\ &\Leftrightarrow \qquad 1 - G(t_q) = -\frac{1-\pi}{\pi}\big(1 - F(t_q)\big) + \frac{1-q}{\pi} \\ &\Leftrightarrow \qquad \mathrm{Sn}_q = -\frac{1-\pi}{\pi}(1 - \mathrm{Sp}_q) + \frac{1-q}{\pi} \end{aligned} \tag{3.7}$$

For a given q, therefore, calibrated thresholds produce pairs of points lying on a common line with slope $-(1-\pi)/\pi$—i.e., with slope which does not vary with q—and with an intercept that varies linearly with q. Setting $q = 0$ slices through $(1,1)$, setting $q = 1$ slices through the origin, and moving q from 0 to 1 moves the slice linearly between the

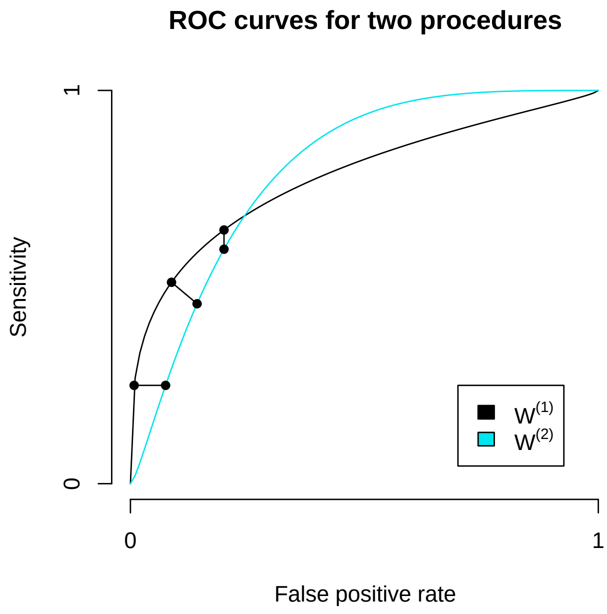

Figure 3.1: Example ROC curves. Three types of "fair" threshold pairs—permitting comparison of the two procedures—are shown. The vertical pair corresponds to matching specificity and comparing sensitivity, and the horizontal pair, vice versa. The diagonal pair is "calibrated" to the .7 quantile of both marginal distributions, with $\pi = .5$ and a slope of -1. By all three methods, $W^{(1)}$ is uniformly superior to $W^{(2)}$ between the origin and the curves' point of intersection.

two. Viewed this way, it is easy to see that the three methods of selecting threshold pairs produce the same points on the curves only at intersections, and that they correspond to a simple visual fact: in a region where $W^{(1)}$ is superior to $W^{(2)}$, its curve can be described as being above, to the left of, or both above and to the left of that of $W^{(2)}$.

In cases where one ROC curve never falls below the other, the method associated with the upper curve matches or improves on the other's sensitivity and specificity for any fair threshold pair. What if the curves cross, however, as shown in Figure 3.1? Traditionally, the area under each curve (AUC) has been used to resolve this issue, and to provide a single-number summary for each curve (Pepe, 2000). Let U and V be independent measurements corresponding to subjects for which $D = 0$ and $D = 1$, respectively. When the ROC curve is estimated from test set data by using empirical distribution functions in (3.5), the area under the curve is, in fact, the normalized Mann-Whitney U-statistic, which estimates $\mathbb{P}(V > U)$ (Hanley and McNeil, 1982); when two empirical ROC curves are to be compared on the basis of their AUC statistics, testing and confidence interval computation may be accomplished via methods for correlated U-statistics (DeLong et al., 1988).

However, comparisons based on AUC suffer from a serious drawback: most of the area comes from the right-hand side of the curves, but these regions are associated with high false positive rates that are not typically tolerated by practitioners. In fact, as shown in Figure 3.1, the curve with the larger AUC may provide uniformly *worse* sensitivity over the acceptable range of false positive rates. A common solution to this problem is to base comparisons on sensitivities at a single false positive rate, or better still, on partial AUCs for which area computed over an acceptable range of false positive rates only. When empirical estimates of two correlated ROC curves are used, inferential techniques for summary statistics based on partial AUCs are also available (Wieand et al., 1989).

3.1.4 Estimation

Both semi-parametric and fully non-parametric methods for estimation of ROC curves, and/or the full or partial area under the curves, are available. As mentioned in the previous section, the simplest approach is to replace the distribution functions in (3.5) with their empirical counterparts. Lloyd (1998), however, suggests that less variable results may be obtained by kernel-based smoothing of the distribution function estimates, and provides formulas for asymptotic variance and bias. The standard semi-parametric approach, on the other hand, assumes that there exists a monotone-increasing function ψ, such that $\psi(U)$ and $\psi(V)$ follow distributions from some location-scale family (Lloyd, 2000). (Because, as shown above, the ROC curve is invariant with respect to such transformations, ψ need not be estimated nor specified.) The classic choice here is the so-called binormal model: if $\psi(U) \sim \mathcal{N}(0,1)$ and $\psi(v) \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\text{ROC}_W = \big\{\big(p, \Phi(\mu + \sigma^{-1}\Phi^{-1}(p))\big) : p \in [0,1]\big\},$$

and the area under the curve is given by $\Phi\big(\mu(1+\sigma^2)^{-1/2}\big)$ (Pepe, 2000). The logistic distribution has sometimes been used in place of the normal; other distribution functions with more appealing theoretical properties exist as well (Lloyd, 2000).

3.2 Gold standards

In many contexts in which one wishes to assess the discriminatory power of a statistical or clinical procedure, or to contrast one procedure against another, a "gold standard" test set does not exist. Even when a reference procedure[2] is available, it is often the case that some putative true positives and true negatives among the test set have been misclassified—due to simple clerical errors, or in many cases, to inherent and unavoidable indeterminacy. In a clinical setting, where definitive determination of disease status may be costly or invasive, this is particularly common. Misclassification is typically not symmetric with respect to the putative true positives and true negatives: one status may be straightforward to establish, but the other, more difficult and subject to a higher error rate. It may also be the case that test set misclassification correlates with the results of the procedure or procedures being evaluated, potentially leading to what has been called "verification bias" in the literature (Begg and Greenes, 1983; Hui and Zhou, 1998; Pepe, 2000; Begg, 2005). Imagine, for example, evaluation of a new procedure based on a biomarker believed to be associated with a certain type of cancer. Subjects thought to be negative based on the reference procedure but who score positively with respect to the marker are likely to undergo additional evaluation and to have their true disease status determined with greater certainty; subjects who were previously thought to be negative and whose marker-based results are also negative, however, will typically not undergo additional screening.

3.2.1 Binary procedures and test set misclassification

Suppose subjects in the test set have been classified as positive or negative for a disease condition based on a reference procedure. We then wish to evaluate a new procedure relative to this reference and produce estimates of the new procedure's sensitivity and specificity, or perhaps to evaluate a set of new procedures against one another. In the simplest case, both the reference and the new procedures produce binary results—with no need for threshold selection—and we may summarize the performance of new procedure relative to the reference in a two-way table (Table 3.1). Let P and R be indicators for the new procedure and reference results, respectively. With perfectly classified test set data, $D_i = R_i$ for each subject i, and a/n_+ and d/n_- provide the natural estimates of sensitivity and specificity for the new procedure.

Suppose, however, that some subjects in the test set have been misclassified. Can sensitivity and specificity estimates for the new procedure—and for the reference procedure as well, since its sensitivity and specificity are no longer assumed to

[2] The term "reference procedure" is sometimes used in the literature for a procedure which establishes definitive disease status. Here and throughout, however, the term refers to the basis for determination of test set classification, which need not be 100% accurate.

	$R_i = 1$	$R_i = 0$	Total
$P_i = 1$	a	b	m_+
$P_i = 0$	c	d	m_-
Total	n_+	n_-	n

Table 3.1: A cross-tabulation comparing a new binary procedure (P) with a reference procedure (R) which splits subjects ($i = 1, \ldots, n$) into two classes: putative true positives (n_+ total subjects), and putative true negatives (n_- total subjects) for a disease condition. With imperfect test set data, $R_i = 1$ need not imply a diseased status ($D_i = 1$) for the subject.

be 100%—be recovered from the data? It is often reasonable to assume that the vector of indicators (P_i, R_i, D_i) is independent (denoted "$\perp\!\!\!\perp$" hereafter) of $(P_{i'}, R_{i'}, D_{i'})$ for $i \neq i'$. Early attempts to address this question made a additional "conditional independence assumption": that P is conditionally independent of R given D, or, when multiple procedures are considered relative to the reference, that $P^{(j)} \perp\!\!\!\perp R|D$ for all j and that $P^{(j)} \perp\!\!\!\perp P^{(j')}|D$ for $j \neq j'$ (e.g., Dawid and Skene, 1979, and de Bock et al., 1994; see discussion in Hui and Zhou, 1998). This additional assumption is typically not realistic (Begg, 1987; Venkatraman and Begg, 1996; Yang and Becker, 1997; Hui and Zhou, 1998), but even with it, the model is not identifiable: the status of a given individual with respect to the new and reference procedures is a multinomial on four possible states, so given the sample size, Table 3.1 contains only three independent cells. The model, however, includes 5 free parameters: letting π denote the prevalence in the population, defining $r_1 \equiv \mathbb{P}(R = 1|D = 1)$, $r_0 \equiv \mathbb{P}(R = 1|D = 0)$, and defining p_1 and p_0 similarly, then

$$\begin{aligned} q_a &= \pi r_1 p_1 + (1-\pi) r_0 p_0 & q_b &= \pi(1-r_1)p_1 + (1-\pi)(1-r_0)p_0 \\ q_c &= \pi r_1 (1-p_1) + (1-\pi) r_0 (1-p_0) & q_d &= 1 - q_a - q_b - q_c. \end{aligned} \tag{3.8}$$

In fact, the model remains unidentifiable even if the prevalence (π) is known.

With extended data sets, however, the model is identifiable. Under the conditional independence assumption, Dawid and Skene (1979) consider two or more new procedures in addition to the reference procedure. Although the authors treat more general circumstances, it is sufficient for illustration purposes to assume that each of the $J > 1$ new procedures is applied to every subject. In this case, the data may be represented as a $(J+1)$-way table of counts for the the joint behavior of R and $P^{(1)}, \ldots, P^{(J)}$. This table has $2^{J+1} - 1$ independent cells given n, which, provided that $J > 1$, are sufficient for estimation of the $2J + 3$ parameters: π, r_1, r_0, $p_1^{(1)}$, $p_0^{(1)}, \ldots, p_1^{(J)}, p_0^{(J)}$. The

	$D = 1$			$D = 0$		
	$R = 1$	$R = 0$	Total	$R = 1$	$R = 0$	Total
$P = 1$	a_1	b_1	$\cdot$	a_0	b_0	$\cdot$
$P = 0$	c_1	$[d_1]$	$\cdot$	c_0	$[d_0]$	$\cdot$
Total	$\cdot$	$\cdot$	$[n_1]$	$\cdot$	$\cdot$	$[n_0]$

Table 3.2: The same comparison shown in Table 3.1, but further disaggregated by true disease status. In the screen-positive setting, true disease status is not established for subjects with $R = 0$ and $P = 0$, so d_1 and d_0 are not observable, nor are n_1 and n_0. The sum $d_1 + d_0$, however, is observable.

parameters here, and in the more general setting as well, may be estimated by treating the true disease status of each subject as a latent variable and applying the EM algorithm.

A second possible extension is to keep $J = 1$, but use samples from two or more populations with distinct prevalence rates and for which the procedures in question are assumed to have a common sensitivity and specificity. For a single new procedure, for example, sampling from two populations yields two independent versions of Table 3.1 with, conditional on the two sample sizes, 6 independent cells. These are now sufficient for estimation of the model's 6 parameters: π_1, π_2, p_1, p_0, r_1, and r_0. When $J = 1$, closed form maximum likelihood estimates are available (Hui and Walter, 1980); when multiple new procedures are contrasted, the EM algorithm may again be used to estimate the various prevalences and the sensitivity and specificity of each procedure under consideration—including the reference (de Bock et al., 1994).

3.2.2 Relative true and false positive rates

As mentioned above, definitive determination of disease status may only be practical or ethical for "screen-positive" subjects—subjects for whom the reference procedure or one of the new procedures under consideration returns a positive result. Suppose, then, that true disease status is known for all screen-positives. Due to incomplete validation of the reference procedure's negative classifications in this case, naive estimates of sensitivity and specificity for the new procedure are subject to verification bias, and the values of these quantities for the reference procedure remain unknown. We have seen that maximum likelihood estimates are possible with extended data sets. Even without such data sets, however, a sensible comparison between the new and reference procedures is still possible—if one is willing to restrict the nature of the question being asked (Schatzkin et al., 1987; Cheng and Macaluso, 1997; Pepe and Alonzo, 2001):

Table 3.2 breaks down the joint behavior of P and R by true disease status. The natural estimate of sensitivity for the new procedure would be $(a_1 + b_1)/n_1$, and for the reference procedure, $(a_1 + c_1)/n_1$. Although a_1, b_1, and c_1 are directly observed, n_1 is not observed because true disease status is not established for $d_0 + d_1$ subjects with a negative score on both procedures. As a consequence, these estimates cannot be computed, and the procedures may not be compared on this basis. Replacing the "1" subscripts with "0" in the forgoing gives false positive rate estimates instead, and the same problem arises.

In some contexts, however, one is only interested in sensitivity or specificity estimates in as much as they permit selection of the superior procedure. In this case, the obvious plug-in estimators for the *relative* sensitivity (rSn) and *relative* false positive rate (rFPR) can be computed using only observable quantities:

$$\widehat{\mathrm{rSn}} = \frac{a_1 + b_1}{a_1 + c_1} \qquad\qquad \widehat{\mathrm{rFPR}} = \frac{a_0 + b_0}{a_0 + c_0}. \tag{3.9}$$

Testing of $H_0\colon$ rSn $= 1$ or $H_0\colon$ rFPR $= 1$ can be accomplished by McNemar's test (Schatzkin et al., 1987), and Cheng and Macaluso (1997) give asymptotic variance estimators which can be used for construction of confidence intervals.

3.3 Pseudo-ROC

3.3.1 Effect of test set misclassification

In this section, we propose a simple ROC-like approach for comparing procedures which generate continuous-valued results, but for which gold standard test sets are not available. This "pseudo-ROC" technique is in the spirit of the relative sensitivity and false positive rate methods of the preceding section—in that comparisons between procedures are possible even if exact quantification of error rates is not.

We have seen in Section 3.1.3 that the traditional ROC curve is a graphical representation of the relationship between F and G—the conditional distribution functions for W when $D = 0$ and $D = 1$, respectively. Of course, F and G are typically unknown in practice and must be estimated. Ideally, a set of observations for which the value of D is known with certainty provides a basis for such estimation. More concretely, assume that this ideal test set consists of $U_1, \ldots, U_m$ i.i.d. as F, and $V_1, \ldots, V_n$ i.i.d. as G. Suppose, however, that some contamination is present in the test set and we are instead presented with $U'_1, \ldots, U'_m$ and $V'_1, \ldots, V'_n$, where U' and V' are distributed as G with probabilities κ and λ, respectively, and as F otherwise. With properly classified test set data, $\kappa = 0$ and $\lambda = 1$; with contamination, $\kappa > 0$ and/or $\lambda < 1$. (We may still assume $\kappa < \lambda$, though, unless the reference procedure is truly pathological.) Under these conditions, U' and V' follow the mixture distributions

$$\begin{aligned} F' &\equiv (1-\kappa)F + \kappa G \\ G' &\equiv (1-\lambda)F + \lambda G. \end{aligned} \tag{3.10}$$

Naive use of the contaminated test set data does not, therefore, lead to an estimate of ROC_W, but rather to one of $\mathrm{ROC}_{W'}$—the ROC curve which describes F' relative to G'. ROC_W are $\mathrm{ROC}_{W'}$ are related to one another, however, in a simple way. Letting $\bar{F}$, for example, denote the survival function $1 - F$,

$$\begin{aligned} \mathrm{ROC}_{W'} &= \left\{ \left(\bar{F}'(t), \bar{G}'(t) \right) : t \in \mathbb{R} \right\} \\ &= \left\{ \left((1-\kappa)\bar{F}(t) + \kappa\bar{G}(t),\ (1-\lambda)\bar{F}(t) + \lambda\bar{G}(t) \right) : t \in \mathbb{R} \right\} \\ &= \left\{ \left(\bar{F}(t), \bar{G}(t) \right) \begin{pmatrix} 1-\kappa & 1-\lambda \\ \kappa & \lambda \end{pmatrix} : t \in \mathbb{R} \right\} \\ &= \left\{ (p,q) \begin{pmatrix} 1-\kappa & 1-\lambda \\ \kappa & \lambda \end{pmatrix} : (p,q) \in \mathrm{ROC}_W \right\}. \end{aligned} \tag{3.11}$$

Thus the pseudo-ROC curve, $\mathrm{ROC}_{W'}$, is just a linear transformation of the true ROC curve for W. This transformation is illustrated in Figure 3.2.

Denote the transformation matrix in (3.11) as M, and observe that M only depends on the misclassification probabilities, not on F and G. As a consequence, if we are interested in comparing two procedures based on a common set of misclassified test data, the same M will describe the transformation of $\mathrm{ROC}_{W^{(1)}}$ to $\mathrm{ROC}_{W'^{(1)}}$ and $\mathrm{ROC}_{W^{(2)}}$ to $\mathrm{ROC}_{W'^{(2)}}$. The common single-number summaries used to score and compare ROC curves—the full or partial AUCs, or the sensitivities at a given false positive rate—are area or distance based, and will therefore be reduced by this transformation, but to the same degree for both curves. Thus, a comparison of $W^{(1)}$ and $W^{(2)}$ using the pseudo-ROC curves will still select the correct procedure as superior. However, we are now using estimated pseudo-ROC curves to detect a reduced effect, but with the same number of observations as would have been the case with proper classification. Obviously, this task is more difficult, and statistical power is lower. Indeed, suppose that the test set data are classified at random, so that $\kappa = \lambda = .5$. In this case, all entries of M equal .5 and the pseudo-ROC curves for $W^{(1)}$ and $W^{(2)}$ both fall along

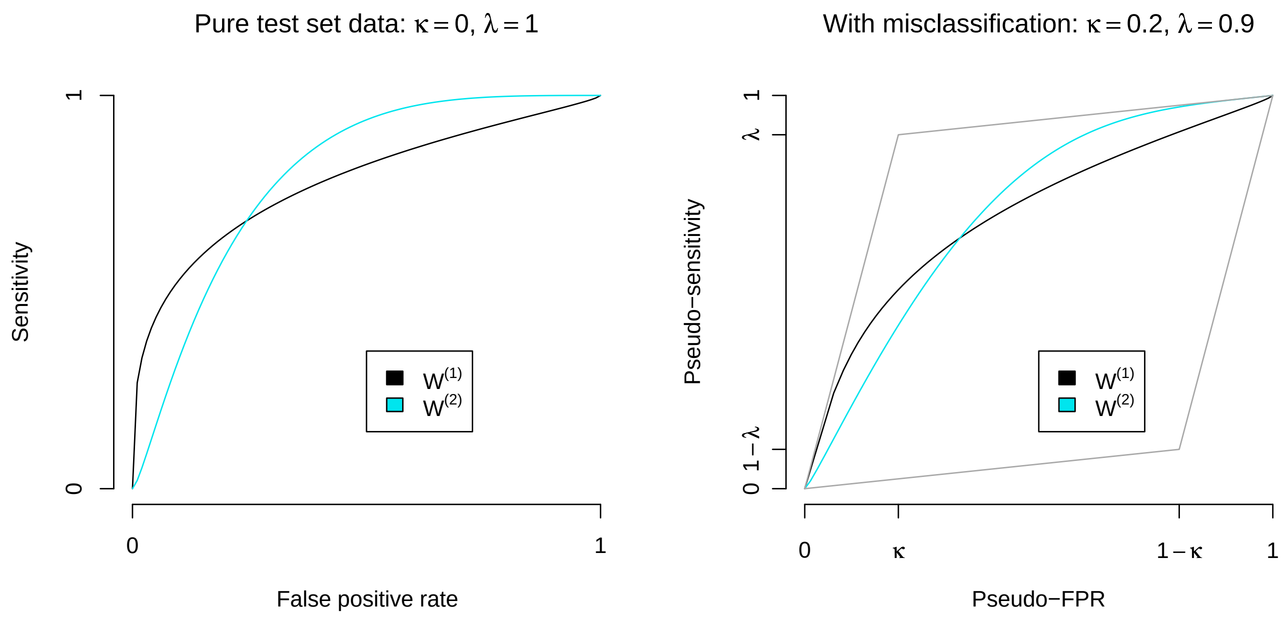

Figure 3.2: The effect of test set misclassification on ROC curves for $W^{(1)}$ and $W^{(2)}$. The "pseudo-ROC" curve for each procedure is a linear transformation of the true ROC curve. The specifics of this transformation are determined by κ and λ, the misclassification probabilities, not by $F^{(1)}$ and $G^{(1)}$ or $F^{(2)}$ and $G^{(2)}$.

the diagonal between $(0,0)$ and $(1,1)$, regardless of the original degree of separation between $\mathrm{ROC}_{W^{(1)}}$ and $\mathrm{ROC}_{W^{(2)}}$.

Note also that the empirical ROC curves—$\widehat{\mathrm{ROC}}_W$, which is unobservable unless $\kappa = 0$ and $\lambda = 1$, and the observable $\widehat{\mathrm{ROC}}_{W'}$—are random graphs; as such, they need *not* be directly related to one another by the same simple relationship. Under mild conditions on F and G, however, both $\widehat{\mathrm{ROC}}_W$ and $\widehat{\mathrm{ROC}}_{W'}$ are strongly consistent estimators of ROC_W and $\mathrm{ROC}_{W'}$ (Hsieh and Turnbull, 1996), so the linear transformation relationship will hold approximately for the empirical curves when m and n are large.

3.3.2 Non-i.i.d. statistics

In the applications discussed in Chapter 4, the test statistics used to estimate pseudo-ROC curves may be reasonably thought of as independent of one another, but will clearly not be identically distributed. The definitions of the ROC and pseudo-ROC curves can, however, be easily extended to accommodate this. Let $\mathcal{F}$ and $\mathcal{G}$ represent families of possible distributions for U and V (the statistics corresponding to test set cases for which $D = 0$ and $D = 1$, respectively). Further, let μ and ν be probability measures on these spaces. If one wishes to consider all further analyses as conditional on the subjects selected for the test set, $\mathcal{F}$ and $\mathcal{G}$ may be finite, μ may simply place mass $1/m$ on each of $F_1, \ldots, F_m$, and ν may place mass $1/n$ on $G_1, \ldots, G_n$; more complicated schemes in which test set subject selection is actually thought of as random can also be accommodated. If we then define

$$F(t) \equiv \int_{\mathcal{F}} \tilde{F}(t)\ d\mu(\tilde{F}) \qquad\qquad G(t) \equiv \int_{\mathcal{G}} \tilde{G}(t)\ d\nu(\tilde{G}), \tag{3.12}$$

the ROC curve can then be defined exactly as in (3.5).

An important, additional complication arises in the non-i.i.d. case if the distribution of U and/or V depends on a subject's misclassification status. Suppose, for instance, that misclassified true negatives (subjects which are true negatives, but which have been deemed positive by the reference procedure) are more likely to have distribution functions drawn from some subset of $\mathcal{F}$ than correctly classified true negatives. In this case, we actually have μ_+ which more heavily weights this subset, with $\mu_- \neq \mu_+$. (The subscript here corresponds to the reference procedure's call.) The same may hold for misclassified vs. correctly classified true positives, yielding $\nu_- \neq \nu_+$. Let $F_-(t) \equiv \int_{\mathcal{F}} \tilde{F}(t)\ d\mu_-(\tilde{F})$, and define F_+, G_-, and G_+ similarly. If perfect test set classification is possible, then $\kappa = 0$ and $\lambda = 1$, so $\text{ROC}_W \equiv \big\{(\bar{F}_-(t), \bar{G}_+(t)) : t \in [-\infty, \infty]\big\}$. The marginal distribution functions for the contaminated test set data, however, are now given by

$$\begin{aligned} F' &\equiv (1-\kappa)F_- + \kappa G_- \\ G' &\equiv (1-\lambda)F_+ + \lambda G_+, \end{aligned} \tag{3.13}$$

which complicates the relationship between ROC_W and $\text{ROC}_{W'}$. Suppressing dependence on t,

$$\begin{aligned} \text{ROC}_{W'} &= \big\{(\bar{F}', \bar{G}')\big\} \\ &= \big\{\big((1-\kappa)\bar{F}_- + \kappa\bar{G}_-,\ (1-\lambda)\bar{F}_+ + \lambda\bar{G}_+\big)\big\} \\ &= \left\{ (\bar{F}_-, \bar{G}_+) \begin{pmatrix} 1-\kappa & 1-\lambda \\ \kappa & \lambda \end{pmatrix} + \big(\kappa(G_+ - G_-), (1-\lambda)(F_- - F_+)\big) \right\}. \end{aligned} \tag{3.14}$$

The left-hand term in the last line of (3.14) is as before: a linear transform (again by M) of the correct ROC curve. The right-hand term, however, may lead to additional shifts. If $\kappa > 0$, i.e., there is some contamination of the nominally "negative" test set data, and further, if the average distribution function corresponding to true positives inadvertently included among the negatives differs from that of correctly classified true positives, the pseudo-ROC curve will be shifted horizontally away from the image of ROC_W under M. If $\lambda < 1$ and the average distribution functions for correctly and incorrectly classified true negatives differ, there will be vertical shifts as well.

These additional shifts are important for application of the pseudo-ROC method because, unlike the i.i.d. case, the transformation which results from test set misclassification now depend on F and G. When two methods are contrasted, a comparison based on the pseudo-ROC curves may in some cases misrepresent the relationship between the true ROC curves. For example, consider two methods with identical ROC curves, and suppose that for both methods, the behavior of true negatives is independent of test set classification, i.e., $F_-^{(1)} = F_+^{(1)}$ and $F_-^{(2)} = F_+^{(2)}$. If $G_-^{(1)} = G_+^{(1)}$, then $\mathrm{ROC}_{W'^{(1)}} = \mathrm{ROC}_{W^{(1)}} \cdot M$, as for the i.i.d. case. Suppose now that for procedure 2, misclassified true positives tend to return weaker scores than correctly classified true positives, i.e., $G_+^{(2)} \leq G_-^{(2)}$ for all t, with strict inequality for at least some t. (Such behavior is quite natural in many contexts: misclassified true positives may have been misclassified precisely because they are more difficult to detect.) If $\kappa > 0$, then (3.14) implies that the pseudo-ROC curve for $W^{(2)}$ is shifted to the left of $\mathrm{ROC}_{W^{(2)}} \cdot M$, and thus to the left of the pseudo-ROC curve for $W^{(1)}$. As a consequence, we incorrectly infer that procedure 2 is superior to procedure 1.

Restating, a decision based on the pseudo-ROC curves for two procedures will agree with one based on the unobservable true ROC curves if (though not only if)

1. When positive test set contamination exists, the average distribution functions for correctly and incorrectly classified true negatives agree; and

2. When negative test set contamination exists, the average distribution functions for correctly and incorrectly classified true positives agree.

If misclassification of test set subjects takes place at random and independently of W, then these conditions are obviously met. They may also be met under less restrictive conditions: if, as is sometimes the case, the test statistics follow a common distribution under $H_0\colon D = 0$, then $|\mathcal{F}| = 1$ and condition 1 is satisfied. In the applications in Chapter 4, uncontaminated positive test set data are difficult to produce, but uncontaminated (or only very mildly contaminated) negative test set data are more readily available. If $\kappa = 0$, then condition 2 is satisfied as well.

Conditions 1 and 2 can be restated in terms of the conditional independence assumption discussed in Section 3.2. Let R denote the result of the reference procedure as before, and let K be an indicator for misclassification. Conditions 1 and 2 together state that $W \perp\!\!\!\perp K | D$. Because K and R are equivalent given D ($K = R$ when $D = 0$, and $K = 1 - R$ when $D = 1$), conditions 1 and 2 therefore require that W be conditionally independent of the reference procedure given D. When two procedures are compared, however, these conditions do *not* require that $W^{(1)}$ and $W^{(2)}$ be conditionally independent of one another given D. As will be discussed in Chapter 4, conditional independence relative to R may be reasonable to assume, but conditional independence between $W^{(1)}$ and $W^{(2)}$ often is not.

3.4 Statistical significance

3.4.1 A test statistic

Venkatraman and Begg (1996) propose a permutation-based method for testing the equality of the ROC curves associated with two continuous random variables, $W^{(1)}$ and $W^{(2)}$, for which observations are paired—taken from the same set of patients, for example, or as will be the case in Chapter 4, derived from two statistical methods applied to the same raw data. The method's p-values are simple to compute, and it avoids the often unrealistic assumption of conditional independence of $W^{(1)}$ and $W^{(2)}$ given D. We briefly review the details here:

Let $t_q^{(1)}$ and $t_q^{(2)}$ be the calibrated thresholds—the q^{th} quantiles of the marginal distributions—described in Section 3.1.2, and let $X_q^{(i)}$ be the indicator variable resulting from comparison of $W^{(i)}$ to $t^{(i)}$. The following random variable contrasts the performance of the two methods by scoring ±1 if one method is correct but the other is not:

$$\mathcal{E}(q) \equiv \begin{cases} +1 & \text{if } X_q^{(1)} = D \neq X_q^{(2)}, \\ -1 & \text{if } X_q^{(1)} \neq D = X_q^{(2)}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.15}$$

Venkatraman and Begg show that $\mathbb{E}\mathcal{E}(q) = 0$ iff $X_q^{(1)}$ and $X_q^{(2)}$ have equal sensitivity and specificity, so $\mathbb{E}\mathcal{E}(q) = 0$ for all q iff $\text{ROC}^{(1)} = \text{ROC}^{(2)}$. In fact,

$$\begin{aligned}\mathbb{E}\mathcal{E}(q) &= \pi\big(\mathbb{P}(X_q^{(1)}=1, X_q^{(2)}=0|D=1) - \mathbb{P}(X_q^{(1)}=0, X_q^{(2)}=1|D=1)\big) + \\ &\qquad (1-\pi)\big(\mathbb{P}(X_q^{(1)}=0, X_q^{(2)}=1|D=0) - \mathbb{P}(X_q^{(1)}=1, X_q^{(2)}=0|D=0)\big) \\ &= \pi\big(\mathbb{P}(X_q^{(1)}=1|D=1) - \mathbb{P}(X_q^{(2)}=1|D=1)\big) + \\ &\qquad (1-\pi)\big(\mathbb{P}(X_q^{(1)}=0|D=0) - \mathbb{P}(X_q^{(2)}=0|D=0)\big) \\ &= \pi(\mathrm{Sn}_{X_q^{(1)}} - \mathrm{Sn}_{X_q^{(2)}}) + (1-\pi)(\mathrm{Sp}_{X_q^{(1)}} - \mathrm{Sp}_{X_q^{(2)}}) \\ &= 2\pi(\mathrm{Sn}_{X_q^{(1)}} - \mathrm{Sn}_{X_q^{(2)}}),\end{aligned} \tag{3.16}$$

where the last line is due to (3.4). If d_q is the length of the segment joining the points on $\mathrm{ROC}^{(1)}$ and $\mathrm{ROC}^{(2)}$ which correspond to $t^{(1)}$ and $t^{(2)}$, the fact that the slope in (3.7) does not depend on q implies that $d_q \propto |\,\mathrm{Sn}_{X_q^{(1)}} - \mathrm{Sn}_{X_q^{(2)}}\,|$, and thus that $d_q \propto |\mathbb{E}\mathcal{E}(q)|$. We omit details, but it is straightforward to then show that $V \equiv \int_0^1 \big|\mathbb{E}\mathcal{E}(q)\big|\ dq$ is proportional to the area *between* $\mathrm{ROC}^{(1)}$ and $\mathrm{ROC}^{(2)}$—a quantity which, in cases where the curves cross one another, provides a better basis for assessing the equivalence of the curves than $\Delta\,\mathrm{AUC}$.

Because the population quantiles and marginal distributions are not typically available, Venkatraman and Begg propose an empirical estimate of the expectation of $\mathcal{E}(q)$, with empirical quantiles used to set the calibrated threshold pair $t^{(1)}$ and $t^{(2)}$. To do this, let n be the total number of test cases (both positive and negative) and define $X_i(k/n)$ to be the indicator for $W_i > W_{(k)}$, for $i = 1, \ldots, n$. Then define

$$\mathcal{E}_i(k/n) \equiv \begin{cases} +1 & \text{if } X_i^{(1)}(k/n) = D_i \neq X_i^{(2)}(k/n), \\ -1 & \text{if } X_i^{(1)}(k/n) \neq D_i = X_i^{(2)}(k/n), \\ \ \ 0 & \text{otherwise.} \end{cases}$$

and

$$\hat{V} \equiv \sum_{k=1}^{n-1} \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{E}_i(k/n) \right|. \tag{3.17}$$

Venkatraman and Begg's decision to base the calibrated thresholds on marginal quantiles makes computation of $\hat{V}$ particularly simple: $n\hat{\mathbb{E}}\mathcal{E}(k/n)$ is just the difference in total number of correct answers given (or, equivalently, errors made) when each method's top $n-k$ scores are called "positive." Suppressing the k/n arguments,

$$\begin{aligned}
n\hat{\mathbb{E}}\mathcal{E} = {} & \#\{X_i^{(1)} = 1, X_i^{(2)} = 0, D_i = 1\} - \#\{X_i^{(1)} = 0, X_i^{(2)} = 1, D_i = 1\} + \\
& \quad \#\{X_i^{(1)} = 0, X_i^{(2)} = 1, D_i = 0\} - \#\{X_i^{(1)} = 1, X_i^{(2)} = 0, D_i = 0\} \\
= {} & \#\{X_i^{(1)} = 1, X_i^{(2)} = 0, D_i = 1\} + \#\{X_i^{(1)} = 1, X_i^{(2)} = 1, D_i = 1\} - \\
& \quad \#\{X_i^{(1)} = 1, X_i^{(2)} = 1, D_i = 1\} - \#\{X_i^{(1)} = 0, X_i^{(2)} = 1, D_i = 1\} + \\
& \quad \#\{X_i^{(1)} = 0, X_i^{(2)} = 1, D_i = 0\} + \#\{X_i^{(1)} = 0, X_i^{(2)} = 0, D_i = 0\} - \\
& \quad \#\{X_i^{(1)} = 0, X_i^{(2)} = 0, D_i = 0\} - \#\{X_i^{(1)} = 1, X_i^{(2)} = 0, D_i = 0\} \\
= {} & \big(\#\{X_i^{(1)} = 1, D_i = 1\} + \#\{X_i^{(1)} = 0, D_i = 0\}\big) - \\
& \quad \big(\#\{X_i^{(2)} = 1, D_i = 1\} + \#\{X_i^{(2)} = 0, D_i = 0\}\big).
\end{aligned} \tag{3.18}$$

While the first sum in (3.17) incorporates the full range of possible thresholds, it would also be natural to begin at $k = k^* > 1$, providing an analog to the partial AUC.

3.4.2 Assessing significance with paired data

Assessing the statistical significance of an observed $\hat{V}$ may be accomplished if one makes an additional assumption, albeit one much weaker than conditional independence of the two procedures. First, note that in general for continuous X and Y distributed as H_X and H_Y, $H_Y(Y)$ has the uniform distribution on $[0,1]$, and so $X \overset{d}{=} H_X^{-1} \circ H_Y(Y)$. In the present context, if we define $\psi = (M^{(1)})^{-1} \circ M^{(2)}$ then $W^{(1)} \overset{d}{=} \psi(W^{(2)})$. Venkatraman and Begg (1996) go on to point out that if $\text{ROC}^{(1)} = \text{ROC}^{(2)}$ the same must hold conditionally; they then make the further assumption that $W^{(1)}$ and $\psi(W^{(2)})$ are exchangeable under the null hypothesis of equal ROC curves. This exchangeability permits a permutation test, which they go on to show to be consistent against the alternative hypothesis of unequal ROC curves.

The test is straightforward to implement. When $W^{(1)}$ and $W^{(2)}$ are on the same scale, their values may be interchanged at random within subject, and $\hat{V}^{(b)}$ computed for $b = 1, \ldots, B$. When $W^{(1)}$ and $W^{(2)}$ are not directly comparable—as is often the case—the original *ranks* may be exchanged instead, with the numerous ties that result broken at random. Software which carries out this procedure is available in the same R package mentioned in Chapter 2.

3.4.3 Application to pseudo-ROC

Application of this approach to pseudo-ROC testing is trivial. First, the pseudo-ROC curve $\text{ROC}_{W'}$ *is* an ROC curve as well, albeit for the distributions F' and

G', which are not of direct interest. Thus, applying Venkatraman and Begg's method to imperfectly classified test set data tests $H_0\colon\ \mathrm{ROC}_{W'^{(1)}} = \mathrm{ROC}_{W'^{(2)}}$. Because, provided conditions 1 and 2 above are met, the same transformation M is applied to both ROC curves, $\mathrm{ROC}_{W^{(1)}} = \mathrm{ROC}_{W^{(2)}}$ iff $\mathrm{ROC}_{W'^{(1)}} = \mathrm{ROC}_{W'^{(2)}}$.

3.5 Summary

Recently, Hall and Zhou (2003) have proposed a fully non-parametric estimation procedure for the distribution functions of random vectors—of dimension $k \geq 3$—which arise from a mixture of two distributions, each of which has independent components. If we discard test set classification labels completely, let $W^{(i)}$ correspond to the i^{th} component of such a vector, and grant independence of the procedures conditional on D, then this method permits estimation of each procedure's conditional distribution functions as well as the prevalence parameter, π. With consistent estimates of the conditional distribution functions in hand, the *true* ROC curves can be consistently estimated as well. Zhou et al. (2004), for example, implement this method to estimate ROC curves for ordinal scale W.

In this chapter, we have presented a new approach to the problem: pseudo-ROC. While pseudo-ROC does not attempt direct estimation of sensitivities and specificities, it does provide a basis for the comparison of two or more procedure—without the need for a full conditional independence assumption. As discussed in Section 3.3.2, a sufficient condition for the validity of pseudo-ROC comparisons—in the sense that they return the same results as would have been obtained from correctly classified test set data and traditional ROC curves—is conditional independence between each procedure under evaluation and the reference procedure. The procedures under evaluation need not be conditionally independent of one another.

To assess the statistical significance of apparent differences between pseudo-ROC curves, we adopt the hypothesis testing approach of Venkatraman and Begg (1996). This approach is also valid when $W^{(1)}$ and $W^{(2)}$ are correlated, and is therefore suitable for data sets for which the conditional independence assumption is unrealistic. Other authors have proposed methods for the construction of simultaneous confidence bands for ROC curves (Campbell, 1994; Hsieh and Turnbull, 1996), and these can be applied directly to pseudo-ROC curves. While it is true that confidence intervals are generally more informative than testing, we feel that the difference is not so important in the present case: our focus is on selection rather than estimation, and the pseudosensitivities and pseudo-specificities are of interest only in as much as they permit identification of the superior procedure.

Chapter 4

Performance of analysis methods

4.1 Introduction

A variety of methods have recently been suggested for analyzing data obtained from ChIP-chip experiments using high-density tiling arrays. As discussed in the Introduction, Cawley et al. (2004) employed a moving-window Wilcoxon rank sum statistic in one of the first papers to utilize the platform. More recently, windowed probe-level t statistics of various forms have been proposed as an alternative to the Wilcoxon rank sum (Buck et al., 2005; Ji and Wong, 2005; Johnson et al., 2006; Keleş et al., 2006; Schwartz et al., 2006), hidden Markov models (HMM) have been suggested as a replacement for moving windows (Ji and Wong, 2005; Li et al., 2005), and some attempts have been made at model-based approaches (Keleş, 2005; Kim et al., 2005), although the latter have yet to be shown workable for whole genome analyses.

In this chapter, we place the majority of these methods within a single statistical framework, and then compare the methods' performance on real and artificial data sets. The statistical framework provides a basis for understanding how the methods differ, and why some are superior to others. In particular, we will consider three issues related to probe behavior: (i) background correction, (ii) systematic variability in probe response, and (iii) probe-level variance estimation.

In addition to taking different approaches to the statistical issues raised above, the methods we consider vary in how they aggregate data from neighboring positions, and how they set thresholds and assess statistical significance. Differences in spatial aggregation will be discussed in Section 4.3.2, below. Setting thresholds and assessing significance are obviously of great practical importance, and deserve consideration in their own right. Permitting each method to set its own thresholds as it saw fit, however, would confound an examination of the method's effectiveness in addressing probe behavior. To finesse this, we have selected the receiver operating characteristic curve—the traditional form and also the pseudo-ROC variant presented in Chapter 3—as our performance metric. Because a single ROC curve depicts the specificity/sensitivity tradeoff for all possible thresholds simultaneously, no single threshold need be established, and the methods' underlying statistics may be compared on a common footing. (As a consequence, methods whose end result is a list of genomic intervals will have been interrupted half-way in this chapter; we will still refer to these methods by name, but with the understanding that it is the set of scores upon which they base their decisions, and not the final decisions themselves, which are under consideration.)

Traditionally, quantitative PCR has been used as the gold standard for array-based enrichment detection (e.g., Johnson et al. (2006)). While qPCR provides relatively (though not completely) unambiguous confirmation, its application is typically restricted to a few dozen sites. Given this, we suggest that the ROC or pseudo-ROC metric, as implemented below, provides a valuable complement to qPCR—one which simultaneously includes hundreds or thousands of positions, and which takes the same genome-wide perspective as the ChIP-chip assay itself.

4.2 Issues related to probe behavior

Before examining the analysis methods in detail, we review three statistical issues, related to probe behavior, which impact results.

4.2.1 Background correction and GC bias

In Chapter 2, we briefly mentioned additive background and its contribution to observed fluorescence intensity, but then set the issue aside; here, we give it more careful consideration. Numerous authors have shown that proper background correction improves the analysis of gene expression with high-density microarrays, mostly by reducing bias in fold change estimation for targets present at low concentrations (Huber et al., 2002; Irizarry et al., 2003; Wu et al., 2004; Affymetrix, Inc., 2005). While they take different approaches to estimation and make different assumptions about the joint distribution of background and error terms, these authors' models can be expressed in a common form:

$$I_{ij} = \alpha_i A_{ij} \epsilon_{ij} + B_{ij}. \tag{4.1}$$

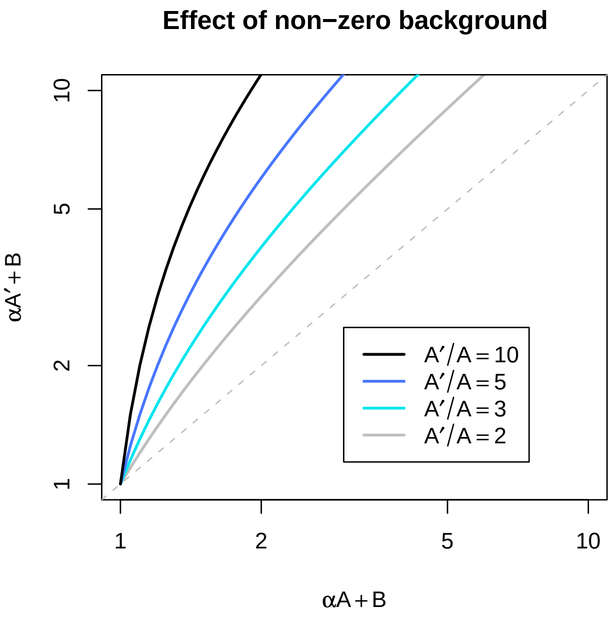

Figure 4.1: To demonstrate the effect of additive background on fold change estimation in the additive background, multiplicative error model, we consider an error free context ($\epsilon \equiv 1$) with constant background ($B \equiv 1$) for all probes, so that observed intensities in both treatment and control are only a function of probe response (α) and target abundance (A' and A, for treatment and control respectively). Plots over a range of α values are shown for various ratios of target abundance between treatment and control. Although the target abundance ratio is fixed for each graph, the perceived fold change—the shift off the diagonal for a log-log plot—varies with probe responsiveness.

As before, the observed fluorescence intensity recorded for probe i on array j is denoted by I_{ij}, the responsiveness of probe i, by α_i, the underlying target abundance, by A_{ij}, and the non-negative multiplicative error, by ϵ_{ij}. This “additive background, multiplicative error” model augments the simpler model of (2.1) with an additional term, denoted B_{ij}, for the combined additive effects of optical noise and non-specific binding. Since nothing about the additive background, multiplicative error model is specific to gene expression applications, we use it as a guide for understanding ChIP-chip data as well.

Ignoring the ϵ noise terms for the moment, if we assume, as in Chapter 2, that that there is no additive background, then the common α_i terms cancel and the ratio of intensities observed for two arrays exactly equals the ratio of target abundances. The presence of positive background complicates this relationship, however, even in the absence of noise. Figure 4.1, for example, assumes constant, non-zero additive background, and plots treatment intensity against control intensity for a range of probe response values and for four different target abundance ratios, demonstrating the degree to which the perceived fold change (the shift off the diagonal for a log-log plot) depends on probe responsiveness. Unresponsive probes with α near 0 appear in the lower left, generating similar treatment and control intensities and a ratio which is biased toward 1; highly responsive probes, for which αA dominates B, appear toward the right and generate observed intensity ratios which approximately match the true ratio of target abundances.

Figures 4.2, based on a BAC spike-in experiment using Affymetrix *D. melanogaster* tiling arrays (described in detail below), shows this effect in practice. Data from two treatment and two control replicates were quantile normalized (Bolstad et al., 2003), but no background correction was applied. In the genomic regions considered, true target abundance was held constant by design in both treatment and control samples, with a known ratio between the two conditions. Figure 4.2 shows that

1. Measured intensity varied widely nonetheless, for both the treatment and control samples;

2. The perceived fold change was too low, even for the most responsive probes (those with highest measured intensity on the control arrays); and

3. The perceived fold change was not constant as expected, but rather varied as a function of measured control intensity.

This last point is explored more carefully in Figure 4.3A, in which we plot the median perceived fold change over all probes with common GC content. It is well-known that GC content serves as a rough proxy for probe responsiveness (Naef et al., 2002; Zhang et al., 2003; Wu et al., 2004). As in Figure 4.2, the medians are not constant as expected, but instead depend on the probes' GC base count; AT-rich *and* GC-rich probes tend to badly under-estimate the true fold change. Median perceived fold change is relatively stable over probes whose GC content falls within the 10 to 14 base range, but restriction to this subset is not compatible with the spatial requirements of a tiling (Mockler et al., 2005).

Should we be concerned about bias in fold change estimation in the ChIP-chip context? While fold change estimates for target abundance have a literal and biologically relevant interpretation in expression analysis, it is not clear what such numbers mean for ChIP-chip: target abundance in the immunoprecipitated and amplified DNA hybridized

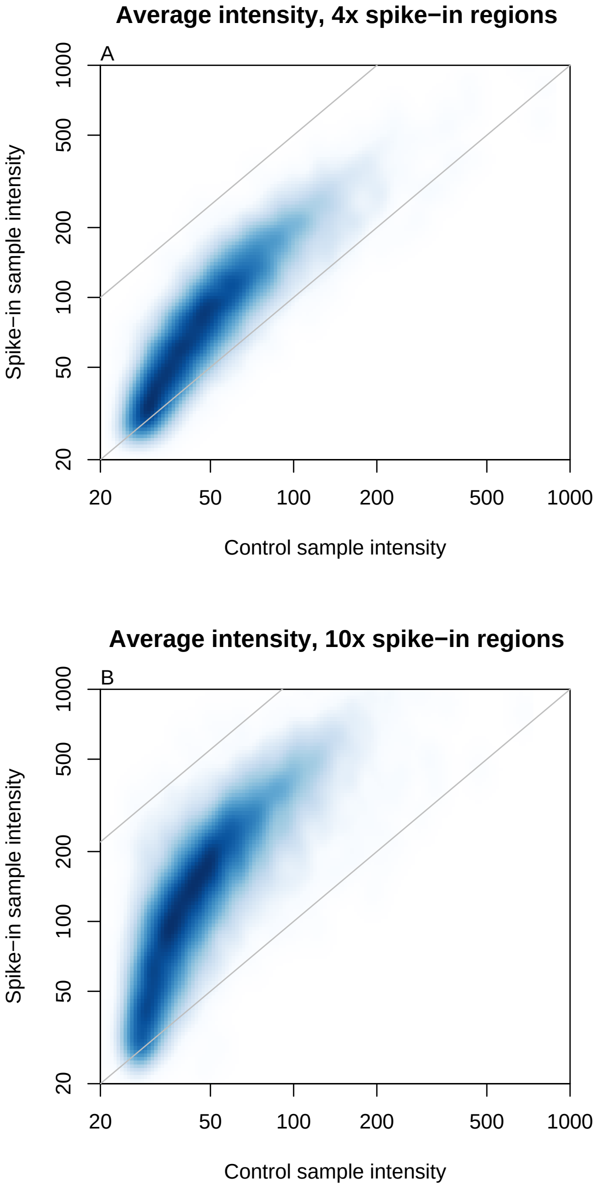

Figure 4.2: The scatter plots (darker colors represent higher frequencies) compare observed PM probe intensities (averaged over two replicates) obtained from spike-in samples—to which BAC DNA from known genomic regions was added at known concentrations—to those obtained from input control samples. Arrays were quantile normalized, but no other preprocessing was applied. The gray diagonals represent a fold change of 1 (i.e., no change) and the nominal fold changes of 5 for the 4x spike (A) and 11 for the 10x spike (B). (The latter is parallel to the diagonal due to the log-log scale.) Without background correction, fold change estimates are biased towards 1, and bias is substantially worse for low-intensity probes.

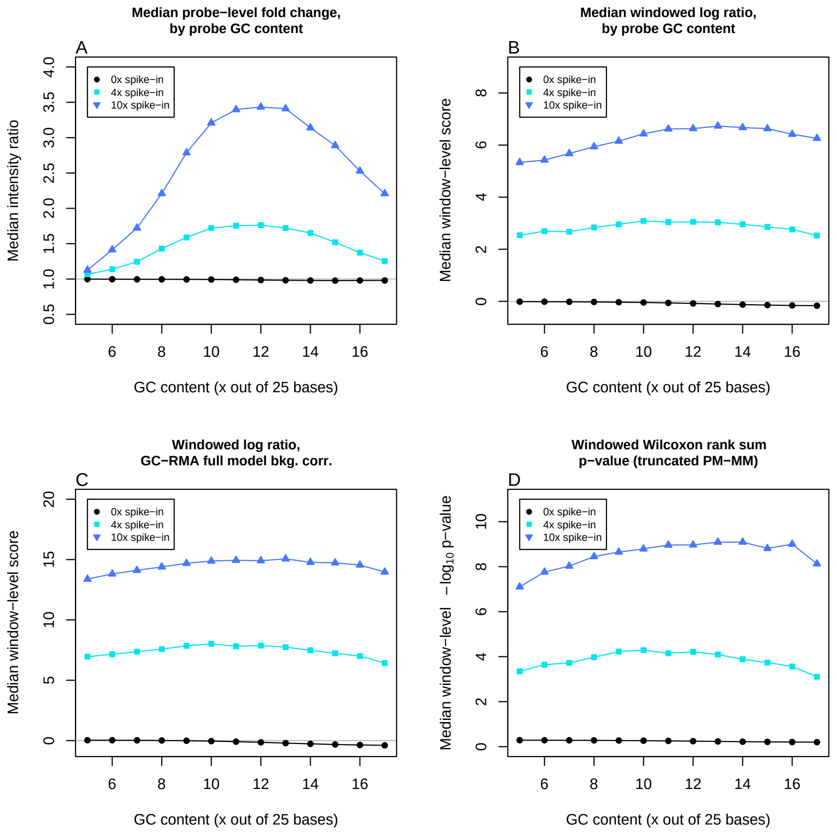

Figure 4.3: For an experiment in which BAC DNA from known genomic regions was added at known concentrations, probes were classified by GC content. For a variety of statistics, the median value per GC bin was computed. (GC bins corresponding to the most extreme 1% are not shown: they contain few probes, have highly variable medians, and distract from the general trend.) (A) Probe-level fold change for PM probe intensities. Arrays were quantile normalized, but no other preprocessing was applied. In 0x (i.e., no spike-in) regions, median fold change was 1.0 as expected, regardless of probe GC content. For the 4x and 10x regions, fold change estimates were substantially lower than the nominal change, with worse bias for high and, especially, low GC content. (B) A moving average (±350 bp) of probe-level log ratios which, for a window with n_i probes, was scaled by $\sqrt{n_i}$. By averaging over a range of contiguous probes, the GC effect was largely—though not completely—eliminated. (C) Preprocessing with the GC-RMA full model, which smoothly integrates mismatch subtraction and a sequence-based correction, further corrected the GC effect for the 10x region and increased the dynamic range relative to panel B. (D) For the TAS windowed (±350 bp) Wilcoxon rank sum procedure, we plot the median value for $-\log_{10} p$ against probe GC content. Some GC effect was still detectable in spite of mismatch subtraction.

to the microarray is a product of antibody-epitope affinity, which may vary from site to site for the same antibody-epitope pair due to differences in local chromatin context; of amplification efficiency, which may also vary from site to site; and of the site occupancy rate in the original biological sample. Nonetheless, background correction may still be relevant. The identification of regions of protein-DNA association depends on perceived differences between treatment and control, so downward bias in fold change estimates may contribute to reduced detection sensitivity. Further, a much-touted advantage that high-density tiling arrays enjoy over their spotted array predecessors is unbiased coverage of the genome. Figure 4.3A, however, suggests that bias in perceived fold change, and thus in detection sensitivity, may vary with GC content in the absence of background correction. It is therefore reasonable to ask if this effect favors the identification of protein-DNA association in some regions at the expense of others—due to long range inhomogeneity in base usage across the genome—or if it produces bias in the inferred weight matrices produced by sequence motif detection algorithms, which are commonly applied to the regions of apparent transcription factor binding identified by ChIP-chip. We address the first of the these questions, at least, below.

4.2.2 Variability in probe response

Equation (4.1) suggests that observed intensity will vary from probe to probe even when target abundance is constant—due to a combination of varying probe responsiveness, measurement error, and fluctuations in background. Figure 4.2 confirms that the magnitude of this variation is large. In the 10x spike-in regions, the signal-to-noise ratio is only 1.30. For the 4x spike-in regions with weaker enrichment, noise exceeds signal in magnitude: the signal-to-noise ratio is only .75.

In Chapter 2 we saw that substantial improvements in ChIP-chip signal-to-noise ratio can be achieved by at least partially canceling the α_i terms via a ratio, and in the multi-sample gene expression context, methods which correct for the probe response effect consistently outperform those which do not (Li and Wong, 2001; Irizarry et al., 2003; Affymetrix, Inc., 2005). In fact, all ChIP-chip analysis methods considered here (see Table 4.1) do address varying probe response, except for Affymetrix's Tiling Analysis Software (TAS); for both its Wilcoxon rank sum p-values and its Hodges-Lehmann signal estimates, TAS treats all probes falling within the moving window as equivalent (Affymetrix, Inc., 2006). Ratio-based methods (e.g., the W_i of Chapter 2, ChIPOTle, TileMap, and Keleş '06) address probe response indirectly through approximate cancelation: under the additive background, multiplicative error model, the intensity ratio is *almost* a noisy version of the abundance fold change. Letting a prime denote treatment, and undecorated symbols denote control,

$$\begin{aligned}\frac{I'}{I} &= \frac{\alpha A'\epsilon' + B'}{\alpha A\epsilon + B} \\ &= \frac{A'}{A} \cdot \frac{\epsilon'}{\epsilon} \cdot \frac{1 + \frac{B'}{\alpha A'\epsilon'}}{1 + \frac{B}{\alpha A\epsilon}} \\ &\approx \frac{A'}{A} \cdot \frac{\epsilon'}{\epsilon},\end{aligned} \tag{4.2}$$

As can be seen from (4.2), this approximation is best when the additive background components (B and B') are small relative to the real signal components (αA and $\alpha A'$). One might therefore expect that preprocessing with a background correction scheme such as VSN (Huber et al., 2002) or GC-RMA (Wu et al., 2004) will improve the cancelation, and we explore this option below.

Other methods attempt to estimate probe response directly or to incorporate it into a probabilistic model. The Li et al. (2005) HMM, for example, estimates probe response parameters from external data. MAT uses probe sequence and a linear model to estimate and then remove the probe response terms (Johnson et al., 2006). HGMM models treatment and control intensities with gamma distributions whose means are correlated between treatment and control for a given probe, but which vary independently from one probe to the next (Keleş, 2005).

4.2.3 Variance estimation

In addition to varying in their responsiveness to target, short oligonucleotide probes may also vary in their accuracy. This has been well-studied in the context of expression arrays; indeed, correction of heteroscedasticity is the main motivation for preprocessing methods such as VSN (Huber et al., 2002). It is reasonable to expect that ChIP-chip data may also sometimes exhibit this behavior, and to suppose that estimation of probe-level variances may improve performance. Heteroscedasticity need not, however, always be the case. In Figure 4.4, for example, we use four arrays from the artificial BAC experiment discussed below, and plot deviation from median versus median log intensity for positions of the genome where there is no enrichment, i.e., no difference between treatment and control. These data are, somewhat surprisingly, homoscedastic.

To show how probe-level variance estimation is typically incorporated, we first present a statistical framework applicable to the models in the first section of Table

Method	*Background*	*Probe response*	*Comments*	*Reference*
W_i	None.	Log ratio.	Moving average of log ratios. Window-level variance adjustment which incorporates spatial correlation. Parametric or non-parametric p-values, adjusted for multiple testing to control FDR.	Chapter 2 above, Schwartz et al. (2006)
ChIPOTle	None.	Log ratio.	Moving average of log ratios. Left-hand tail used for parametric p-values via normal CDF. Probe-level statistics assumed independent. Bonferroni adjustment for multiple testing.	Buck et al. (2005)
MAT	None.	Sequence-based probe response estimate, typically combined with log ratio.	Smoothed variance estimate, computed by binning probes with a similar estimated response parameter. Resulting t statistics are combined by moving trimmed mean, with p-value derived from normal approximation.	Johnson et al. (2006)
TileMap	None.	Log ratio.	Empirical Bayes model smoothes variance estimates. Resulting t statistics are combined by a moving average or HMM, with significance assessed by Unbalanced Mixture Subtraction.	Ji and Wong (2005)
Keleş '06	None.	Log ratio.	Standard two-sample t statistics, combined by moving average. Significance by normal approximation and parametric bootstrap, with various multiple testing correction options.	Keleş et al. (2006)
Chipper	Background correction and normalization via variance-stabilizing transformation.	Generalized log ratio.	Intended for spotted arrays, but easily adapted to one-channel high-density arrays. Left-hand tail used to estimate a Gaussian null; p-values adjusted for multiple testing to control FDR.	Gibbons et al. (2005)
GC-RMA/W_i	Sequence-based background correction, optionally combined with mismatch subtraction.	Log ratio.	GC-RMA was developed for expression arrays, but its background correction scheme may be used as preprocessing for, e.g., W_i, TAS, TileMap, or Keleş '06.	Wu et al. (2004)
TAS	Mismatch subtraction.	Probe response assumed equal for all probes.	Wilcoxon rank sum test applied to a moving window. Asymptotic approximation to the distribution of test statistic used for p-values.	Cawley et al. (2004)
HGMM	None.	Mixture model permits probe-specific intensity distributions.	A common coefficient of variation is assumed for all probes on a given array. Estimation by EM algorithm. FDR is controlled via a "direct posterior probability."	Keleş (2005)
Li '05	Mismatch subtraction.	Probe-specific behavior parameters estimated from putative control data external to the experiment.	Two-state HMM applied to probe-level statistics for identification of enriched regions.	Li et al. (2005)

Table 4.1: An overview of recently proposed methods for analyzing data from ChIP-chip with high-density oligonucleotide tiling arrays. Methods in the first set are based on probe-level statistics consistent with Equation 4.3; see Table 4.2 for more details. Methods in the second set take distinct approaches.

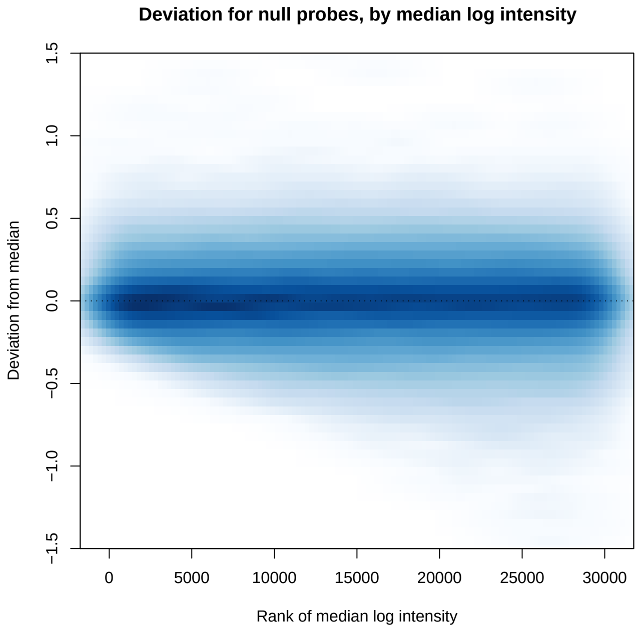

Figure 4.4: Four arrays from the BAC spike-in experiment, in which DNA from known regions was added at known concentrations, were quantile normalized. No other preprocessing steps were applied. For regions known to exhibit no enrichment, a median log intensity was computed, as well as the four deviations from this median. We produce a smoothed scatter plot (darker colors indicate higher frequencies) of deviation vs. the rank of the median for each position falling within such regions, and observe that magnitude of deviation is largely independent of the median value.

4.1. Each is based on probe-level statistics which are, fundamentally, a logged ratio of intensities—either raw intensities or, in some cases, intensities which have been adjusted in an effort to correct for background and/or probe response terms, adjust for heteroscedasticity, etc. Replicates are combined by an arithmetic mean on the log scale, or equivalently, by a geometric mean on the natural scale. In the former form, the probe-level statistics for each method may be expressed as follows:

$$Y_i = \frac{1}{n'} \sum_{j=1}^{n'} \frac{\log \tilde{I}'_{ij}}{s'_{ij}} - \frac{1}{n} \sum_{j=1}^{n} \frac{\log \tilde{I}_{ij}}{s_{ij}}. \tag{4.3}$$

Here, a prime again distinguishes treatment from control, n denotes the number of arrays, $\tilde{I}_{ij}$ denotes a transformed perfect match probe intensity, and s_{ij} denotes a (potentially) probe- and array-specific standard deviation estimate. The methods differ, of course, in their choice of values for these components. The windowed log ratios of

Chapter 2, for example, make no intensity transformation (i.e., $\tilde{I}_{ij} = I_{ij}$) and no probe-specific variance estimates (i.e., $s_{ij} = 1$ for all i and j). (They do go on to estimate a *window-specific* variance, but this is required by irregularity in probe spacing, not by assumed differences in probe-level variability.) TileMap uses unadjusted intensities in the numerators, but then employs an empirical Bayes model to produce probe-specific variance estimates which are moderated by a global estimate—especially when few arrays are used. (A slight variant on this model has previously been used to estimate variances for expression arrays (Lönnstedt and Speed, 2002; Smyth, 2004).) MAT, on the other hand, substantially corrects the raw intensities by dividing out a sequence-based probe response term estimate; it also makes array- and probe-specific variance estimates. Table 4.2 summarizes these differences.

4.3 Data and Methods

4.3.1 Experiments

The methods comparisons which follow are based on data from real and artificial ChIP-chip experiments carried out by the Berkeley *Drosophila* Transcription Network Project.

Zeste and Pol II ChIP-chip

Chromatin was derived from stage 11 (Zeste) or stage 10 (Pol II) wild type embryos after formaldehyde crosslinking. For both experiments, three different sample types were prepared: (i) a treatment sample obtained by ChIP using a specific antibody, (ii) a mock-IP control sample obtained using an IgG antibody with no known affinity for *Drosophila* proteins, and (iii) an input DNA control sample that contained genomic DNA. Samples were then amplified by random-primed PCR, and labeled and hybridized to Affymetrix *D. melanogaster* tiling arrays per standard protocol. The arrays are as described in Chapter 2: over 3 million 5μ perfect-match features (plus an equal number of mismatch features) which interrogate essentially all non-repetitive euchromatic sequence, spaced with a median gap size of 11 bp between interrogated 25-mers.

For the Zeste experiment, two biological replicates were prepared for the treatment and mock IP samples, whereas a single biological replicate was prepared for the input sample. For all three sample types, aliquots were hybridized in triplicate, yielding 6 treatment, 6 mock IP, and 3 input control data sets. For the Pol II experiment,

Method	$\tilde{I}_{ij}$	s_{ij}	*Comments*
W_i	I_{ij}	1	No intensity adjustment or probe-specific variance estimates.
MAT	$(I_{ij}/\hat{\alpha}_{ij})$	$s_{\text{bin}(i),j}$	Arrays are processed individually and probe sequence is used to estimate the probe response term α. Binned variance estimates—for sets of probes with similar $\hat{\alpha}$ estimates—are also computed one array at a time.
TileMap	I_{ij}	$\tilde{\sigma}_i\sqrt{1/n' + 1/n}$	No intensity adjustment. Empirical Bayes probe-specific variance estimates are computed by fitting all arrays simultaneously.
Keleş '06	I_{ij}	$\sqrt{(s'_i)^2/n' + (s_i)^2/n}$	A standard two-sample t statistic is compute for each probe.
Chipper	$(I_{ij} - a_i) + \sqrt{(I_{ij} - a_i) + b_i}$	1	With $\tilde{I}_{ij}$ defined as shown, $\log \tilde{I}_{ij}$ is a generalized logarithm of I_{ij}. For single-channel high-density arrays, the probe-specific a_i and b_i may be estimated from all arrays simultaneously.
GC-RMA/W_i	$I_{ij} - \hat{B}_{ij}$	1	Probe sequence and, optionally, mismatch partner intensities are used to correct for additive background. Background correction is applied to each array separately.

Table 4.2: A summary of perfect match intensity transformations and probe-specific variance estimates used by various analysis methods, prior to evaluation of Equation 4.3. Notation has been changed from original papers for consistency. Here and throughout, a prime is used to distinguish treatment from control when necessary. Chipper was originally proposed for spotted arrays, but adapts naturally to high-density tiling arrays. GC-RMA applies to expression arrays, but its background correction may be used as a preprocessing step for the windowed log ratios, or for other methods.

two biological replicates were prepared for each condition, but technical replicates were omitted.

BAC spike-in artificial data

To assist in evaluating statistical analysis methods, input DNA representing the whole genome was spiked with *Drosophila* DNA fragments maintained in bacterial artificial chromosomes (BACs), yielding samples in which known regions of chromosomes 2 and 3 (ranging from 148 to 193 kb in length) were enriched at known concentrations. In total, seven different spike-in concentrations were considered, although we only show data from the 1x, 4x, and 10x regions here. Before labeling and hybridization, DNA for both BAC spike-in and input control samples was amplified by random-primed PCR.

The generative model presented in Chapter 2 suggests that in experiments targeting, say, a transcription factor protein or unphosphorylated Pol II, ChIP-enrichment should be highly localized, and should tail off as one moves away from sites of protein-DNA interaction. Clearly, the BAC spike-in samples' uniform enrichment over relatively large genomic regions does not replicate this pattern; nonetheless, the BAC data permit exact specification of enriched and unenriched regions of the genome, providing a useful complement to actual ChIP-chip data.

4.3.2 Analysis methods

Prior to applying any analysis method, probes whose 25 base sequence mapped (MUMmer, Kurtz et al., 2004) to more than one location in the *Drosophila* genome (release 4.3) were discarded. Then, all arrays—IP and input control—were quantile normalized as a set (Bolstad et al., 2003). While one might expect this to dampen enrichment signal somewhat, we have found negligible negative impact in most cases, and a substantial positive impact in the odd case where differing hybridization or scanning conditions have changed the shape of the intensity distributions for a subset of the arrays (data not shown).

The log ratios of Chapter 2 may be represented in the form of (4.3), with $\tilde{I}_{ij} = I_{ij}$ and $s_{ij} = 1$ for all i and j. The resulting probe-level Y_i were then smoothed with a ±350 bp moving average to obtain window-level W_i. We then computed probe-level statistics for the other methods listed in Table 4.1. MAT, TileMap, and the Keleş '06 method also have probe-level statistics which are consistent with (4.3), differing only in their choice of $\tilde{I}_{ij}$ and s_{ij}. (See Table 4.2.) In Chapter 2, we proposed a window-level scaling procedure which produced $\tilde{W}_i$ with approximately equal variance for all non-enriched regions. The moving-average version of the TileMap algorithm and the Keleş '06 method take a much simpler approach to averaging: they fix the number of probes in a window rather than the genomic size, and ignore spatial correlation among the probe-level statistics. MAT, on the other hand, focuses on window size rather than probe count, but also ignores correlation among probe-level statistics: Johnson et al. (2006) suggest scaling each window-level trimmed mean by $\sqrt{n_i}$, where n_i denotes the number of values averaged after trimming. This multiplicity of averaging methods leaves us with two options:

1. Allow TileMap and the Keleş '06 method to average based on probe index rather than genomic position. Neither method applies scaling, since none is necessary under the (incorrect) assumption of independence of probe-level statistics. Allow MAT to apply trimmed means, scaling the results as described above. Similarly, compute the $\tilde{W}_i$ as described above.

2. Apply a common windowing procedure the each method's probe-level statistics.

While both options produce interesting comparisons, we chose to focus on the performance of probe-level statistics; accordingly, we used a common windowing strategy to avoid confounding. Specifically, for each of these "generalized t statistic" methods, we smoothed the probe-level statistics with the same ±350 bp moving average and scaled the results by $\sqrt{n_i}$. (Omitting scaling entirely would give windows with fewer probes undue influence; $\sqrt{n_i}$ scaling balances windows with respect to probe count when probe-level statistics are independent, but favors windows with *more* probes when probe-level statistics are positively correlated.) Further, we smoothed MAT's t values with a simple mean rather than the trimmed mean suggested in the original algorithm. In fact, MAT applies its trimmed means to individual arrays first, and then averages the resulting window-level "MAT scores" across replicates; as a consequence, different sets of probes may end up being trimmed on each array, and (4.3) is not strictly applicable. Switching to a simple mean makes this small difference in order irrelevant, leads to better comparability for our purposes here, and brings MAT in line with the other generalized t methods.

The windowed log ratio approach can easily be combined with background correction or other preprocessing steps. In addition to the standard W_i described in the preceding paragraph, we considered two different GC-RMA background corrections (Wu et al., 2004), as implemented in the R `gcrma` package. We also considered the Chipper (Gibbons et al., 2005) approach, using the vsn package in R to apply a variance-stabilizing transformation to the raw perfect-match intensities. This transformation serves to both normalize the intensity data and correct for additive background (Huber et al., 2002). Chipper is intended for two-color arrays and it therefore estimates transformation parameters one array (i.e., two channels) at a time; with single-channel Affymetrix arrays, it was more natural to estimate transformation parameters from all arrays simultaneously.

To implement TAS, which does not fall within the generalized t statistic framework, we used the Wilcoxon rank sum p-values produced by Affymetrix's software, again with a bandwidth of ±350 bp. As discussed in Chapter 2, the distributional assumptions for the rank sum test are not typically satisfied by ChIP-chip data, but the reported "p-values" are just a monotone transformation of the rank sum scores, and as such, still permit the ranking needed for the construction of ROC curves.

HGMM requires that the genome be divided into disjoint "regions," and produces a single test statistic—the posterior probability of a "peak" in that region—for each of these. For the Pol II pseudo-ROC analysis, we created 850 bp regions centered on the chromosome 2L pseudo-positive and pseudo-negative intervals described below, and then supplied the algorithm with normalized intensities from these regions. (Analysis was restricted to chromosome 2L due to computational efficiency issues.) We specified a uniform prior on "peak" size, first ranging from 10 to 19 positions, then ranging from 3 to 19 positions. (Results were similar in both cases.) HGMM's default values were used to initialize the EM algorithm.

ChIPOTle's probe-level and window-level statistics are identical to the LR_i and W_i of Chapter 2. This method takes a different approach to assessing statistical significance and correcting for multiple testing, but since these latter issues are not relevant for ROC-based comparisons, we do not present ChIPOTle as a distinct method in what follows. The HMM approach of Li et al. (2005) was also omitted: it seems to have been supplanted by MAT, which was developed by the same group. Moreover, it requires external data for estimation of null probe-level mean and variance parameters, and proposes an enriched-state emission model with probe-level parameters which are rather arbitrary functions of these null model estimates. The authors cite empirical evidence for this model for SNP genotyping arrays, but it seems unlikely to be appropriate for ChIP-chip, particularly given the content of Chapter 2.

4.3.3 ROC and pseudo-ROC comparisons

For the BAC spike-in data, we have a set of regions R in which the treatment sample is known not to be enriched relative to the control sample, and a set of regions S in which the treatment sample is known to be enriched. (1x spike-in regions were used to define S.) If we denote the test statistic which an analysis method associates with a window falling in one of these regions as U_k for $k \in R$ and V_k for $k \in S$, then the empirical estimate of the method's receiver operating characteristic curve is obtained by plotting

$$\widehat{\text{ROC}} = \left\{ \left(\hat{\mathbb{P}}(U > t), \hat{\mathbb{P}}(V > t) \right) : t \in (-\infty, \infty) \right\}. \tag{4.4}$$

For simplicity, 5000 positions were sampled from each of the BAC R and S sets, and the window-level statistics associated with these positions were used to construct the estimated ROC curves.

For the Zeste and Pol II experiments, however, R and S are not readily available. In the case of Zeste, a small number of binding sites are known, and sequence from these binding sites can be used to construct a model for other binding sites. Unfortunately, short motifs consistent with such a model appear far too frequently in the genome, and most do not represent sites of actual Zeste binding. Further, it is possible that many *real* binding sites recruit Zeste by alternative mechanisms, and as a consequence are inconsistent with the model. Thus, any sequence-based attempt to define S is likely to be heavily contaminated with negative regions; and defining R based on a failure to match the motif model is likely to let some true positives slip through.

For Pol II, the situation is similar: even if gene annotation in *Drosophila* were perfect, many annotated genes are not actually expressed in a given sample, and thus will not exhibit Pol II localization at their transcription start sites. As a consequence, defining S based on annotated transcription start sites will produce substantial contamination by regions which are transcribed in some tissues under some conditions, but which are negative in the sample being considered. Defining R to be a set of regions far from any annotated transcription start site is likely to produce a reasonable true negative set, but even here, some contamination is expected—due, for example, to the existence of unannotated genes or of alternative 5' ends for annotated genes.

To address these problems, we appeal to the pseudo-ROC approach presented in Chapter 3. Definitions for pseudo-positive and pseudo-negative regions, for both the Zeste and Pol II experiments, are given next. While the compression of the pseudo-ROC curves that results from contamination of the test set data—especially of the pseudo-positive sets—makes it more difficult to detect small quality differences between methods, we will see below that such differences can still be seen, and that pseudo-ROC analyses using real Zeste or Pol II ChIP-chip data essentially corroborate the full ROC analyses using the artificial BAC data.

Zeste intervals

To construct pseudo-ROC plots for the Zeste ChIP-chip data, we began with extended promoter regions spanning -2000 to +500 bp relative to the annotated transcription start sites for FlyBase genes. These regions were deemed to be pseudo-positive candidates if the average score (PATSER, Hertz and Stormo, 1999) for the 5 top-scoring 10-mers within the region—relative to a position weight matrix for the Zeste binding motif constructed from 41 known binding site— fell within the top quartile for such scores.

Candidate pseudo-positive regions were then filtered to remove overlaps, yielding a total of 535 disjoint regions. To select 1,054 pseudo-negative regions, we required a top-five average motif score in the bottom quartile, and again filtered to remove overlapping regions. Because both pseudo-positive and pseudo-negative regions corresponded to promoters for annotated genes, they did not differ significantly in terms of average probe GC content or probe density. This suggests, though it does not guarantee, that conditions 1 and 2 of Section 3.3.2—which are sufficient to ensure that pseudo-ROC and proper ROC analyses lead to the same ranking of procedures—may hold. The fact that the pseudo-ROC analyses presented below come to the same conclusions as proper ROC analyses—based, however, on different data—lends further support in this respect.

The pseudo-positive and pseudo-negative regions so defined always included multiple probes, and thus multiple window-level scores. We could have, as with the BAC data, constructed pseudo-ROC curves from a set of non-overlapping windows found within the pseudo-positive and pseudo-negative regions. Because the Zeste regions are more biologically meaningful than in the BAC case, we chose instead to assign a single score to each region: the score from the single highest (or lowest, for TAS p-values) scoring window that it contained. In this way, each 2500 bp promoter region was associated with a positive call if and only if it contained at least one positive binding site call.

Pol II intervals

For the Pol II ChIP-chip data, short proximal promoter regions from -150 to -1 bp relative to the annotated transcription start sites for FlyBase genes were selected as pseudo-positive candidates. Pseudo-negative candidates were then created by matching pseudo-positives to intergenic regions of the same size, with matching probe density and similar GC content. (This matched pairs approach again attempts to ensure that the statistical behavior of misclassified intervals is similar to that of their correctly classified counterparts.) The pseudo-positive and pseudo-negative sets were filtered to remove overlapping regions, yielding 14,149 pseudo-positives and a similar number of pseudo-negatives. As with the Zeste regions, the best window-level score associated with any position in a region was then attributed to that region.

4.4 Results

4.4.1 Background correction and GC bias

Figure 4.3A demonstrates that perceived fold change at the probe level depended strongly on probe GC content, which may be taken as a rough proxy for the probe response term α. Figure 4.3B shows, however, that this effect was largely gone at the window level. As discussed in Chapter 2 and, e.g., Ji and Wong (2005) and Keleş et al. (2006), averaging across contiguous positions was originally motivated by the expected size of regions of enrichment and by the desire to reduce the impact of isolated aberrant probes. We see here, though, that it also served to balance probes with varying response characteristics against one another. Some mild downward bias is still visible, nonetheless, for window-level statistics associated with probes with the lowest GC content.

Equation (4.2) suggests that background correction may help reduce any residual downward bias in fold change by reducing the magnitude of additive background relative to real signal. Figures 4.3C and 4.3D show the median window-level signal for two methods implementing mismatch-based background correction, which is the most aggressive: GC-RMA full model preprocessing combined with the W_i windowed log ratios, and Affymetrix's TAS windowed Wilcoxon rank sum test. The former, which smoothly combines mismatch subtraction with a sequence-based correction, marginally reduced GC bias for probes in the 10x spike-in regions; the latter was more variable across the range, and with respect to GC sensitivity, was equivalent to or slightly worse than the windowed log ratio results. Other background correction approaches considered—GC-RMA affinities-only preprocessing and Chipper's variance stabilizing transformation—yielded results very similar to those depicted for the windowed log ratios in Figure 4.3B.

Although GC-RMA full model background correction achieved a minor improvement with respect to GC bias, this came at a cost. Figure 4.5A gives ROC curves contrasting the performance of five different background correction approaches in the BAC experiment. Both TAS and the GC-RMA full model performed significantly worse than approaches which ignored all (or most, in the case of the GC-RMA affinities method) MM probes. The milder forms of background correction, on the other hand, achieve sensitivities that were as good as, but not better than, those of the uncorrected windowed log ratios. In Figure 4.5D we select a single false positive rate (1%), zoom in on this slice of the ROC curves in Figure 4.5A, and examine sensitivity as a function of the GC content of each window's central probe. At this false positive rate, all methods—even those implementing the most aggressive background correction—still exhibited a slight loss of sensitivity at low GC content positions. (Results for other false positive rates were qualitatively the same.)

Figure 4.6A shows similar results in a more realistic context, albeit one in which a large set of perfectly classified test regions is not readily available. (As a consequence, the plot depicts empirical estimates of pseudo-ROC curves—which have been compressed in both the horizontal and vertical directions, relative to the true ROC curves, as a function of the unknown test set contamination fractions.) For the Pol II experiment, the milder background correction provided by VSN led to a slight improvement over the windowed log ratios, which ignored background correction. GC-RMA's sequence-based model performed essentially the same as the windowed log ratios, but the mismatch-based GC-RMA full model and TAS fared worse, just as they did with the BAC spike-in data set.

Overall, the methods in Figure 4.6A appear to be more closely grouped than was the case for the BAC data, even if we allow for the misclassification compression. (By taking the smallest parallelogram which contains the estimated pseudo-ROC curves, one can, up to estimation error, roughly bound the pseudo-positive misclassification fraction below 40 or 50%, and the pseudo-negative misclassification fraction below 5 or 10%. The differences between the curves in Figure 4.6A have been more greatly reduced, relative to their counterparts in Figure 4.5A, than these percentages would suggest.) The choice of true positives in the BAC experiment, however, provides an

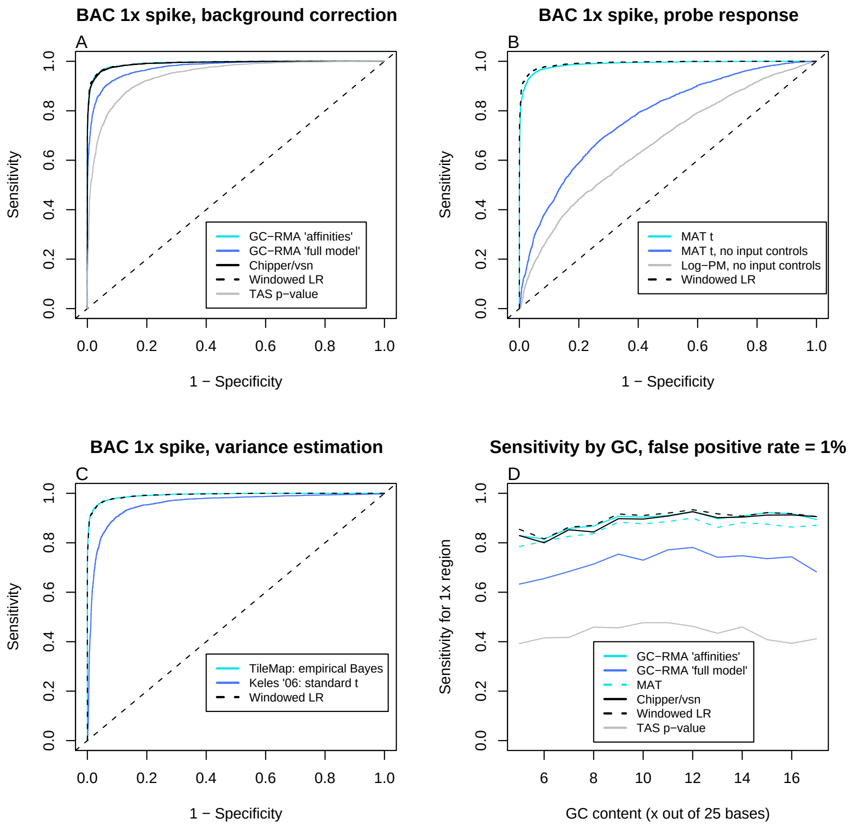

Figure 4.5: ROC curves using BAC spike-in 0x and 1x regions. Probe-level signal was smoothed with a moving window (±350 bp) in all cases, and scaled by $\sqrt{n_i}$ for generalized t methods. (A) Use of MM probes for background correction—by TAS or GC-RMA full model preprocessing—did significantly worse. Neither the milder GC-RMA "affinities" background correction, nor the Chipper/VSN approach produced improvement. (B) MAT uses probe sequence to correct for varying feature response. Combining this with a log-scale difference leads to results which were marginally worse. If input control arrays were ignored, the MAT sequence-based correction was better than uncorrected log-PM intensities; but both were substantially worse than ratio-based results. (C) TileMap's empirical Bayes probe-level variance estimation yielded results which were indistinguishable from the windowed log ratio procedure. The two-sample t statistic did substantially worse. (D) Disaggregation by GC content for a single false positive rate (1%) from the ROC curves in panel A. Mild background correction (GC-RMA affinities or VSN) had a negligible effect; mismatch-based background correction was detrimental. Although MAT's sequence-based model addresses probe response rather than additive background, we have included it here for comparison. For all methods except TAS, sensitivity was slightly lower for GC-poor probes.

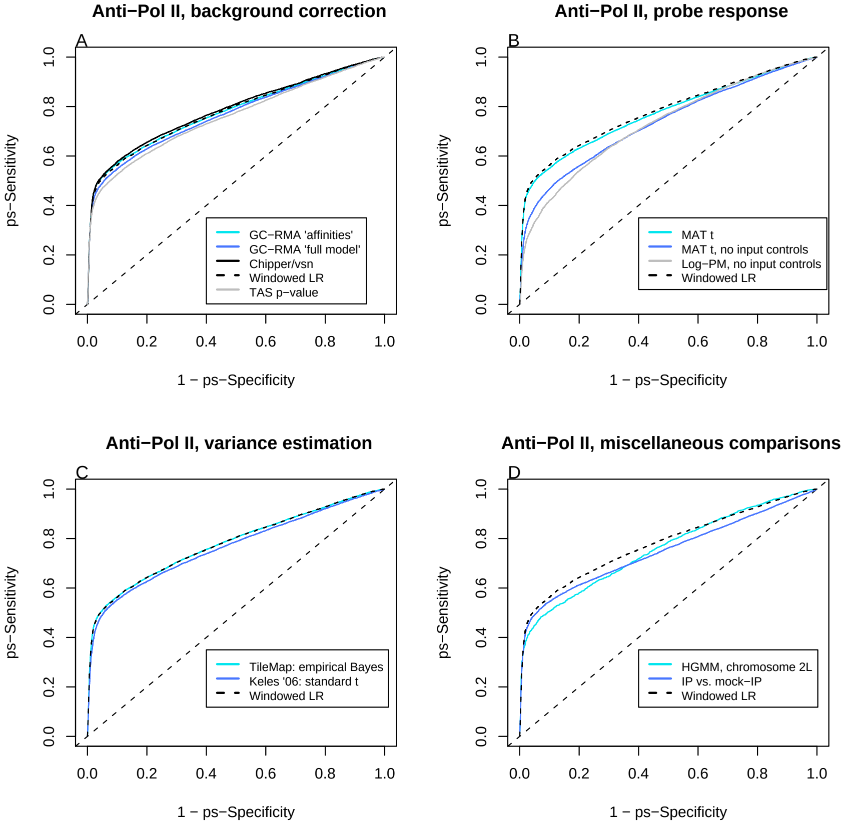

Figure 4.6: Pseudo-ROC curves based on a comparison of 150 bp promoter intervals vs. 150 bp intergenic intervals matched for GC content and probe density. Probe-level signal was smoothed with a moving window (±350 bp) in all cases, and scaled by $\sqrt{n_i}$ for generalized t methods. (A) TAS's mismatch subtraction and the mismatch-based GC-RMA full model did worse; the sequence-based GC-RMA affinities model preprocessing had little impact; Chipper/VSN background correction yielded a slight improvement. (B) MAT's sequence-based adjustment for varying feature response, followed by a log-scale difference, was marginally worse than using a log ratio alone. When input control arrays were ignored, MAT's sequence-based correction improved on uncorrected log-intensities, but both were worse than results obtained using ratios. (C) TileMap's empirical Bayes probe-level variance estimation gave results indistinguishable from the simple windowed log ratio. The traditional, two-sample t statistic was somewhat worse. (D) Using mock-IP controls instead of input controls in the log ratio denominator was detrimental. The HGMM model, with a flat prior on “peak” sizes ranging from 3 to 19 positions, was also substantially worse.

explanation for this effect: the true fold changes in the Pol II experiment are likely to have been greater than the 2:1 ratio for the BAC data positive regions, and differences between the methods may be less relevant when signal is stronger and more easily detectable.

4.4.2 Model-based probe response estimation

While most approaches considered here rely on a ratio to partially correct for varying probe response, MAT takes a model-based approach and attempts to estimate the α_i using probe sequence. HGMM also explicitly incorporates probe behavior within a probabilistic model.

To assess the effectiveness of these methods, we contrasted windowed log ratios with MAT scores (computed, as described above, with a standard rather than trimmed mean) in two different contexts. First, we considered a difference between treatment and input control MAT t values, which is consistent with Equation 4.3 and is the authors' recommended procedure when control arrays are available. In Figure 4.5B, we see that the MAT transformations, when combined with a log-scale average difference, produced a negligible decrease in performance for the BAC spike-in data. We also considered using treatment MAT scores without any input control data: if input controls largely serve to estimate and correct for varying probe response, and the sequence-based model can accomplish this correction by other means, the input controls may not be required. When input controls were ignored, Figure 4.5B shows that the MAT transformation significantly improved on raw log-scale PM intensities: the sequence-based model *was* capturing some aspects of variability in probe response. However, both treatment-only approaches yielded results which were substantially worse than those from treatment-to-control ratios. Figure 4.6B shows similar results in the Pol II experiment, although here the difference between the two treatment-only approaches appears to be smaller.

In Figure 4.6D, we contrast the performance of windowed log ratios with that of HGMM, using the subset of Pol II pseudo-positive and pseudo-negative intervals located on chromosome 2L. The performance of HGMM was significantly worse on these data, but given the broad differences between moving window methods and HGMM, this drop cannot be wholly attributed to HGMM's approach to variability in probe response.

Factor/target	*Input*	*Mock-IP*
Pol II (round 1)	.89	.85
Pol II (round 2)	.86	.84
Knirps	.87	.86
Paired	.89	.78
Hunchback	.85	.83
Bicoid	.86	.85

Table 4.3: Inter-replicate correlation coefficients were computed for six different ChIP-chip experiments, each using two input control arrays and two mock-IP control arrays. Mock-IP data consistently exhibited slightly lower correlation coefficients. Were mock-IP experiments detecting systematic enrichment or depletion—which would presumably also be present in the real IP data—we would expect the opposite to hold.

4.4.3 Control type and probe response estimation

If, as is suggested above, a major role for control data is the estimation of, and correction for, the varying probe response coefficients in (4.1), then the quality of these data will have an impact on the sensitivity and specificity of any method. Typically, mock-immunoprecipitations—performed without an antibody, with a different antibody believed to have no specific targets in the organism under consideration, or with chromatin derived from cells in which the protein of interest lacks the necessary epitope tag, or is missing entirely (Horak et al., 2002; Mao et al., 2003; Cawley et al., 2004; Euskirchen et al., 2004; Odom et al., 2004)—produce little precipitate. As a consequence, more amplification is required to obtain DNA is quantities sufficient for hybridization, and one expects noisier results. Table 4.3, for example, gives the inter-replicate correlation coefficients for both input and mock-IP controls, for the Pol II data discussed in this chapter as well as for several other ChIP-chip experiments performed by BDTNP. While the differences are not large, inter-replicate correlation was always lower between mock-IP controls than between input controls; further, since experiments are carried out by transcription factor rather than by control type, this is not a batch effect. In Figure 4.6D, the windowed log ratio approach of Chapter 2 was applied using both input control and mock-IP controls in the denominators, and there was clear performance degradation when the latter was used.

In fact, control data—in particular, mock-IP controls—play a more complicated role in ChIP-chip analyses. It has been suggested that mock-IP controls, for example, provide a basis for detection of and correction for regions which are systematically enriched as a consequence of off-target antibody binding, or which are systematically depleted by the IP or amplification processes. The data presented in this chapter do not, however, seem to exhibit such phenomena: mock-IP vs. input comparisons typically exhibit no statistically significant enrichment or depletion (data not shown). These phenomena may, however, be present in other data sets, or when other experimental protocols are followed.

4.4.4 Variance estimation

In Figure 4.5C, we contrast three different approaches to variance estimation in Equation (4.3). While the windowed log ratio procedure described in Chapter 2 adjusts variance for each window, this adjustment is due to differences in probe density and spacing within windows, not to any assumed difference in probe-level variance. In other words, the windowed log ratio approach assumes that there is a single, global probe-level variance term. This term can be estimated parametrically, if the normal CDF is used to assign p-values; alternatively, if the non-parametric symmetric-null method is used, it is never explicitly estimated, but rather is incorporated implicitly into $\hat{F}_0$. The Keleş '06 and TileMap methods, on the other hand, do assume heteroscedasticity at the probe level. The Keleş '06 approach computes a distinct variance estimate for each probe, based on the replicated observations in the two conditions. TileMap smoothes evenly between the global and the local, in a data adaptive fashion: the B parameter in its empirical Bayes model is estimated from the data, and determines the extent to which probe-level variance estimates derive from a single global estimate rather than from the probe-specific estimates of the Keleş '06 t statistic.

Figure 4.5C shows that with the BAC data, the windowed log ratios and the TileMap results were comparable, while the traditional two-sample t statistic fared worse. Because only two treatment and two input control arrays were used in the BAC spike-in experiment, traditional probe-specific variance estimates were very noisy; it is therefore not surprising that the Keleş '06 t statistic encountered difficulties. (Indeed, one would be unlikely to actually use such a statistic when $n' = n = 2$.) The similarity between the other two approaches is also not surprising: TileMap estimated $\hat{B} = .79$,

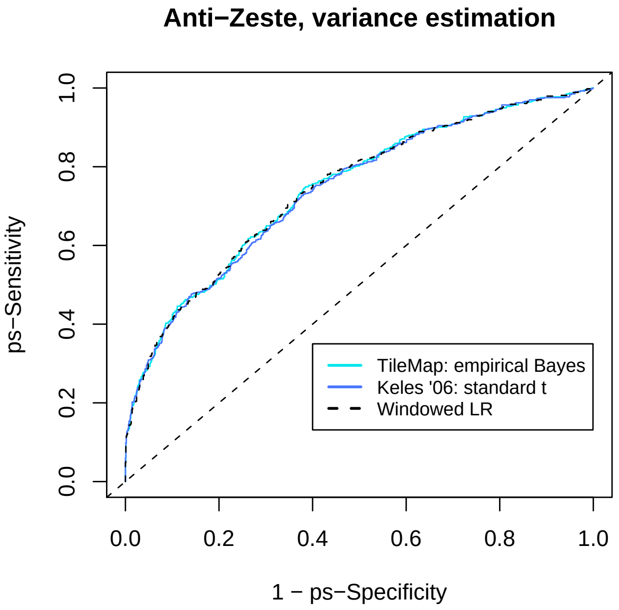

Figure 4.7: The BAC spike-in and anti-Pol2 data sets of Figures 4.5 and 4.6 did not provide enough arrays for effective probe-level variance estimation. The anti-Zeste data set contained 6 treatment (2 biological replicates × 3 technical replicates) and 3 input control (technical replicates) arrays. With more data, all three probe-level variance estimation approaches—the windowed log ratios, TileMap, and the Keles '06 two-sample t—yielded essentially equivalent results.

and so placed most weight on a global variance estimate.

In Figure 4.6C we see similar results for the Pol II data, although the performance of the Keleş '06 t statistic appears to have edged closer to that of the other two methods—again, more so than can be explained by misclassification compression alone. This is especially surprising given that the Pol II experiment also included only two treatment and two input control arrays. To give probe-specific variance estimation methods a better chance, we also constructed pseudo-ROC curves for a Zeste data set in which a total of 9 arrays were used. (With these data, TileMap estimated $\hat{B} = .40$.) For Zeste, we see in Figure 4.7 that there was no significant difference between the pseudo-ROC curves for the three methods.

4.5 Discussion

The analysis of ChIP-chip data, using either traditional two-color arrays or high-density tiling arrays, may be broken into two parts: the conversion of raw intensity measurements into processed test statistics, and the selection of a threshold for these statistics. While both parts contribute to the sensitivity and specificity of the end results, in this chapter we have chosen to focus on the first. Specifically, we have considered the manners in which published methods for high-density tiling ChIP-chip address additive background and differences in probe responsiveness and variability. We have also considered approaches which were originally proposed for expression arrays or spotted array ChIP-chip, but which relate to these same issues and carry over easily.

In searching for protein-DNA interaction, one is most interested in good detection sensitivity and a low false positive rate; the numerical accuracy of fold-change quantification is, at best, a secondary issue. Given this, additive background is relevant only to the extent that it (i) negatively impacts sensitivity or (ii) produces spatial bias in the location of identified positions. With respect to sensitivity, we have shown that mismatch-based background correction does more harm than good. On the other hand, statistical approaches which largely or completely avoid the MM data—such as the GC-RMA affinities model or VSN—might help slightly (Figure 4.6A) or hurt slightly (Figure 4.5D), but typically achieve sensitivities which are very similar to those obtained when background correction is ignored altogether (at least for the data sets under consideration). Furthermore, neither of these PM-focused methods has much impact on the residual GC effect shown for the simple windowed log ratios in Figure 4.3B. The BDTNP has observed that, for the regions exhibiting the strongest evidence of transcription factor binding, base composition is often significantly biased towards guanine and cytosine. One hopes, of course, that there are real biological reasons for this, but the presence of some bias in this direction in the probe-level statistics from the BAC experiment—where any preference for one base over another must be artifactual—was cause for concern. Based on the data shown here, however, our current opinion is that smoothing over windows is sufficient to reduce artifactual effects to a level well below that observed for actual transcription factor binding.

With respect to variability in probe responsiveness, out data suggest that ratio-based methods are effective at canceling (at least partially) the α_i terms in the multiplicative noise, additive background model, but that sequence-based affinity estimation does not provide any additional improvement over what is achieved by a ratio. If control arrays are not available, MAT's sequence-based adjustment *does* clearly reduces probe-to-probe variability relative to unadjusted log-scale perfect match intensities. However, this reduction is no substitute for ratios (or, equivalently, log-scale differences) when control arrays are available. While sequence-based models may one day successfully predict and correct differences in hybridization, synthesis and amplification efficiency—and thereby obviate the need for costly control hybridizations—the current state of the art cannot yet replace empirical estimation using control samples. Further, while mock-IP samples seem to provide a more natural control than input DNA—they more faithfully represent the procedures to which the treatment sample has been subjected—the data presented here suggest that use of an input sample as a direct reference is superior, perhaps because the additional amplification steps required for mock-IP controls make estimation of probe responsiveness terms more difficult. For other experimental protocols, however, the addition level of control afforded by mock-IP data—for other systematic effect not related to probe responsiveness—may justify the introduction of additional noise.

Heteroscedasticity in probe-level statistics is common for expression arrays, and it was somewhat surprising at first to see that probe-specific variance estimates did not improve on a single, global estimate. In fact, the simple two-sample t statistic was typically counterproductive when only a small number of arrays were available, as is often the case. Figure 4.4, of course, provides an explanation, at least for the BAC spike-in data. Averaging over neighboring positions likely contributed as well, reducing the effect of probe-to-probe differences in variability, just as it reduced the impact of varying GC content and probe responsiveness.

At first glance, the windowed log ratio approach—which ignores additive background and differences in probe-level variability—seems naive relative to more involved methods that have recently been proposed for the analysis of high-density tiling array ChIP-chip data. Nonetheless, its performance was very competitive on two complementary data sets—an artificial spike-in experiment in which the true enrichment state was known for all positions, and more natural ChIP-chip experiments for which positive and negative enrichment regions could not be specified with full certainty. We do not doubt, however, that a more thorough understanding of probe behavior will lead to improvement on these results in turn, and hope that the framework presented in this chapter accelerates the process.

Chapter 5

Conclusion

Each of the preceding chapters concluded with a review of its contents, so here we only recapitulate in broad terms. In this document,

- We have presented a generative model for the data obtained from chromatin immunoprecipitation experiments using high density tiling arrays. This model makes several important predictions about the nature of the probe-level statistics which are produced; some of these have been observed informally by other authors (e.g., signal shape near binding sites), while others have not previously been acknowledged (e.g., spatial correlation in probe-level signal). We have shown that the details of the model are consistent with observed behavior in real ChIP-chip data.

- We have show how the generative model guides the selection of a window-level statistic with tractable properties, for which thresholds can be set in a rigorous way. Again, correlation in probe-level signal away from binding sites has not previously been acknowledged or incorporated into enrichment detection procedures, so p-values or posterior probabilities reported by existing methods are not likely to be correct. The p-values and adjusted p-values associated with the $\tilde{W}_i$ of Chapter 2, on the other hand, should behave as advertised.

- We have formally introduced a variant on receiver operating characteristic analysis: pseudo-ROC. Methods comparisons based on the same intuition—that differential behavior on sets which have somehow been enriched for true positives is indicative of real gains is sensitivity and/or specificity—have previously been used, informally, throughout the computational biology literature. The connection between such approaches and the existing research on error rate estimation in the absence of a gold standard, however, has rarely been made. More importantly, the conditions under which this type of comparison is valid have rarely be articulated or checked.

- We have shown how an existing method for assessing the statistical significance of the difference between two correlated empirical ROC curve estimates can be extended to the pseudo-ROC context.
- We have provided a framework which links several recently proposed methods for the analysis of high-density tiling array ChIP-chip data, and then directly compared the probe-level statistics produced by these, and other, methods. Perhaps surprisingly, more elaborate methods, which explicitly address aspects of probe behavior believed to be important for performance, did not actually perform better than simple log ratios. In fact, they performed worse in some cases.

As alluded to at the end of Chapter 2, the methods presented in this document—and indeed the full set of methods presented in the literature to date—still come up short to some extent. In particular, we see two clear avenues for future improvement. The first is related to the fixed bandwidth of moving window methods. Chapter 2 makes it plain that, even in an idealized setting where all binding events are point-like, the size of regions of enriched signal should vary from event to event. For large regions, larger bandwidths have better power. Often, however, the researcher is interested in more subtle events which were not previously detectable; but larger bandwidths will tend to wash these out, by averaging in surrounding unenriched signal. Narrower bandwidths are better suited to smaller regions—and we may be willing to sacrifice some detection power against large regions, since these are often easy to detect in any case—but narrower bandwidths are also more susceptible to noise. Hidden Markov model proponents claim to have resolved this issue, since an HMM only specifies the *expected* duration within a state, not the actual duration in any instance, which remains random. Current HMM implementations, however, only model two states: enriched and unenriched (Ji and Wong, 2005; Li et al., 2005). The same reasoning which shows a single bandwidth to be inappropriate shows a single enriched state distribution to be equally inappropriate. It would be straightforward to carry out a comparison of the degree to which moving windows and two-state HMMs err when presented with signal which varies in both magnitude and duration.

Addressing the issue directly, however, would be better still. The hierarchical gamma mixture model in Keleş (2005) attempts to do this, by explicitly permitting enriched regions of varying size. In Chapter 4, however, this method was seen to underperform; excessive computation time also makes the current implementation practical on large computing clusters only. If, on the other hand, actual binding events are point-like and moderately well separated, Chapter 2 provides a clear basis for a hidden semi-Markov model, in which (i) duration in the enriched state and magnitude of the enrichment may be explicitly modeled, and (ii) the joint distribution for emitted values during the enriched state need not exhibit conditional independence. Peak-like forms of varying sizes can then be explicitly modeled within by a single state. Passing from an HMM to an HSMM would add an extra level of computational complexity, but this should not be prohibitive. (We have prepared details for such an model, although implementation and testing could not be completed in time for inclusion here. It also remains to be seen if extension to the correlated noise context is feasible; without this, the posterior probabilities on the state space will again be incorrect.)

A second, related area in which there is clear room for improvement is in the estimation of error rates. While the methods of Chapter 2 correctly compute these at the window level, error rate estimates are most useful, and most interpretable, at the event level. Correct estimation of event-level error rates is, as already mentioned, complicated by their random boundaries and non-monotonicity with respect to thresholds; it is further complicated by the fact that "events" themselves may not always be unambiguously defined in a biological sense. In Schwartz et al. (2006), for example, ChIP-chip using an antibody against Polycomb identified numerous broad domains over which window-level signal was significantly enriched; multiple sharp peaks, however, could be found within these domains, and these peaks co-localized with more traditional point-like binding sites identified for two other factors. Do such domains each represent a single event, or multiple events? Attempts at event-level error rate estimation so far have utilized putative null data—obtained from mock-IP vs. input comparisons, or, using the same symmetric null hypothesis we make above, from apparent enrichment in favor to the input data—and have simply counted called intervals of enrichment. This approach provides an estimation-focused counterpart to the testing-focused methods of Chapter 2. The behavior of such estimates needs to be explored; if handled carefully, they may yield usable results.

References

Affymetrix, Inc. Guide to probe logarithmic intensity error (PLIER) estimation. Technical note, Affymetrix, Inc., 2005. URL `http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf`.

Affymetrix, Inc. *Affymetrix Tiling Analysis Software Version 1.1 – User Guide*, 2006. URL `http://www.affymetrix.com/support/developer/downloads/TilingArrayTools/TileArray.pdf`.

C. B. Begg. Biases in the assessment of diagnostic tests. *Stat Med*, 6(4):411–423, Jun 1987.

C. B. Begg and R. A. Greenes. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, 39(1):207–215, Mar 1983.

Colin B. Begg. Evaluation of diagnostic tests. In *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd., 2005. On-line edition.

Bradley E Bernstein, Michael Kamal, Kerstin Lindblad-Toh, Stefan Bekiranov, Dione K Bailey, Dana J Huebert, Scott McMahon, Elinor K Karlsson, Edward J Kulbokas, Thomas R Gingeras, Stuart L Schreiber, and Eric S Lander. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2): 169–181, Jan 2005.

Paul Bertone, Mark Gerstein, and Michael Snyder. Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chromosome Res*, 13(3):259–274, 2005.

David R. Bickel. Robust estimators of the mode and skewness of continuous data. *Computational Statistics and Data Analysis*, 39:153–163, 2002.

B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan 2003.

Peter J Brockwell and Richard A Davis. *Introduction to Time Series and Forecasting.* Springer-Verlag, New York, second edition, 2002.

Alexander S Brodsky, Clifford A Meyer, Ian A Swinburne, Giles Hall, Benjamin J Keenan, Xiaole S Liu, Edward A Fox, and Pamela A Silver. Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biol*, 6(8):R64, 2005.

Michael J Buck and Jason D Lieb. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360, Mar 2004.

Michael J Buck, Andrew B Nobel, and Jason D Lieb. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol*, 6(11):R97, 2005.

Gregory Campbell. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine*, 13:499–508, 1994.

Jason S Carroll, X. Shirley Liu, Alexander S Brodsky, Wei Li, Clifford A Meyer, Anna J Szary, Jerome Eeckhoute, Wenlin Shao, Eli V Hestermann, Timothy R Geistlinger, Edward A Fox, Pamela A Silver, and Myles Brown. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, 122(1):33–43, Jul 2005.

Simon Cawley, Stefan Bekiranov, Huck H Ng, Philipp Kapranov, Edward A Sekinger, Dione Kampa, Antonio Piccolboni, Victor Sementchenko, Jill Cheng, Alan J Williams, Raymond Wheeler, Brant Wong, Jorg Drenkow, Mark Yamanaka, Sandeep Patel, Shane Brubaker, Hari Tammana, Gregg Helt, Kevin Struhl, and Thomas R Gingeras. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, 116 (4):499–509, Feb 2004.

H. Cheng and M. Macaluso. Comparison of the accuracy of two tests with a confirmatory procedure limited to positive results. *Epidemiology*, 8(1):104–106, Jan 1997.

Lior David, Wolfgang Huber, Marina Granovskaia, Joern Toedling, Curtis J Palm, Lee Bofkin, Ted Jones, Ronald W Davis, and Lars M Steinmetz. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A*, 103(14):5320–5325, Apr 2006.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1):20–28, 1979. ISSN 00359254.

G. H. de Bock, J. J. Houwing-Duistermaat, M. P. Springer, J. Kievit, and J. C. van Houwelingen. Sensitivity and specificity of diagnostic tests in acute maxillary sinusitis determined by maximum likelihood in the absence of an external standard. *J Clin Epidemiol*, 47(12):1343–1352, Dec 1994.

E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, Sep 1988.

B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18 Suppl 1:S105–S110, 2002.

Bradley Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *JASA*, 99(465):96–104, Mar 2004.

Ghia Euskirchen, Thomas E Royce, Paul Bertone, Rebecca Martone, John L Rinn, F. Kenneth Nelson, Fred Sayward, Nicholas M Luscombe, Perry Miller, Mark Gerstein, Sherman Weissman, and Michael Snyder. CREB binds to multiple loci on human chromosome 22. *Mol Cell Biol*, 24(9):3804–3814, May 2004.

Francis D Gibbons, Markus Proft, Kevin Struhl, and Frederick P Roth. Chipper: discovering transcription-factor targets from chromatin immunoprecipitation microarrays using variance stabilization. *Genome Biology*, 6(11):R96, 2005.

Peter Hall and Xiao-Hua Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *The Annals of Statistics*, 31(1):201–224, 2003.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, Apr 1982.

Christopher T Harbison, D. Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B

Reynolds, Jane Yoo, Ezra G Jennings, Julia Zeitlinger, Dmitry K Pokholok, Manolis Kellis, P. Alex Rolfe, Ken T Takusagawa, Eric S Lander, David K Gifford, Ernest Fraenkel, and Richard A Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, Sep 2004.

S. Blair Hedges and Prachi Shah. Comparison of mode estimation methods and application in molecular clock analysis. *BMC Bioinformatics*, 4:31, Jul 2003.

G. Z. Hertz and G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–577, 1999.

Christine E Horak, Milind C Mahajan, Nicholas M Luscombe, Mark Gerstein, Sherman M Weissman, and Michael Snyder. GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proc Natl Acad Sci U S A*, 99 (5):2924–2929, Mar 2002.

Fushing Hsieh and Bruce W. Turnbull. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, 24(1):25–40, 1996.

Wolfgang Huber, Anja von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, 2002.

S. L. Hui and S. D. Walter. Estimating the error rates of diagnostic tests. *Biometrics*, 36(1):167–171, Mar 1980.

Siu L. Hui and Xiao H. Zhou. Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research*, 7:354–370, 1998.

Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2): 249–264, Apr 2003.

V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder, and P. O. Brown. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, 409 (6819):533–538, Jan 2001.

Hongkai Ji and Wing Hung Wong. TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, 21(18):3629–3636, Sep 2005.

W. Evan Johnson, Wei Li, Clifford A Meyer, Raphael Gottardo, Jason S Carroll, Myles Brown, and X. Shirley Liu. Model-based analysis of tiling-arrays for chip-chip. *Proc Natl Acad Sci U S A*, 103(33):12457–12462, Aug 2006.

Sündüz Keleş. Mixture modeling of genome-wide localization of transcription factors. Technical Report 189, University of Wisconsin, Madison, Biostatistics Program, 2005. URL http://www.stat.wisc.edu/~keles/ggcc.v1.pdf.

Sündüz Keleş, Mark J van der Laan, Sandrine Dudoit, and Simon E Cawley. Multiple testing methods for ChIP-chip high density oligonucleotide array data. *J Comput Biol*, 13(3):579–613, Apr 2006.

Tae Hoon Kim, Leah O Barrera, Ming Zheng, Chunxu Qu, Michael A Singer, Todd A Richmond, Yingnian Wu, Roland D Green, and Bing Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880, Aug 2005.

Antonis Kirmizis and Peggy J Farnham. Genomic approaches that aid in the identification of transcription factor target genes. *Exp Biol Med (Maywood)*, 229(8):705–721, Sep 2004.

Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. Versatile and open software for comparing large genomes. *Genome Biol*, 5(2):R12, 2004.

Tong Ihn Lee, Richard G Jenner, Laurie A Boyer, Matthew G Guenther, Stuart S Levine, Roshan M Kumar, Brett Chevalier, Sarah E Johnstone, Megan F Cole, Kyo ichi Isono, Haruhiko Koseki, Takuya Fuchikami, Kuniya Abe, Heather L Murray, Jacob P Zucker, Bingbing Yuan, George W Bell, Elizabeth Herbolsheimer, Nancy M Hannett, Kaiming Sun, Duncan T Odom, Arie P Otte, Thomas L Volkert, David P Bartel, Douglas A Melton, David K Gifford, Rudolf Jaenisch, and Richard A Young. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell*, 125(2):301–313, Apr 2006.

C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98(1):31–36, Jan 2001.

Wei Li, Clifford A. Meyer, and X. Shirley Liu. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, 21 Suppl. 1:i274–i282, Jun 2005.

Chih Long Liu, Stuart L Schreiber, and Bradley E Bernstein. Development and validation of a T7 based linear amplification for genomic DNA. *BMC Genomics*, 4(1):19, May 2003.

Chris J. Lloyd. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, 93 (444):1356–1364, Dec. 1998.

Chris J. Lloyd. Regression models for convex ROC curves. *Biometrics*, 56:862–867, Sep. 2000.

Ingrid Lönnstedt and Terence P. Speed. Replicated microarray data. *Statistica Sinica*, 12:31–46, 2002.

David M MacAlpine, Heather K Rodrguez, and Stephen P Bell. Coordination of replication and transcription along a *Drosophila* chromosome. *Genes Dev*, 18(24):3094–3105, Dec 2004.

Daniel Y L Mao, John D Watson, Pearlly S Yan, Dalia Barsyte-Lovejoy, Fereshteh Khosravi, W. Wei-Lynn Wong, Peggy J Farnham, Tim H-M Huang, and Linda Z Penn. Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. *Curr Biol*, 13(10):882–886, May 2003.

Rebecca Martone, Ghia Euskirchen, Paul Bertone, Stephen Hartman, Thomas E Royce, Nicholas M Luscombe, John L Rinn, F. Kenneth Nelson, Perry Miller, Mark Gerstein, Sherman Weissman, and Michael Snyder. Distribution of NF-κB-binding sites across human chromosome 22. *Proc Natl Acad Sci U S A*, 100(21):12247–12252, Oct 2003.

Todd C Mockler, Simon Chan, Ambika Sundaresan, Huaming Chen, Steven E Jacobsen, and Joseph R Ecker. Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, 85(1):1–15, Jan 2005.

Zarmik Moqtaderi and Kevin Struhl. Genome-wide occupancy profile of the RNA polymerase III machinery in *Saccharomyces cerevisiae* reveals loci with incomplete transcription complexes. *Mol Cell Biol*, 24(10):4118–4127, May 2004.

Felix Naef and Marcelo O Magnasco. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68(1 Pt 1):011906, Jul 2003.

Felix Naef, Daniel A Lim, Nila Patil, and Marcelo Magnasco. DNA hybridization to mismatched templates: a chip study. *Phys Rev E Stat Nonlin Soft Matter Phys*, 65 (4 Pt 1):040902, Apr 2002.

Duncan T Odom, Nora Zizlsperger, D. Benjamin Gordon, George W Bell, Nicola J Rinaldi, Heather L Murray, Tom L Volkert, Jörg Schreiber, P. Alexander Rolfe, David K Gifford, Ernest Fraenkel, Graeme I Bell, and Richard A Young. Control of pancreas and liver gene expression by HNF transcription factors. *Science*, 303(5662):1378–1381, Feb 2004.

Margaret Sullivan Pepe. Receiver operating characteristic methodology. *Journal of the American Statistical Association*, 95(449):308–311, Mar. 2000.

Margaret Sullivan Pepe and Todd A. Alonzo. Comparing disease screening tests when true disease status is ascertained only for screen positives. *Biostatistics*, 2(3):249–260, Sep 2001.

Dmitry K Pokholok, Christopher T Harbison, Stuart Levine, Megan Cole, Nancy M Hannett, Tong Ihn Lee, George W Bell, Kimberly Walker, P. Alex Rolfe, Elizabeth Herbolsheimer, Julia Zeitlinger, Fran Lewitter, David K Gifford, and Richard A Young. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, 122(4):517–527, Aug 2005.

Yuan Qi, Alex Rolfe, Kenzie D MacIsaac, Georg K Gerber, Dmitry Pokholok, Julia Zeitlinger, Timothy Danford, Robin D Dowell, Ernest Fraenkel, Tommi S Jaakkola, Richard A Young, and David K Gifford. High-resolution computational models of genome binding events. *Nat Biotechnol*, 24(8):963–970, Aug 2006.

B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290 (5500):2306–2309, Dec 2000.

Bing Ren, Hieu Cam, Yasuhiko Takahashi, Thomas Volkert, Jolyon Terragni, Richard A Young, and Brian David Dynlacht. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev*, 16(2):245–256, Jan 2002.

A. Schatzkin, R. J. Connor, P. R. Taylor, and B. Bunnag. Comparing new and old screening tests when a reference procedure cannot be performed on all screenees. Example of automated cytometry for early detection of cervical cancer. *Am J Epidemiol*, 125(4):672–678, Apr 1987.

Yuri B Schwartz, Tatyana G Kahn, and Vincenzo Pirrotta. Characteristic low density and shear sensitivity of cross-linked chromatin containing polycomb complexes. *Mol Cell Biol*, 25(1):432–439, Jan 2005.

Yuri B Schwartz, Tatyana G Kahn, David A Nix, Xiao-Yong Li, Richard Bourgon, Mark Biggin, and Vincenzo Pirrotta. Genome-wide analysis of Polycomb targets in *Drosophila melanogaster*. *Nat Genet*, 38(6):700–705, Jun 2006.

Chad A Shaw. Theoretical consideration of amplification strategies. *Neurochem Res*, 27(10):1123–1131, Oct 2002.

Devanjan Sikder and Thomas Kodadek. Genomic studies of transcription factor-DNA interactions. *Curr Opin Chem Biol*, 9(1):38–45, Feb 2005.

Gordon K Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(1):Article3, 2004.

John D. Storey, Jonathan E. Taylor, and David Siegmund. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society, Series B*, 66:187–205, 2004.

Pavel Tomancak, Amy Beaton, Richard Weiszmann, Elaine Kwan, ShengQiang Shu, Suzanna E Lewis, Stephen Richards, Michael Ashburner, Volker Hartenstein, Susan E Celniker, and Gerald M Rubin. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol*, 3(12):RESEARCH0088, 2002.

E. S. Venkatraman and Colin B. Begg. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*, 83(4): 835–848, Dec 1996.

Amy S Weinmann and Peggy J Farnham. Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. *Methods*, 26(1): 37–47, Jan 2002.

Sam Wieand, Mitchell H. Gail, Barry R. James, and Kang L. James. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 76(3):585–592, Sep 1989. ISSN 0006-3444.

Zhijin Wu, Rafael A. Irizarry, Robert Gentleman, Francisco Martinez-Murillo, and Forrest Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99:909–917, Dec 2004.

Kayoko Yamada, Jun Lim, Joseph M Dale, Huaming Chen, Paul Shinn, Curtis J Palm, Audrey M Southwick, Hank C Wu, Christopher Kim, Michelle Nguyen, Paul Pham, Rosa Cheuk, George Karlin-Newmann, Shirley X Liu, Bao Lam, Hitomi Sakano, Troy Wu, Guixia Yu, Molly Miranda, Hong L Quach, Matthew Tripp, Charlie H Chang, Jeong M Lee, Mitsue Toriumi, Marie M H Chan, Carolyn C Tang, Courtney S Onodera, Justine M Deng, Kenji Akiyama, Yasser Ansari, Takahiro Arakawa, Jenny Banh, Fumika Banno, Leah Bowser, Shelise Brooks, Piero Carninci, Qimin Chao, Nathan Choy, Akiko Enju, Andrew D Goldsmith, Mani Gurjal, Nancy F Hansen, Yoshihide Hayashizaki, Chanda Johnson-Hopson, Vickie W Hsuan, Kei Iida, Meagan Karnes, Shehnaz Khan, Eric Koesema, Junko Ishida, Paul X Jiang, Ted Jones, Jun Kawai, Asako Kamiya, Cristina Meyers, Maiko Nakajima, Mari Narusaka, Motoaki Seki, Tetsuya Sakurai, Masakazu Satou, Racquel Tamse, Maria Vaysberg, Erika K Wallender, Cecilia Wong, Yuki Yamamura, Shiaulou Yuan, Kazuo Shinozaki, Ronald W Davis, Athanasios Theologis, and Joseph R Ecker. Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science*, 302(5646):842–846, Oct 2003.

Ilsoon Yang and Mark P. Becker. Latent variable modeling of diagnostic accuracy. *Biometrics*, 53(3):948–958, Sep. 1997. ISSN 0006341x.

Guo-Cheng Yuan, Yuen-Jong Liu, Michael F Dion, Michael D Slack, Lani F Wu, Steven J Altschuler, and Oliver J Rando. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science*, 309(5734):626–630, Jul 2005.

Li Zhang, Michael F Miles, and Kenneth D Aldape. A model of molecular interactions on short oligonucleotide microarrays. *Nat Biotechnol*, 21(7):818–821, Jul 2003.

Xiao-Hua Zhou, Pete Castelluccio, and Chuan Zhou. Non-parametric estimation of ROC curves in the absence of a gold standard. Paper 231, UW Biostatistics Working Paper Series, University of Washington, 2004. URL `http://www.bepress.com/uwbiostat/`.