University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2004 *Paper* 147

Multiple Testing Methods For ChIP-Chip High Density Oligonucleotide Array Data

Sunduz Keles* Mark J. van der Laan†
Sandrine Dudoit‡ Simon E. Cawley**

*Dept. of Statistics & Biostatistics & Medical Informatics, University of Wisconsin, Madison, keles@stat.wisc.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

‡Division of Biostatistics, School of Public Health, University of California, Berkeley, sandrine@stat.berkeley.edu

**Affymetrix, 3380 Central Expressway, Santa Clara, CA 95051

Multiple Testing Methods For ChIP-Chip High Density Oligonucleotide Array Data

Sunduz Keles, Mark J. van der Laan, Sandrine Dudoit, and Simon E. Cawley

Abstract

Cawley et al. (2004) have recently mapped the locations of binding sites for three transcription factors along human chromosomes 21 and 22 using ChIP-Chip experiments. ChIP-Chip experiments are a new approach to the genome-wide identification of transcription factor binding sites and consist of chromatin (Ch) immunoprecipitation (IP) of transcription factor-bound genomic DNA followed by high density oligonucleotide hybridization (Chip) of the IP-enriched DNA. We investigate the ChIP-Chip data structure and propose methods for inferring the location of transcription factor binding sites from these data. The proposed methods involve testing for each probe whether it is part of a bound sequence or not using a scan statistic that takes into account the spatial structure of the data. Different multiple testing procedures are considered for controlling the family-wise error rate and false discovery rate. A nested-Bonferroni adjustment, that is more powerful than the traditional Bonferroni adjustment when the test statistics are dependent, is discussed. Simulation studies show that taking into account the spatial structure of the data substantially improves the sensitivity of the multiple testing procedures. Application of the proposed methods to ChIP-Chip data for transcription factor p53 identified many potential target binding regions along human chromosomes 21 and 22. Among these identified regions, 18% fall within a 3kb vicinity of the 5'UTR of a known gene or CpG island, 31% fall between the codon start site and the codon end site of a known gene but not inside an exon. More than half of these potential target sequences contain the p53 consensus binding site or very close matches to it. Moreover, these target segments include the 13 experimentally verified p53 binding regions of Cawley et al. (2004), as well as 49 additional regions that show higher hybridization signal than these 13 experimentally verified regions.

1 Introduction

Chromatin (Ch) immunoprecipitation (IP) of transcription factor-bound genomic DNA followed by microarray hybridization (Chip) of IP-enriched DNA is an exciting technology that allows genome-wide analysis of transcription factor binding (see Lee *et al.* (2002) and references therein for the application of this technology with two-color spotted arrays). The data produced by this technology are often referred to as *ChIP-Chip data*. Recently, Cawley *et al.* (2004) performed ChIP-Chip experiments using high density oligonucleotide arrays to identify the locations of binding sites for three transcription factors on human chromosomes 21 and 22. Application of this technology with high density oligonucleotide arrays allows the scanning of whole or parts of a genome at a higher resolution than with spotted microarrays. Here, we investigate this new type of genomic data and propose statistical methods for finding transcription factor-bound sequences. ChIP-Chip data are different than classical microarray gene expression data obtained by measuring mRNA levels or DNA copy numbers (comparative genome hybridization (CGH)) in several respects. For our purpose, there are two probe sequence classes that are of interest. The first one is the class of *bound* sequences, i.e., the group of sequences that are bound *in vivo* by the transcription factor of interest, thus are targets of it. The second one is the *unbound* class, composed of sequences that are not bound by the transcription factor. The goal of analyzing ChIP-Chip data is to first identify the bound sequence fragments and then search for common regulatory elements in these sequences.

Here, we firstly make the observation that there is a *spatial structure* in this type of genomic data due to the fact that IP-enriched DNA fragments bind to multiple adjacent probe sequences. Specifically, the bound probe sequences occur as small *clusters* or *blips*. We propose an appropriate scan test statistic and apply multiple testing procedures to identify these blips of bound sequences. The scan test statistic exploits the spatial structure in the data by combining intensity measures across probes within a certain window size. A cross-validation based method is proposed to select the window size or blip size. We consider multiple testing procedures that control the family-wise error rate (FWER), tail probabilities for the proportion of false positives (TPPFP) among the rejected hypotheses, and the false discovery rate (FDR). Simulation studies suggest that incorporating information about the spatial data structure improves the sensitivity of these procedures. We analyzed the ChIP-Chip data for transcription factor p53 and identified many potential target binding sequences that are located within plausible regulatory regions on human chromosomes 21 and 22. For about 18% of these target regions, enrichment for locations proximal to 5' exons and CpG islands is detected. About 31% of the identified regions are located inside annotated genes but not exons. More than half of these target sequences contain the known p53 consensus binding sequence or very close matches to it. Moreover, these segments include the 13 experimentally verified targets of Cawley *et al.* (2004), as well as 49 additional targets that show higher hybridization signal than these 13 experimentally verified target regions.

This paper is organized as follows: In the next section, we briefly describe ChIP-Chip high density oligonucleotide array experiments and illustrate the spatial structure in the data. We also present the p53 ChIP-Chip dataset of Cawley *et al.* (2004) and discuss preprocessing issues as well as previous work on the analysis of this new type of genomic data. In Section 3, we consider multiple testing procedures that use a scan statistic to capture the spatial structure of the data and discuss the control of various false positive rates. Finally, Section 4 is dedicated to simulation studies and application of our proposed methods to ChIP-Chip data for transcription factor p53.

2 ChIP-Chip high density oligonucleotide array data

2.1 ChIP-Chip high density oligonucleotide array experiments

ChIP-Chip data are produced by the following two steps:

- A DNA binding protein, i.e., transcription factor, is cross-linked to its genomic DNA targets *in vivo* and the chromatin is isolated. Then, the DNA with the bound proteins is sheared by sonication into small fragments of average length $\sim$ 1kb. After an immunoprecipitation step, where the protein-bound DNA is precipitated using an antibody specific to the protein of interest, the DNA is separated from the protein. The resulting solution (IP-enriched DNA) is amplified with polymerase chain reaction (PCR) and the $\sim$ 1kb regions are fragmented into segments of 50-100bps.
- The IP-enriched DNA is fluorescently labelled and hybridized to a chip containing perfect match (PM) and mismatch (MM) probe-pairs of length 25bps. The average center to center distance between each consecutive probe-pair along the genome is about 35bps.

In summary, transcription factor targets appear in fragments of average length 1kb in the IP-enriched solution and after being fragmented into smaller pieces are hybridized to chips containing 25mer probes. As a result of this process, the set of probes that maps to a target fragment of the transcription factor hybridizes to that fragment. This process, as illustrated in Figure 1, causes the data to have a spatial structure, that is, the bound probes are expected to occur in small clusters which we refer to as *blips*.

[Figure 1 about here.]

An empirical plot of the test statistics (introduced in Section 3 and computed from probe-level data) is displayed in Figure 2. This plot illustrates the predicted blip structure. A simple calculation suggests that the bound probes are expected to occur, on the average, in groups of approximately 30 probes. This calculation is carried out as follows. Since the average distance between the mid points of two adjacent probes is 35bps, the average distance between the closest end points of these adjacent probes is 10bps ($35 - 2 \times (25/2)$). Hence, solving for the blip size w, i.e., number of probe-pairs in a blip, in $25w + 10(w - 1) = 1000$ gives us an average blip size of approximately 30 probes. However, empirical plots of the data, such as the plots of simple test statistics in Figure 2, suggest a much smaller blip size ($\sim$ 10 probes). We propose a cross-validation based method to select the blip size for the downstream statistical analysis in Section 3.2.

[Figure 2 about here.]

2.2 ChIP-Chip high density oligonucleotide array data for transcription factor p53

The ChIP-Chip experiments, reported in Cawley *et al.* (2004), aimed to identify binding sites on human chromosomes 21 and 22 for three transcription factors, namely, cMyc, sp1, and p53. Probe-level data were obtained on non-repeat genomic sequences of chromosomes 21 and 22 for $N \approx 1.1$ million PM (perfect match) and MM (mismatch) pairs, distributed on three high density oligonucleotide chips (chips A, B, and C). We focus here only on transcription factor p53, for which the ChIP-Chip experiments were performed using DNA samples from the cell line HCT1116. Three hybridization replicates were performed for each of two IP replicates resulting in a total of six technical hybridizations. The IP replicates were obtained by dividing the DNA extraction sample from cell line HCT1116 into two before the IP-step.

There were also two types of control experiments: *whole cell extraction*, which skips the IP-step, so that hybridization is expected to occur at every probe on the array (positive control), and *controlGST*, which uses a bacterial antibody in the IP-step, so that the resulting DNA solution does not have any transcription factor-bound fragments and hence hybridization is not expected to take place anywhere on the array (negative control). Here, we use *controlGST* as the control group.

2.3 Pre-processing

After hybridization of IP-enriched DNA, the probe-pairs on the chip were mapped to their genomic locations on the June 2002 assembly of the human genome, by simply aligning each 25mer to any exact matching 25mer in this new assembly. As a result of this process, and the fact that repeat masking programs might fail to detect all the repeats, various types of replication of the probe-pairs may occur in the final dataset, i.e., the final dataset may contain cases where the measurement for a single 25mer is repeated for a set of genomic locations. These repeats can be classified as *local* versus *distant* repeats and can occur as explained in detail in Figures 3 and 4.

[Figure 3 about here.]

[Figure 4 about here.]

Table 1 gives an example of a local repeat, where data for two different 25mers are duplicated over a genomic region of 188bps. The data presented in the table correspond to rows 194671 to 194680 of the data file. For instance, if one averages over a window size of ten, from row 194671 to row 194680, to compute the mean intensity over these ten probe-pairs, there are only two independent observations rather than ten. Ignoring this local repeat structure leads to incorrect treatment of dependent data as independent and, as a result, may generate blips which are invalid.

[Table 1 about here.]

Next, we give an example of distant repeat regions in Table 2. There are two different measurements for the same 25mer and each maps to a total of two different genomic locations. However, each measurement is duplicated and paired along with both of these locations. In contrast to the local repeat regions, these two unique locations are far away from each other on the genome. A graphical display of a portion of the repeat regions from chip A is given in Figure 5.

[Table 2 about here.]

[Figure 5 about here.]

We decided to filter out local repeat regions but keep distant repeat regions, because distant repeats carry information about valid targets, with the complication that there are more than one possibility in terms of the location. A simple filtering procedure for local repeats is described in the Appendix. As a result of this filtering process, 1550 (0.47%), 6684 (1.81%), and 7657 (2.32%) probe-pairs were filtered out from chips A, B, and C, respectively.

2.4 Previous work

Cawley *et al.* (2004) also use a multiple hypothesis testing procedure for analyzing ChIP-Chip high density oligonucleotide array data. Specifically, these authors use a Wilcoxon rank sum statistic to test the null hypothesis of equality of the population distribution functions of the hybridization intensities for IP-enriched DNA (treatment group) and control DNA (positive or negative control group). The test is performed for each probe-pair by combining data within a window of 1kb. As a result, a sliding window is applied across the genome, creating as many test statistics as the number of probe-pairs. In order to control the family-wise error rate, a cut-off for the p-values of the Wilcoxon rank sum statistic is obtained based on randomization of the data. The randomizations are performed as follows: The genomic positions associated with each set of 12 (PM, MM) pairs (6 treatment and 6 control) are randomized. The randomized data are used as input to the method for detecting bound probes, i.e., a Wilcoxon rank sum statistic is computed for each randomized genomic position as outlined above. This is repeated 100 times and a p-value cut-off that results in less than one false positive on average is chosen. We note that this randomization procedure aims to generate data under a null distribution with no spatial structure. However, it fails to accommodate the null hypothesis that the treatment and control groups come from the same distribution.

Our approach is different from the above method in several aspects. Firstly, the test statistics in our method operate on a *fixed* number of probe-pairs, whereas Cawley *et al.* (2004) employ a fixed window size of 1kb, which may correspond to *varying* numbers of probe-pairs in different regions of the two chromosomes considered. One potential pitfall of the latter approach is that each Wilcoxon rank sum statistic may depend on a different number of observations, i.e., different numbers of probe-pairs may contribute to each test statistic. This might lead to high variability among the test statistics. Secondly, since the fragment length of 1kb is an average number, most of the fragments might actually be shorter than this average. For example, if the fragment length distribution is an exponential distribution with rate 1/1000bps, many of the fragments will actually be shorter than the average fragment length of 1kb. Hence, it is important to investigate the effect of average fragment length, i.e., allow different window sizes, as we do for the blip size parameter. Finally, our approach tests for differences in the mean hybridization intensities of the IP-enriched and control DNA populations and the null distribution of the test statistics is selected accordingly.

3 Methods

3.1 Model based multiple testing

Let $Y_j = (Y_{j,i} : i \in \{1, \cdots, N\}) \sim P_j, \quad j = 1, 2$, denote a random vector of quantile normalized $\log_2(PM)$ in an IP-enriched DNA (treatment) hybridization ($j = 2$) and a control DNA (controlGST or whole cell extract) hybridization ($j = 1$). The reader is referred to Irizarry *et al.* (2003) for a detailed discussion on pre-processing of high density oligonucleotide array data. Here, P_j are the data generating distributions and we have $n_j = 6$ realizations of each random vector Y_j, $j = 1, 2$. Let $\mu_j = (\mu_{j,i}, i \in \{1, \cdots, N\})$, $j = 1, 2$, denote the corresponding mean vectors in the control and treatment populations. Let $\mu_i = \mu_{2,i} - \mu_{1,i}$ represent the difference in mean $\log_2(PM)$ for treatment and control hybridizations of probe i, $i \in \{1, \cdots, N\}$.

For the negative control (controlGST), we are interested in testing, for each probe i, the null hypothesis $H_{0,i} = I(\mu_i \leq 0)$ versus the alternative hypothesis $H_{1,i} = I(\mu_i > 0)$. Here, we consider one-sided alternative hypotheses, since we expect higher intensity levels in the IP-enriched hybridizations than in

the negative control hybridizations. Let $\bar{Y}_{j,i}$ and $\hat{\sigma}^2_{j,i}$, $i \in \{1, \cdots, N\}$, $j \in \{1, 2\}$, denote the empirical mean and variance of $\log_2(PM)$ for probe i in hybridizations from population j across n_j observations. A test statistic of interest for each probe i is the standard two-sample Welch t-statistic given by

$$T_{i,n} = \frac{\bar{Y}_{2,i} - \bar{Y}_{1,i}}{\sqrt{\hat{\sigma}^2_{1,i}/n_1 + \hat{\sigma}^2_{2,i}/n_2}}. \tag{1}$$

To take into account the blip structure, consider the following *scan statistics*:

$$T^*_{i,n} = \frac{1}{w} \sum_{h=i}^{i+w-1} T_{h,n}, \quad i \in \{1, \cdots, N - w + 1\},$$

where $T_{h,n}$ is the two-sample Welch t-statistic for probe h given in (1). This test statistic aims to borrow strength across w neighboring probes when testing the null hypothesis for a given probe. That is, if the probe is in the neighborhood of bound probes, i.e., within a blip, the scan statistic makes it easier to reject the null hypothesis. Conversely, if the probe is in the vicinity of unbound probes, it becomes harder to reject the null hypothesis.

Multiple testing procedures consist of choosing a vector of cut-offs for the test statistics such that a suitably defined false positive rate is controlled at an a priori specified level α. Since the test statistics typically have an unknown joint distribution, the cut-offs are computed under a *null joint distribution* $\mathcal{Q}_0$. The choice of the null distribution needs to be such that control of the error rate under this *assumed* distribution does indeed imply control of the error rate under the *actual* data generating distribution. In our current setting, the numbers of IP-enriched and control hybridization samples are too small (only 6 observations in each group and about 1.1 million tests) for applying resampling-based approaches to estimate the joint null distribution of the test statistics. For this reason, we turn our attention to model-based multiple testing.

3.1.1 Controlling the family-wise error rate: nested-Bonferroni adjustment

The family-wise error rate (FWER) is the probability of rejecting at least one true null hypothesis, i.e., the probability of making at least one Type I error. In order to control the FWER at level α, we select a common cut-off c for the scan statistics so that

$$P_{\mathcal{Q}_0}\left(\max_{i \in \{1, \cdots, N-w+1\}} T^*_{i,n} > c\right) \quad \leq \quad \alpha.$$

Bonferroni adjustment. It is common practice to control the FWER by a Bonferroni adjustment on the significance level. Here, we assume that the joint distribution $\mathcal{Q}_0$ of $T_{i,n}$ is such that the scan statistics $T^*_{i,n}$ are identically distributed with marginal distribution function $\mathcal{F}_0$. The Bonferroni-adjusted common cut-off is given by

$$c_B = c_B(\alpha; N, w, \mathcal{F}_0) \equiv \mathcal{F}_0^{-1}\left(1 - \frac{\alpha}{(N - w + 1)}\right). \tag{2}$$

This adjustment relies on Boole's inequality to derive an upper bound for the FWER, which is based only on the marginal distribution $\mathcal{F}_0$ of the scan statistics and hence ignores any dependence among these statistics. As a result, the Bonferroni adjustment can be very conservative.

Nested-Bonferroni adjustment. To gain some improvement on the Bonferroni adjustment, we consider applying this adjustment in a nested fashion. We firstly define random variable $Z_{k,n}$ as maxima of scan statistics over blocks of w probes. That is,

$$Z_{k,n} = \max_{i\in\{k,k+1,\cdots,k+w-1\}} T^*_{i,n},$$

where $k \in \mathcal{K} = \{1, w+1, 2w+1, \cdots, m_1\}$ and $m_1 = \lceil (N-w+1)/w \rceil$. Here, the notation $\lceil a \rceil$ refers to ceiling of a, i.e., the smallest integer that is greater than or equal to a. We assume that the statistics $Z_{k,n}$ are identically distributed with marginal distribution function $\mathcal{G}_0$. Next, we note the following relation between the distribution functions of the random variables $Z_{k,n}$ and $T^*_{i,n}$, under the test statistics null distribution $\mathcal{Q}_0$:

$$\begin{aligned} P_{\mathcal{Q}_0}(Z_{k,n} > c) &= P_{\mathcal{Q}_0}(\max_{i\in\{k,k+1,\cdots,k+w-1\}} T^*_{i,n} > c) \\ &\leq \sum_{i=k}^{k+w-1} P_{\mathcal{Q}_0}(T^*_{i,n} > c) \qquad (3) \\ \implies 1 - \mathcal{G}_0(c) &\leq w(1-\mathcal{F}_0(c)), \qquad (4) \end{aligned}$$

where the upper bound (3) follows from Boole's inequality. We have

$$\begin{aligned} P_{\mathcal{Q}_0}\left(\max_{i\in\{1,\cdots,N-w+1\}} T^*_{i,n} > c\right) &= P_{\mathcal{Q}_0}\left(\max_{k\in\mathcal{K}} Z_{k,n} > c\right) \\ &\leq \sum_{k\in\mathcal{K}} P_{\mathcal{Q}_0}(Z_{k,n} > c) \qquad (5) \\ &= |\mathcal{K}|\,(1-\mathcal{G}_0(c)), \qquad (6) \end{aligned}$$

where $|\mathcal{K}| = m_1 = \lceil (N-w+1)/w \rceil$ and the upper bound (5) again follows from Boole's inequality. This gives us the nested-Bonferroni adjusted common cut-off

$$c_{NB} = c_{NB}(\alpha; N, w, \mathcal{G}_0) \equiv \mathcal{G}_0^{-1}\left(1 - \frac{\alpha}{\lceil (N-w+1)/w \rceil}\right). \qquad (7)$$

Comparison of the Bonferroni and nested-Bonferroni adjustments. We note that, from the inequality in Equation (4), if $\lceil (N-w+1)/w \rceil \times w = N-w+1$, then

$$\left\lceil \frac{N-w+1}{w} \right\rceil (1-\mathcal{G}_0(c)) \leq \left\lceil \frac{N-w+1}{w} \right\rceil w(1-\mathcal{F}_0(c)) \leq (N-w+1)(1-\mathcal{F}_0(c)). \qquad (8)$$

Moreover, we have that $\lceil (N-w+1)/w \rceil \times w \leq N+1$, hence the conservative upper bound for the nested-Bonferroni adjustment is

$$\left\lceil \frac{N-w+1}{w} \right\rceil (1-\mathcal{G}_0(c)) \leq \left\lceil \frac{N-w+1}{w} \right\rceil w(1-\mathcal{F}_0(c)) \leq (N+1)(1-\mathcal{F}_0(c)).$$

Here, note that if $N-w+1$ is not an exact multiple of w, the last $Z_{k,n}$ statistic to be constructed for $k = m_1$ is based on fewer than w scan statistics and hence has a different null distribution than $\mathcal{G}_0$. However, this is just one test statistic among $N \approx 1.1$ million test statistics. Thus, without loss of generality we assume that $\lceil (N-w+1)/w \rceil \times w = N-w+1$. We define nominal (Boole's upper bound) Type I error rates for the nested-Bonferroni and Bonferroni procedures as

$$\begin{aligned} FWER_{NB}(c \mid N, w, \mathcal{G}_0) &= \left\lceil \frac{N-w+1}{w} \right\rceil (1-\mathcal{G}_0(c)), \qquad (9) \\ FWER_{B}(c \mid N, w, \mathcal{F}_0) &= (N-w+1)(1-\mathcal{F}_0(c)). \qquad (10) \end{aligned}$$

Note that solving for c for a given α in $FWER_{NB}(c \mid N, w, \mathcal{G}_0) = \alpha$ gives us cut-off c_{NB} and in $FWER_B(c \mid N, w, \mathcal{F}_0) = \alpha$ returns c_B. Hence, from (8), we have $c_{NB} \leq c_B$ and the nested-Bonferroni adjustment is less conservative than the Bonferroni adjustment.

In Figure 6, values of the nominal Type I error rates $FWER_B$ and $FWER_{NB}$ are plotted for a range of cut-off values c based on simulated data. In this simulation, $T_{i,n}, i \in \{1, \cdots, N\}$, are independent two-sample Welch t-statistics based on $n_1 = 6$ and $n_2 = 6$ independent observations from a normal distribution with mean 0 and standard deviation 1. The marginal distribution functions $\mathcal{F}_0$ and $\mathcal{G}_0$ are estimated by their empirical distributions based on $B = 10,000,000$ simulated observations using $w = 10$.

[Figure 6 about here.]

In the plots of Figure 6, the abscissa x corresponding to the ordinate $y = 0.05$ (horizontal green line) is the cut-off value of a given procedure at nominal level $\alpha = 0.05$. The cut-offs obtained by the two procedures at level $\alpha = 0.05$ with different numbers of probes (N), hence different numbers of test statistics, are summarized in Table 3. We observe that, for all numbers of test statistics, the nested-Bonferroni adjusted cut-off is smaller than the Bonferroni adjusted cut-off. The presented results are for blip size $w = 10$, but similar results were obtained for other values of w, e.g., $w \in \{2, 5, 10, 20\}$. This illustrates that the nested-Bonferroni adjustment has slightly more power than the Bonferroni adjustment. For $N = 1,000,000$ tests, this result does not seem to hold uniformly in α, i.e., the ranking of the curves changes for very small α levels. However, this is much related to the fact that the two distribution functions $\mathcal{F}_0$ and $\mathcal{G}_0$ need to be estimated based on a larger number of simulated observations to accurately estimate the small tail probabilities. One critical point in this setting is that since $T^*_{i,n}$, $i \in \{1, \cdots, N - w + 1\}$, are at least locally dependent, i.e., $T^*_{i,n}$ and $T^*_{i+j,n}$ are dependent if $j < w$, then so are $Z_{k,n}$, $k \in \mathcal{K}$. In the appendix, we investigate the independence scenario where the set of $T^*_{i,n}$, and hence the set of $Z_{k,n}$ are independent, and show that the two adjustments are equivalent for small blip sizes w.

[Table 3 about here.]

Estimating the null distribution of the test statistics $Z_{k,n}$. We now describe how to estimate the null distribution $\mathcal{G}_0$ of $Z_{k,n}$ for a given blip size w using the parametric bootstrap:

1. Fit normal distributions $\hat{\mathcal{Q}}^c_0$ and $\hat{\mathcal{Q}}^t_0$ to the centered control and IP-enriched hybridization observations $\{Y_{j,i,k} - \bar{Y}_{j,i} : i \in \{1, \cdots, N\}\}$, $j = 1, 2$, respectively. This step pools the observations of all N probes across n_j replicates when fitting a normal distribution for the j-th hybridization group.
2. For $b = 1$ to $b = B$ repeat: Generate $2wn_1$ control observations from the distribution $\hat{\mathcal{Q}}^c_0$ and $2wn_2$ treatment observations from the distribution $\hat{\mathcal{Q}}^t_0$. Compute the first w scan statistics from these $2w$ observations and set $Z_{b,n}$ to the maximum of these.

The common cut-off for nested-Bonferroni adjusted control of the FWER at level α is given by

$$c_{NB}(\alpha; N, w, \hat{\mathcal{G}}_0) = \hat{\mathcal{G}}_0^{-1}(1 - \alpha/\lceil (N - w + 1)/w \rceil),$$

where $\hat{\mathcal{G}}_0$ represents the empirical distribution of B simulated test statistics $\{Z_{1,n}, \cdots, Z_{B,n}\}$. Hence, the set of rejections, identifying bound probes, is given by

$$\mathcal{R}_n(\alpha; w) = \left\{ i \in \{1, \cdots, N - w + 1\} : T^*_{i,n} > c_{NB}(\alpha; N, w, \hat{\mathcal{G}}_0) \right\}.$$

Alternatively, the nonparametric bootstrap might also be employed, by pooling the centered intensity measures for each hybridization j, $j = 1, 2$, in order to estimate the null distribution $\mathcal{G}_0$.

Remark. Another complication of the ChIP-Chip dataset that we have yet to mention is the so-called *gap structure*. This gap structure refers to the fact that the probe-pairs are not evenly spaced throughout the two chromosomes. Since the probes on the array represent non-repeating genomic locations on two chromosomes, there might be large gaps between the genomic locations of any two adjacent probes, i.e., when going from one probe to the next closest in terms of genomic location, there might be more than a 1kb jump. This may cause the scan statistic to be computed across a gap which by definition cannot represent a valid blip (recall that a blip spans a continuous region of average length 1kb). Moreover, the filtering process described in Section 2.3 generates more such gaps in the data. When taking into account gaps, the total number of scan test statistics is $N - M(w-1)$, where M represents the total number of disjoint chunks generated by marking the locations where the gaps occur. Hence, the corresponding nested-Bonferroni adjusted common cut-off is given by

$$c^g_{NB}(\alpha; N, M, w, \mathcal{G}_0) = \mathcal{G}_0^{-1}\left(1 - \frac{\alpha}{\lceil (N - M(w-1))/w \rceil}\right). \tag{11}$$

3.1.2 Controlling tail probabilities for the proportion of false positives

van der Laan *et al.* (2004) recently proposed an augmentation of FWER-controlling procedures for control of the tail probability, TPPFP(q), that the proportion of false positives among the rejected hypotheses exceeds a user supplied value $q \in (0,1)$. We use this procedure to augment the above nested-Bonferroni adjustment procedure as follows: Let $R_n(\alpha; w)$ denote the number of rejected hypotheses, i.e., identified bound probes, by controlling the FWER at level α for a given blip size w. For a given q, define

$$A_n(q, \alpha; w) = \max\left\{j \in \{0, \cdots, N - R_n(\alpha; w)\} : \frac{j}{j + R_n(\alpha; w)} \leq q\right\}. \tag{12}$$

The set of rejections $\mathcal{R}_n(\alpha; w)$ is then augmented as follows: Consider all the hypotheses that are not rejected by the FWER controlling procedure and choose the first $A_n(q, \alpha; w)$ hypotheses with the largest test statistics. Denote this set by $\mathcal{A}_n(q, \alpha; w) \subseteq \mathcal{R}_n(\alpha; w)^c$. Then, the final set of rejected hypotheses for controlling TPPFP(q) at level α for a user supplied q is given by

$$\mathcal{R}^+_n(q, \alpha; w) = \mathcal{R}_n(\alpha; w) \cup \mathcal{A}_n(q, \alpha; w).$$

3.1.3 Controlling the false discovery rate

For controlling the false discovery rate (FDR), i.e., the expected proportion of false positives among the rejected hypotheses, we apply the Benjamini and Hochberg (1995) step-up method. This method takes as input only the marginal p-values for each probe. The p-values corresponding to the scan statistics can be estimated by assuming a standard Student's t-distribution for the Welch two-sample t-statistics and then making a normal approximation for the sum of w such test statistics (analytical approximation to the null distribution $\mathcal{F}_0$ of the scan statistics). Alternatively, the parametric bootstrap procedure described in Section 3.1.1 might also be used. Let $p_{(1)}, \cdots, p_{(N-w+1)}$ be the ordered marginal p-values for the $N - w + 1$ scan statistics $T^*_{i,n}$ and denote the corresponding indices by $r(1), \cdots, r(N-w+1)$. Then, the set of rejections identified by the Benjamini and Hochberg (1995) procedure is given by

$$\mathcal{R}^{BH}_n(q; w) = \left\{r(i), i \in \{1, \cdots, N - w + 1\} : p_{(i)} \leq \frac{q}{N - w + 1} i\right\},$$

where q is the a priori specified nominal false discovery rate.

3.2 Cross-validation in a piecewise constant mean regression model to choose the blip size

The aforementioned multiple testing approaches rely on the blip size parameter w. The theoretical calculations in Section 2.1 suggest a blip size of about 30 probes; however, visualizing the experimental data suggests a smaller blip size of around 10 probes. We use a regression-based strategy in combination with Monte-Carlo cross-validation to determine an appropriate blip size.

Let $\{Y_i, i \in \{1, \cdots, N\}\}$ represent statistics of interest for N probes. In our context, Y_i represents the two-sample Welch t-statistic corresponding to probe i, that is, $Y_i \equiv T_{i,n}$. We assume that there are two groups in the data, i.e., the *bound* and *unbound* groups as mentioned earlier, and that the observations Y_i are independent and identically distributed within each group. Let $\mathcal{A}^* = \{\theta_1, \cdots, \theta_L\}$ represent the start positions of L blips of bound probes, i.e., probes $\{\theta_i, \cdots, \theta_i + w - 1\}$, $i \in \{1, \cdots, L\}$, are in the bound group. Define

$$\mathcal{A} = \bigcup_{i=1}^{L} \{\theta_i, \cdots, \theta_i + w - 1\}$$

to be the set of bound probes. We consider the following piecewise constant mean regression model

$$E[Y_i] = I(i \notin \mathcal{A})\mu_1 + I(i \in \mathcal{A})\mu_2, \tag{13}$$

where μ_1 and μ_2 represent the population means of the unbound and bound groups, respectively. We need to estimate the population means μ_1 and μ_2 and the start positions $\mathcal{A}^* = \{\theta_1, \cdots, \theta_L\}$ in this model. Given $\mathcal{A}^*$, the least squares estimates of the means μ_1 and μ_2 are simply the empirical means of the observations in each group. Since the blip start positions, $\mathcal{A}^*$, are nondifferentiable parameters, we use a forward stepwise selection algorithm to choose among the set of all possible start positions. The algorithm starts with zero blips and adds blips of size w one at a time. Adding a blip corresponds to putting w consecutive probes in the bound group. The blips are added according to their improvement on the residual sum of squares, i.e., the empirical risk for the squared error loss function. Specifically, at a given iteration, the algorithm considers all possible blip start positions. For each candidate position, it reestimates the mean parameters by placing a blip starting at that particular position and then chooses the position which decreases the residual sum of squares the most. Such stepwise algorithms, in general, require a stopping criterion or must go up to an a priori specified number of blips. We use an upper bound on the number of blips as the stopping criteria. Then, Monte-Carlo cross-validation is employed to select the number of blips as well as the blip size from a set of candidate blips sizes. Details on this cross-validation procedure are given in Section 4 with the p53 data.

4 Results

4.1 Simulations

We performed simulation studies to investigate the properties of the proposed methods for analyzing ChIP-Chip data. Specifically, we considered the following four multiple testing procedures:

- NB-FWER: control of the FWER by the nested-Bonferroni adjustment,
- B-FWER: control of the FWER by the Bonferroni adjustment,
- VDP-TPPFP: augmentation method of van der Laan *et al.* (2004) for control of the tail probability of the proportion of false positives, applied to the NB-FWER procedure,

- BH-FDR: Benjamini and Hochberg (1995) step-up procedure for the control of the false discovery rate.

Comparison of the NB-FWER, B-FWER, VDP-TPPFP, and BH-FDR procedures in terms of Type I error control. We firstly compared the four approaches above in terms of their *actual* Type I error rates. Such a comparison between NB-FWER and B-FWER provides guidance as to which procedure is more or less conservative since they both aim to control the FWER. In the first simulation study (Simulation 0), unbound probe-level intensity data, i.e., $Y_{j,i}$, $i \in \{1, \cdots, N\}$, for $j = 1$, were generated from a normal distribution with mean 0 and variance 1, and bound probe-level intensity data from a normal distribution with mean 2 and variance 1. A total of 500 independent simulated datasets were generated. Each dataset contained about $N = 2000$ probes[1], and $n_1 = 6$ treatment and $n_2 = 6$ control samples were generated for each probe. A true blip size of $w = 10$ was used and each dataset contained 12 such blips. In Table 4, we report the actual Type I error rates for the four multiple testing procedures for assumed blip sizes $w \in \{1, 2, 5, 10, 20\}$. The nominal FWER for both the B-FWER and NB-FWER procedures is $\alpha = 0.05$. The nominal Type I error rate for the VDP-TPPFP and BH-FDR procedures is $q = 0.05$. For the Bonferroni adjustment (B-FWER) and BH-FDR procedure, two approaches were implemented for estimating the marginal null distribution, $\mathcal{F}_0$, of the scan test statistic: (1) parametric bootstrap and (2) normal approximation. This simulation study illustrates a couple of interesting results. Firstly, we observe that, at small blip sizes, e.g., $w \in \{2, 5\}$, estimating the null distribution by a normal approximation causes the actual Type I error rates of the B-FWER and BH-FDR procedures to be much larger than the nominal Type I error rate. As a result, both of these procedures seem to be less conservative, i.e., identify a larger number of bound probes, than the other two procedures. For example, for the B-FWER procedure, the actual error rate is 0.326 with $w = 2$, whereas the nominal family-wise error rate is 0.05. It seems that the effect of the normal approximation is less dramatic on BH-FDR. When the parametric bootstrap is used for estimating the marginal null distribution $\mathcal{F}_0$, we see that NB-FWER is, in general, slightly less conservative than B-FWER. The VDP-TPPFP procedure turns out to be quite conservative, i.e., the actual Type I error rate is much smaller than the nominal Type I error rate of $q = 0.05$. However, if we look at the actual number of rejections and correct rejections for these procedures (Figures 7, 8, 9, 10, and 11), we observe that the VDP-TPPFP procedure has larger numbers of rejections and correct rejections than NB-FWER since it is augmenting this procedure. Additionally, the numbers of rejections and correct rejections are slightly higher for NB-FWER than for B-FWER.

[Table 4 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

[1]When generating the probe-level data, sets of bound and unbound probes are generated in a consecutive manner, based on a Bernoulli random variable indicating whether the next set is a bound or unbound sequence of probes, until the desired total number of blips is reached. The number of probes in each set of unbound probes is uniformly distributed over the range $[100, 200]$.

Comparison of the NB-FWER, VDP-TPPFP, and BH-FDR procedures in terms of power. In the following second set of simulations, we compared the three methods, NB-FWER, VDP-TPPFP, and BH-FDR, under various models, in terms of their performance at finding correct blips. In all of the simulations, the probe-level intensity data of the unbound group were generated from a normal distribution with parameters (μ_1, σ_1) and the bound group probe-level data were generated from a normal distribution with parameters (μ_2, σ_2). Another parameter of interest in these simulations is the true blip size w. In practice, not all blips are of the same size. For this reason, we also investigated the case where the blip sizes are variable. As performance measures, we report the number of correctly identified blips as well as the number of probes for which the null hypotheses were correctly rejected, since for practical purposes it is sufficient to report the locations of the blips rather than their exact boundaries, i.e., the identified locations could be slightly shifted compared to the true boundaries. For all simulations, $n_1 = 6$ control and $n_2 = 6$ treatment intensity measures were generated for each probe. All of the results are based on $B = 100$ independently generated datasets, where the null distribution $\mathcal{G}_0$ of the test statistic $Z_{k,n}$ is estimated based on $100,000$ independent observations of the random variable Z. The nominal FWER, TPPFP, and FDR are set at $\alpha = 0.05$, $q = 0.05$, and $q = 0.05$, respectively. The simulation setting and results are summarized below. Tables 5, 6, 7, and 8 present these results in detail. Moreover, specificity and sensitivity measures for each of the multiple testing procedures, at all considered blip sizes w, are displayed in Figure 12. Simulation IV includes about 3000 probe, while all other simulations include about 2000 probes.

[Figure 12 about here.]

- **Simulation I, fixed blip size, high separation:** In this simulation, the blip size was kept constant at $w = 10$ and probe-level intensity data for 12 blips were generated from a normal distribution with parameters $\mu_2 = 2$ and $\sigma_2 = 0.75$, whereas non-blip probe-level intensity data were generated from a normal distribution with parameters $\mu_1 = 0$ and $\sigma_1 = 1$. These two distributions are quite separated, so that the blips are visible when the probe-level test statistics are plotted against their corresponding genomic locations. The results of this simulation are summarized in Table 5, where the mean numbers of rejected hypotheses, correctly rejected hypotheses, identified blips, and correctly identified blips are reported along with their corresponding standard deviations over $B = 100$ independent simulated datasets. We observe that both NB-FWER and VDP-TPPFP perform very well in identifying the correct blips with all but an assumed blip size $w = 1$, for which rejections require higher test statistics. Interestingly, the BH-FDR procedure performs well at all blip sizes in terms of finding the correct blips; however, as expected, the number of false positives is much higher compared to the FWER-controlling procedure NB-FWER.

- **Simulation II, fixed blip size, low separation:** In this simulation, the blip and non-blip prob-level intensity distributions were chosen close to each other, making it harder to visualize the blips. Probe-level intensity data for a total of 12 blips of size $w = 10$ were generated from a normal distribution with parameters $\mu_2 = 1.5$ and $\sigma_2 = 1$, and the non-blip probe-level intensity data $Y_{1,j}$ were generated from a normal distribution with parameters $\mu_1 = 0$ and $\sigma_1 = 1$. The results of this simulation are summarized in Table 6. Contrary to the first simulation, where the distributions of the blip and non-blip probe-level intensities were well separated, all three procedures failed at identifying the correct blips with an assumed blip size of $w = 1$. This result illustrates the necessity of taking into account the spatial structure of the data in multiple testing procedures. The performance of BH-FDR increases dramatically with a higher assumed blip size and all blips are correctly identified even with an assumed blip size as low as $w = 2$. However, the number of false positives increases, too. All three methods have a good performance in identifying

the correct blips, with assumed blip sizes $w = 5$ and $w = 10$. Their performance starts decreasing with a higher blip size of $w = 20$, since at this higher blip size the test statistics are averaging over ten blip and ten non-blip probes.

- **Simulation III, variable blip size (uniform blip size distribution), low separation:** In this simulation, we investigated the effect of variable blip size. We randomly generated 12 blip sizes from a uniform blip size distribution with support $[5, 16]$, hence with a mean blip size of 10.5. The data generating distributions of blip and non-blip probe-level intensity data are the same as in Simulation II. The results of this simulation study are reported in Table 7. The overall performance of all three methods is comparable with the results of Simulation II, with the following two exceptions. Firstly, with varying blip size, the standard errors of the mean numbers of correctly identified blips increase for all the three methods. Secondly, procedures with an assumed blip size of 5, which is smaller than the true mean blip size of 10.5, are performing slightly better than those based on the true mean blip size. This might suggest that when the true blip sizes are variable, as is usually the case in practice, using a slightly conservative assumed blip size might yield better results than using the true average blip size.

- **Simulation IV, variable blip size (heavy right-tailed blip size distribution), low separation:** In this final simulation, we investigated the effect of having a heavy right-tailed blip size distribution. We varied the blip size distribution to be a truncated gamma distribution with mean and standard deviation of ten probes in a blip (the truncation is for the purpose of generating an integer number for the blip size). A total of ~ 3000 probes, using the probe-level intensity data distributions of Simulation II, were generated for each of the $n_1 = 6$ and $n_2 = 6$ treatment and control samples and these included a total of 20 blips. The results of this simulation study are summarized in Table 8. These results do not deviate significantly from the results of Simulation III, where a uniform blip size distribution was employed. One interesting result that again emerges is that an assumed blip size of 5, which is smaller than the true mean blip size, performs slightly better than the mean blip size.

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

[Table 8 about here.]

4.2 Analysis of p53 ChIP-Chip high density oligonucleotide array data

4.2.1 Monte-Carlo cross-validation chooses a blip size of 10 probes

Before applying the multiple testing procedures discussed in Section 3.1, we firstly choose an appropriate blip size using cross-validation with a piecewise constant mean regression model as described in Section 3.2. For the p53 ChIP-Chip dataset, there are a total of 6 replicate experiments in both the treatment and control groups. Among the 6 replicates, there are 3 hybridization replicates for each of two IP replicates. To accommodate the small sample size and the nested structure of the replicates, we apply the following cross-validation scheme for both the control and treatment group probe-level data. We use Monte-Carlo cross-validation and divide the six observations into a validation set of size 2 and a training set of size 4. Each validation sample consists of one hybridization replicate from the first IP sample and one from the second IP sample. All possible hybridization combinations ($3 \times 3 = 9$) were

covered. At each step of the cross-validation, the blip boundaries and the two population means μ_1 and μ_2 are estimated on the training sample and the corresponding empirical risk with squared error loss function, i.e., mean squared error, is computed over the validation sample. The average cross-validated risk for different blip sizes as a function of the number of blips is displayed in the left panel of Figure 13 for chip A, which covers about 2/3 of chromosome 21 (blip size selection is based on only chip A for computational reasons). We see that the cross-validated risk keeps decreasing as more and more blips are added. This suggests that if we were to select the number of blips based on this cross-validated risk criterion, we would potentially choose as many blips as possible (the maximum allowed size is 500 in this example) for blip sizes such as 1, 2, and 10. However, if we look at the right panel of Figure 13, which zooms into the first 30 blips, we see that the ranking of the blip sizes according to cross-validated risk becomes constant after about the first 25 blips. This suggests that if the true number of blips is smaller than ~ 16, a blip size of 10 will be selected as the best blip size by cross-validation.

To get an idea about the number of blips in chip A, we applied the three multiple testing procedures, NB-FWER, VDP-TPPFP, and BH-FDR, using only chip A and visualized all of the identified blips, for each blip size, to determine how many *looked like real blips*. Our notion of *real blips*, as mentioned in the introduction, consists of a cloud of probes (> 1 probes) that have higher test statistics than their surroundings. The plots of the blips identified by the NB-FWER procedure, for blip sizes $w \in \{1, 2, 10, 20, 30\}$, are given in Figures 14, 15, 16, 17, and 18, respectively. These plots suggest that there are a maximum of 13 *real blips* identified by the multiple testing procedures. Table 9 summarizes the results for all the blip plots. Now, if we look at the right panel of Figure 13, we see that the blip size of $w = 10$ has the minimum cross-validated risk when the total number of blips is 13. In conclusion, this cross-validation approach in a piecewise constant mean regression model provides a simple rough guide for choosing the blip size for the multiple testing procedures.

[Table 9 about here.]

[Figure 13 about here.]

[Figure 14 about here.]

[Figure 15 about here.]

[Figure 16 about here.]

[Figure 17 about here.]

[Figure 18 about here.]

4.2.2 Multiple testing procedures NB-FWER, VDP-TPPFP, and BH-FDR identify 254, 269, and 719 blips with a blip size of $w = 10$

We applied the multiple testing methods discussed in Section 3.1 with four different blip sizes $w \in \{1, 10, 20, 30\}$ to analyze the ChIP-Chip data for transcription factor p53. Note that using a blip size of one corresponds to ignoring the spatial data structure. The nominal false positive rate α is set to 0.05 for controlling the FWER and q is also set to 0.05 for controlling the TPPFP and FDR.

In order to investigate the effect of the gap structure in the data, we applied the multiple testing procedures in the following two set-ups: (1) assuming no gaps, such that every scan statistic represents

a measurement for a valid potential blip and (2) acknowledging the existence of gaps and hence fragmenting the data so that no scan statistic corresponds to an invalid blip. In the first analysis, we have $N = 1,029,389 - w + 1$ scan statistics for each assumed blip size w. In the second approach, given that there are M segments, we have $N - M(w-1)$ scan statistics using a blip size of w. If a given fragment includes a smaller number of probes than the blip size, then the scan statistics cannot be computed; hence this causes the total number of scan statistics to be very different for various blip sizes, i.e., $w = 1$ versus $w = 30$. In order to avoid such a situation, we discarded fragments that had a smaller number of probes than the maximum employed blip size of 30. This left a total of $M = 5,584$ fragments and a total number of $997,377$ scan statistics with a blip size of $w = 1$.

When the gap structure is ignored, the NB-FWER procedure identified 128, 254, 188, and 145 blips, for blip sizes of 1, 10, 20, and 30, respectively (Table 10). In contrast, the BH-FDR procedure identified 553, 719, 355, and 225 blips. When the gap structure is taken into account, NB-FWER identified 121, 230, 154, and 112 blips, for blip sizes of 1, 10, 20, and 30, respectively; BH-FDR identified 531, 651, 306, and 178 blips (Table 11). The blips identified by accommodating the gap structure are among the blips identified by ignoring the gap structure. This indicates that ignoring the gap structure is not causing any blips to be missed.

4.2.3 Enrichment for sequences proximal to 5' exons, near or within CpG islands, and within coding regions

To further investigate the blips identified by the NB-FWER procedure for control of the FWER, we annotated the blip locations using the *UCSC Genome Browser* at `www.genome.ucsc.edu` (Kent *et al.*, 2002). The conventional model for transcriptional regulation implies that the regulatory elements are generally located in the 5' end promoter region of the genes. Moreover, many human promoters are located near CpG islands (Ioshikhes and Zhang, 2000) and distal modules (enhancers and silencers) can lie many kbs downstream (3' end) of coding regions and within introns. Tables 10 and 11 report the detailed annotation of the blips identified by ignoring the gap structure and by taking into account the gap structure, respectively. We found that using an assumed blip size of $w = 10$ identified the largest number of blips (a total of 14) falling within 3kb of the 5' end of a known gene. Similarly, the largest number of blips within 3kb of known CpG islands was identified with an assumed blip size of $w = 10$. A blip size of $w = 10$ was also selected as the best blip size by the Monte-Carlo cross-validation procedure of Section 4.2.1. For the BH-FDR procedure, we observed that a substantial proportion of the identified blips (35.26% with $w = 1$; 32.13% with $w = 10$; 31.55% with $w = 20$; 28.00% with $w = 30$) fell within a known/annotated gene but a very small percentage (3.25% with $w = 1$; 2.09% with $w = 10$; 1.13% with $w = 20$; 0.44% with $w = 30$) fell within exons. These percentages were even smaller for the NB-FWER procedure.

The chromosomal distribution of the blips identified by the NB-FWER procedure is given in Table 12. We further filtered out the 254 blips identified under the no gap assumption. This filtering only kept the blips for which the proportion of unique 25mers is at least 0.8. The purpose of such a filtering is to account for local repeats that might have been missed out by the filtering process discussed in Section 2.3. This left a total of 221 blips to identify regulatory motifs. Below are LocusLink (`www.ncbi.nih.gov/LocusLink`) descriptions of the genes that are within 1kb upstream or downstream of blips identified by the NB-FWER procedure.

[Table 10 about here.]

[Table 11 about here.]

[Table 12 about here.]

Genes whose 5' ends are within 1kb of a blip.

- BC022865 (LocusID=539): ATP5O: ATP synthase, H+ transporting, mitochondrial F1 complex, O subunit (oligomycin sensitivity conferring protein).
- AL137448 (LocusID=54101): ANKRD3: ankyrin repeat domain 3.
- BC007658 (LocusID=23786): BCL2L13 BCL2-like 13 (apoptosis facilitator).
- M30474 (LocusID=2679): GGT2 gamma-glutamyltransferase 2.
- L20493 (LocusID=2678): GGT1 gamma-glutamyltransferase 1.
- BC014912 (LocusID=8664): EIF3S7 eukaryotic translation initiation factor 3, subunit 7 zeta, 66/67kDa.

Genes whose 3' ends are within 1kb of a blip.

- AB001535 (LocusID=7226): TRPM2 transient receptor potential cation channel, subfamily M, member .

4.2.4 Enrichment for p53 consensus binding sequence in identified blips

The p53 DNA binding site primarily consists of the consensus sequence `RRRCW` (will be represented by $\triangleright$) and its reverse complement `WGYYY` (will be represented by $\triangleleft$), arranged as `RRRCWWGYYY` ($\triangleright\triangleleft$). This palindromic sequence commonly occurs in a pair, where the members of the pair are separated by a spacer of length 0 to 15bps. The variable nature of the p53 consensus sequence complicates the identification of the binding sites (Hoh *et al.*, 2002). Wang *et al.* (1995) showed that the tetrameric p53 protein can bind to various arrangements of multiple copies of the consensus `RRRCW`. Their results indicate that the efficiency of DNA binding is proportional to the number of repeats and related to the orientation of the consensus sequence `RRRCW`. In particular, four repeats yield the highest efficiency and three adjacent consensus sequences perform better than three non-adjacent consensus sequences.

We focus on the 221 blips identified by the NB-FWER multiple testing procedure controlling the FWER at nominal level 0.05 using a blip size of $w = 10$. Among these 221 blips only 4 contain an exact match to the consensus `RRRCWWGYYYN {0-15} RRRCWWGYYY`, where $\{0-15\}$ represents a variable length spacer between the two dimers. However, more than half of the blips contain one or more copies of the 10mer `RRRCWWGYYY`.

We compared our identified blips to the 14 experimentally verified transcription factor binding site regions of Cawley *et al.* (2004). Our blips include 13 of these regions and the remaining region is among the ones identified by the procedures controlling the TPPFP and FDR. Overall, 23 of the 48 blips in Cawley *et al.* (2004) are among ours, even though Cawley *et al.* (2004) used whole cell extract as the control group and full length p53 protein `p53FL` in the ChIP-Chip experiments. Using their 13 experimentally verified regions, which are also among ours, we further filtered our 221 blips according to the following criteria:

- Blips should have mean two-sample Welch t-statistics at least as large as the minimum of the mean test statistics for the 13 experimentally verified blips.

- Blips should have mean scan statistics at least as large as the minimum of the mean scan statistics for the 13 experimentally verified blips.
- Blips should be composed of unique probes, no local repetitions are allowed.

Applying these criteria identified 49 blips out of $221 - 13 = 208$ blips. Annotation information for these blips is given in Table 13. For regulatory motif analysis, the start and end sites of the blips are extended so that the total blip region covers 2kb.

[Table 13 about here.]

We then searched the blips identified by our NB-FWER procedure for consensus arrangements listed in Wang *et al.* (1995). Various arrangements used to scan the identified blips are as follows:

- $\triangleright \triangleleft \triangleright$ represents the pattern `RRRCWWGYYYRRRCW`,
- $\triangleleft \triangleright \triangleleft$ represents the pattern `WGYYYRRRCWWGYYY`,
- $\triangleright \triangleleft \{0-15\} \triangleleft$ corresponds to `RRRCWWGYYY` and `WGYYY` separated by 0 to 15bps,
- $\triangleright \triangleleft \{0-15\} \triangleright$ corresponds to `RRRCWWGYYY` and `RRRCW` separated by 0 to 15bps,
- $\triangleleft \{0-15\} \triangleright \triangleleft$ corresponds to `WGYYY` and `RRRCWWGYYY` separated by 0 to 15bps,
- $\triangleright \{0-15\} \triangleright \triangleleft$ corresponds to `RRRCW` and `RRRCWWGYYY` separated by 0 to 15bps.

Table 14 displays the exact number of blips that have any of the above iterated patterns among the 13 blips experimentally *verified* by Cawley *et al.* (2004), our 49 *filtered* blips, and *all* of the 221 identified blips. We observe that 67.35% of the 49 blips have at least one copy of the consensus 10mer, whereas this percentage is 61.5% for the 13 experimentally verified blips and 53.39% for all 221 identified blips. It is also interesting to note that none of the experimentally verified blips have an exact match to the 20mer consensus and more than 50% of the sequences with at least one copy of the 10mer have a copy of the 5mer `RRRCW` or its reverse complement in the immediate vicinity. Inga *et al.* (2002) investigated p53 transactivation capacity for 26 different p53 response elements (DNA binding sites) under special conditions. Their results indicate that DNA sites with as many as 4bp mismatches to the 20mer consensus can be functional and enable high levels of transactivation. This work also shows that having a `CAGT` core in both parts of the dimer, i.e., at positions 4, 5, 6, 7, and 14, 15, 16, 17, leads to tighter p53 tetramer binding by affecting the bending properties of the DNA. We investigated whether the blips contained sequences showing these characteristics and the results are summarized in Table 15. About 64% of the 221 identified blips, 71% of the 49 filtered blips, and 38% of the 13 experimentally verified blips have sequences that have at most 2 mismatches to the 20mer consensus.

We also ran the MDscan (Liu *et al.*, 2002) and BioProspector (Liu *et al.*, 2001) motif finding methods separately for (1) the 13 experimentally verified binding site regions, (2) our 49 filtered blips, and (3) all of the 221 blips. In none of these three cases, were the methods able to produce a pattern that resembled the p53 consensus binding site. Recently, Yin *et al.* (2003) showed that the 12bp site `CCCCACGTGAGG` was crucial for induction of the PAC1 promoter activity by p53. None of our blips have exact matches to this site and the above motif finding programs did not identify position weight matrices that could generate sites similar to this site. However, this finding about the PAC1 and p53 relation suggests that there might be other consensus sequences for p53 which are very different than the ones identified so far in the literature.

[Table 14 about here.]

[Table 15 about here.]

5 Conclusions and future work

We have proposed multiple testing methods for analyzing ChIP-Chip data from high density oligonucleotide array experiments. In particular, we propose to use a scan statistic in order to take into account the spatial structure of this genomic data. Simulation studies illustrated that incorporating the spatial information facilitates the detection of the bound probes substantially. For control of the family-wise error rate, we derived a nested-Bonferroni adjustment that is slightly less conservative than the traditional Bonferroni adjustment when the test statistics are dependent. Another insight from the simulation studies is that the proposed methods adapt well to variable blip sizes. That is, procedures using a fixed blip size, when in truth the blip size is variable, perform reasonably well in detecting true blips. Application of these methods identified more potential targets for the transcription factor p53 than the approach of Cawley *et al.* (2004) and included their 14 experimentally verified blips. Even though only four of our blips have an exact match to the p53 consensus binding sequence, more than half have matches to the consensus sequence with at most two mismatches. In addition, the identified blips are preferentially located in potential regulatory regions, i.e., 5' exons, CpG islands, and within introns.

A number of issues remain to be addressed. The first one is the comparison of the two control experiments, whole cell extract and ControlGST. Here, we have only used the negative control, ControlGST, however it is of interest to see how these two controls compare and if there are ways to combine information from both. Secondly, it would be worthwhile to investigate whether there is any systematic organization of the pentamer `RRRCW` and its reverse complement, other than the consensus identified in the literature, that might be causing the binding activity in the identified sequences. This might lead to a slightly different representation of the p53 consensus sequence. A third issue is investigating various types of scan statistics. One alternative could be a weighted version of the scan statistic, with weights inversely related to the genomic distance, e.g., using a Gaussian kernel.

Acknowledgements

The authors are grateful to Tom Gingeras for providing early access to the ChIP-Chip data, thank Stefan Bekiranov for his comments on an earlier version of this manuscript, and acknowledge Siew Leng Teng for MDScan and Bioprospector runs on the identified blips as well as critical reading of the manuscript.

Appendix

Procedure for filtering the local repeats of Section 2.3

We use the following procedure to systematically filter out locally repeated probe-pairs. The general scheme of this filtering procedure is to discard probe-pairs for which exact replicate measurements are found within ~ 30 probes (within ~ 1kb). The procedure is as follows.

- For each 25mer that has more than one copy in the dataset, extract the data file indices and genomic locations of all occurrences. Compute the data file index distances (DDI), genomic location

distances (DGL) between each consecutive occurrence, and the number of unique measurements among these occurrences (UM). The number of unique measurements is based on the actual PM and MM measurements.

- The probes can be filtered out for two reasons:
  - **Type I filtering: Delete probes with UM=1 and DDI> -30.** If any of the DDI values are greater than -30 and UM is 1, then delete all the occurrences of this probe.
  - **Type II filtering: Delete probes with UM> 1, DGL> 0, and DDI> -30.** If UM is greater than 1, this means that the same 25mer with different measurements is repeated in the vicinity of the same genomic location. In this case, we want the repeated pairs to be far away from each other. Hence, we delete the probes, among the ones with a DGL greater than 0, that have DDI greater than -30.

Example of type I filtering.

```
25mer: "AAGGCCCTGTACCAACACAGATACA"

Data file index of the occurrences: 955271 955432 955454 955761
955851

Genomic locations of the occurrences: 41125518 41126057 41126131
41127138 41127459

DDI: -161  -22 -307  -90

DGL:  -539   -74 -1007  -321

PM values of the 5 occurrences: 1078 1078 1078 1078 1078

MM values of the 5 occurrence: 581 581 581 581 581

UM: 1
```

Example of type II filtering.

```
Example 1:

25mer : "GCCTCCAACACAGGAGGCTTCAGTA"

Data file index of the occurrences:  630    632 708351 708352
881545

Genomic locations of the occurrences: 7715959  7715999 14061570
14061610 35374379

DDI:   -2   -707719  -1 -173193
```

```
DGL:  -40  -6345571 -40 -21312769

PM values of the 5 occurrences:  709 709 550 550 550

MM values of the 5 occurrence: 664 664 712 712 712

UM: 2 (Number of unique measurements among the 5 occurrences).

Example 2:

25mer : "AAATATTGATTAACAGTGATTTATT"

Data file index of the occurrences:  8210   8211 244518 244519
244528 244529


Genomic locations of the occurrences: 11332183 11332183 25957216
25957216 25957341 25957341

DDI: -1 -236307      -1      -9      -1

DGL: 0 -14625033         0      -125         0

PM values of the 5 occurrences: 337 364 337 364 337 364

MM values of the 5 occurrence: 386 452 386 452 386 452

UM: 2
```

Nested-Bonferroni adjustment when $T^*_{i,n}, i \in \{1, \cdots, N-w+1\}$ are independent

We now investigate the case when $T^*_{i,n}$, $i \in \{1, \cdots, N-w+1\}$, are independent. Obviously, this condition never holds for our choice of scan statistics $T^*_{i,n}$. Under independence, we have

$$\mathcal{G}_0(.) = \mathcal{F}_0^w(.) = (1 - \bar{\mathcal{F}}_0(.))^w,$$

where $\bar{\mathcal{F}}_0(.) = 1 - \mathcal{F}_0(.)$ denotes the marginal survivor function of the $T^*_{i,n}$. Accordingly, (6) becomes

$$Pr\left(\max_{k \in \mathcal{K}} Z_{k,n} > c\right) \leq \left\lceil \frac{N-w+1)}{w} \right\rceil (1 - \mathcal{G}_0(c)) = \left\lceil \frac{N-w+1)}{w} \right\rceil (1 - \mathcal{F}_0^w(c)).$$

Moreover, if w is small, i.e., $w << N$, then by the binomial expansion we have

$$1 - \mathcal{G}_0(c) \approx 1 - w\bar{\mathcal{F}}_0(c) \Longrightarrow \mathcal{G}_0(c) \approx w(1 - \mathcal{F}_0(c)). \tag{14}$$

We now consider the following two functions for identifying the cut-offs of the nested-Bonferroni and Bonferroni procedures at various nominal α levels:

$$FWER_{NB}(c \mid N, w, \mathcal{F}_0) \quad = \quad \left\lceil \frac{N-w+1}{w} \right\rceil (1 - \mathcal{F}_0^w(c)), \tag{15}$$

$$FWER_B(c \mid N, w, \mathcal{F}_0) \quad = \quad (N - w + 1)(1 - \mathcal{F}_0(c)). \tag{16}$$

Using (14), we have

$$FWER_{NB}(. \mid N, w, \mathcal{F}_0) \approx \left\lceil \frac{N - w + 1}{w} \right\rceil w(1 - \mathcal{F}_0(.)) \approx FWER_B(. \mid N, w, \mathcal{F}_0). \tag{17}$$

Figures 19 and 20 display plots of the functions $FWER_{NB}(c \mid N, w, \mathcal{F}_0)$ and $FWER_B(c \mid N, w, \mathcal{F}_0)$ as a function of c for various N values with $\mathcal{F}_0 \sim N(0, 1)$, for $w = 10$ and $w = 5$, respectively. We observe that, as shown in (17), the two functions give the same cut-off for different values of α ($\alpha = 0.05$ is marked with the $y = 0.05$ line).

[Figure 19 about here.]

[Figure 20 about here.]

References

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57, 289–300.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccilboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T. R. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116, 499–509.

Hoh, J., Jin, S., Parrado, T., Edington, J., Levine, A. J., and Ott, J. 2002. The p53MH algorithm and its application in detecting p53-responsive genes. *Proceedings of National Academy of Sciences* 99, 8467–8472.

Inga, A., Francesca, S., Darden, T. A., and Resnick, M. A. 2002. Differential tranactivation by the p53 transcription factor is highly dependent on p53 level and promoter target sequence. *Molecular and Cellular Biology* 22, 8612–8625.

Ioshikhes, I. P. and Zhang, M. Q. 2000. Large-scale human promoter mapping using CpG islands. *Nature Genetics* 26, 61–63.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., , and Haussler, D. 2002. The Human Genome Browser at UCSC. *Genome Research* 12, 996–1006.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. R., Thompso n, C. M., I., S., J., Z., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J., L., V. T., Fr aenkel, E., K., G. D., and Young, R. A. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae* . *Science* 298, 799–804.

Liu, X., Brutlag, D. L., and Liu, J. S. 2001. BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Proceedings of the Pacific Symposium of Biocomputing*, pages 127–138.

Liu, X. S., Brutlag, D. L., and Liu, J. S. 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin immunoprecipitation microarray experiments. *Nature Biotechnology* 20, 853–839.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. 2004. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology.* url: `www.bepress.com/ucbbiostat/paper141`. (To appear).

Wang, Y., Schwedes, J. F., Parks, D., Mann, K., and Tegtmeyer, P. 1995. Interaction of p53 with its consensus DNA-binding site. *Molecular and Cellular Biology* 15, 2157–2165.

Yin, Y., Liu, Y., Jin, Y. J., Hall, E. J., and Barrett, J. C. 2003. PAC1 phosphatase is a transcription target of p53 in signalling apoptosis and growth suppression. *Nature* 422, 527–531.

List of Figures

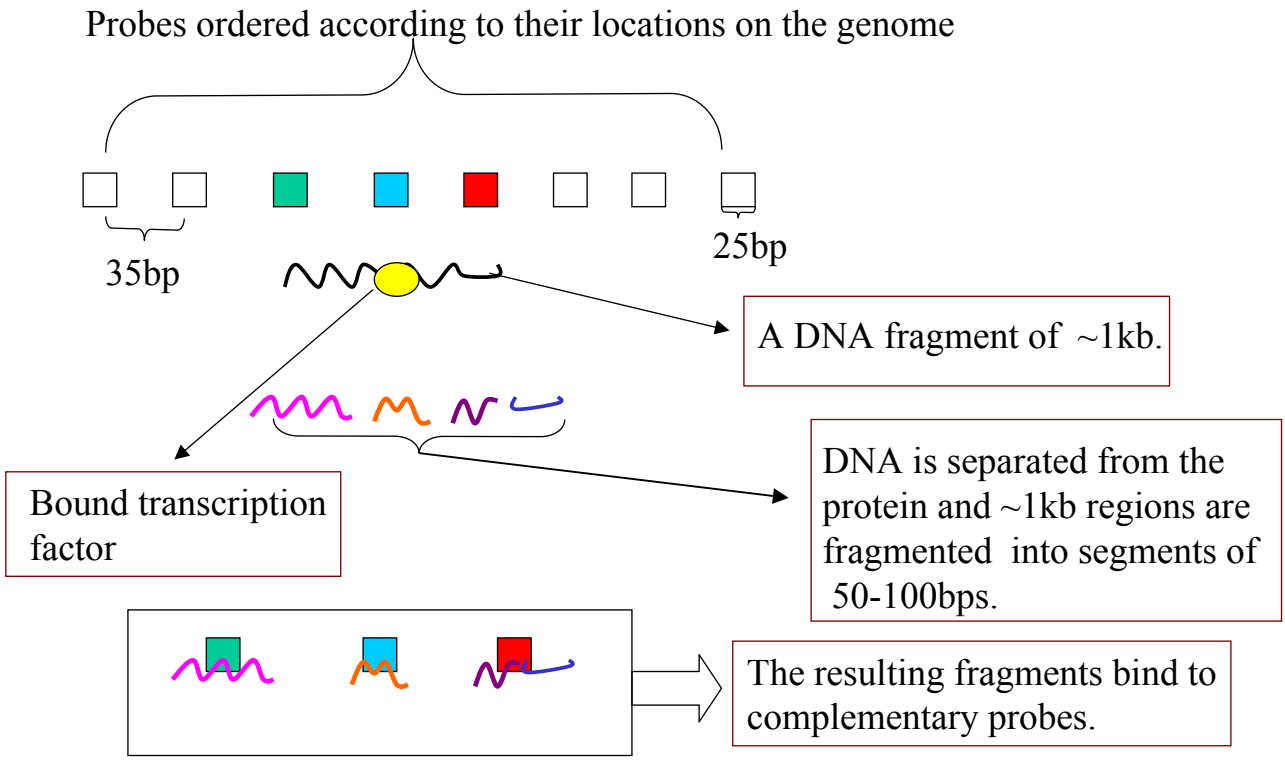

Figure 1: *Details of IP-enriched DNA hybridization at the probe-level.*

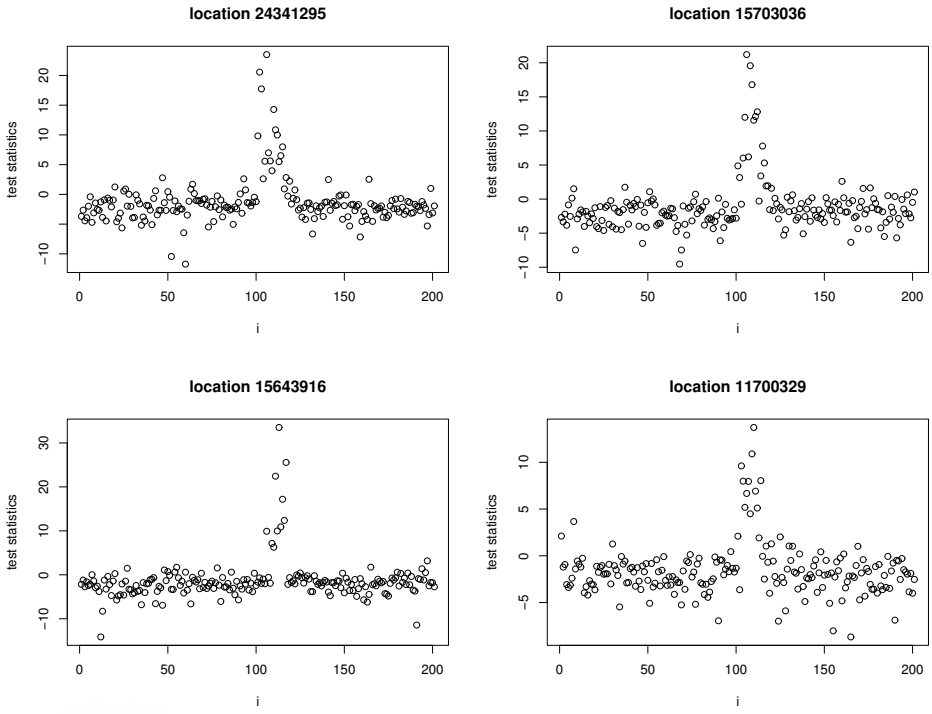

Figure 2: *p53 ChIP-Chip data.* Plot of 200 two-sample Welch t-statistics around four different locations on chromosome 21.

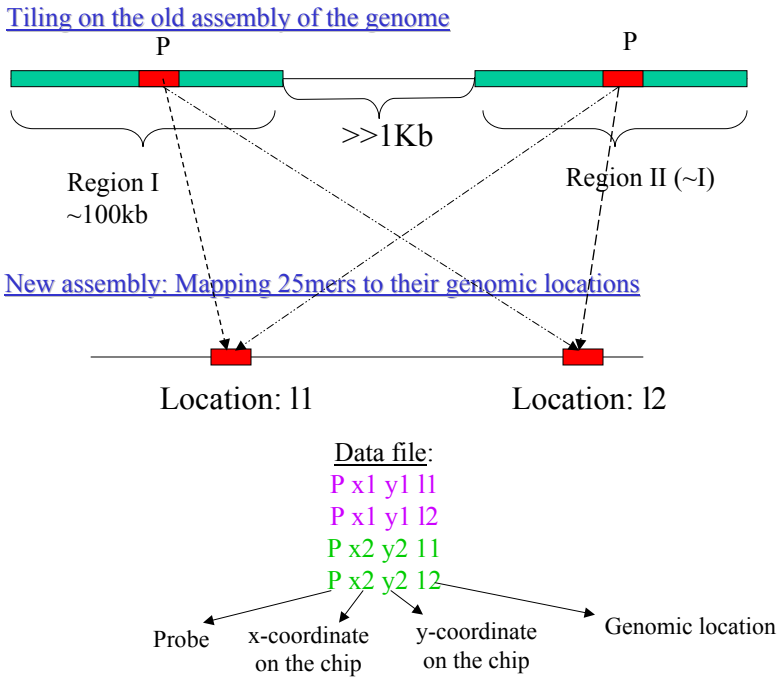

Figure 3: *Distant repeats.* Regions I and II are two distant repeat regions that RepeatMasker failed to mask at the time when the corresponding version of the genome was tiled. Two copies of the same 25mer P is used on the chip and two independent observations are measured. When the probes are mapped to the new version (June 2002 freeze) of the genome assembly, each 25mer will match to both positions l_1 and l_2. Hence ,the data corresponding to each will be duplicated, creating four observations instead of two.

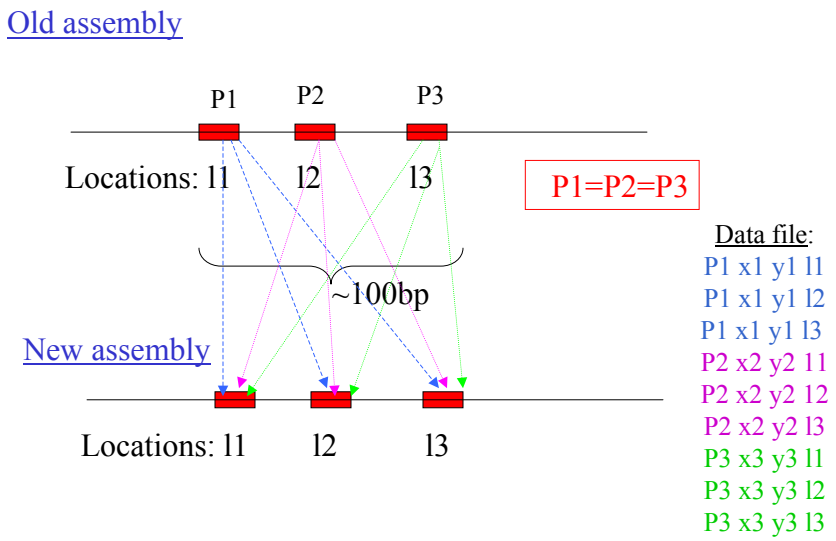

Figure 4: *Local repeats.* Local repeats are generated as follows. When there are short repeat regions with non-unique 25mers in the genome, the tiling process might pick exactly the same 25mer from different parts of this region, e.g., P_1, P_2, and P_3 are the same. Subsequently, three independent observations are measured for these probes. However, when these 25mers are mapped to the updated assembly of the genome (June 2002 freeze), each 25mer will map to all the three locations. Hence, the three probes will be represented by a total of nine data points, only three of which are unique. In the data file, observations mapping to the same genomic location occur in a consecutive manner.

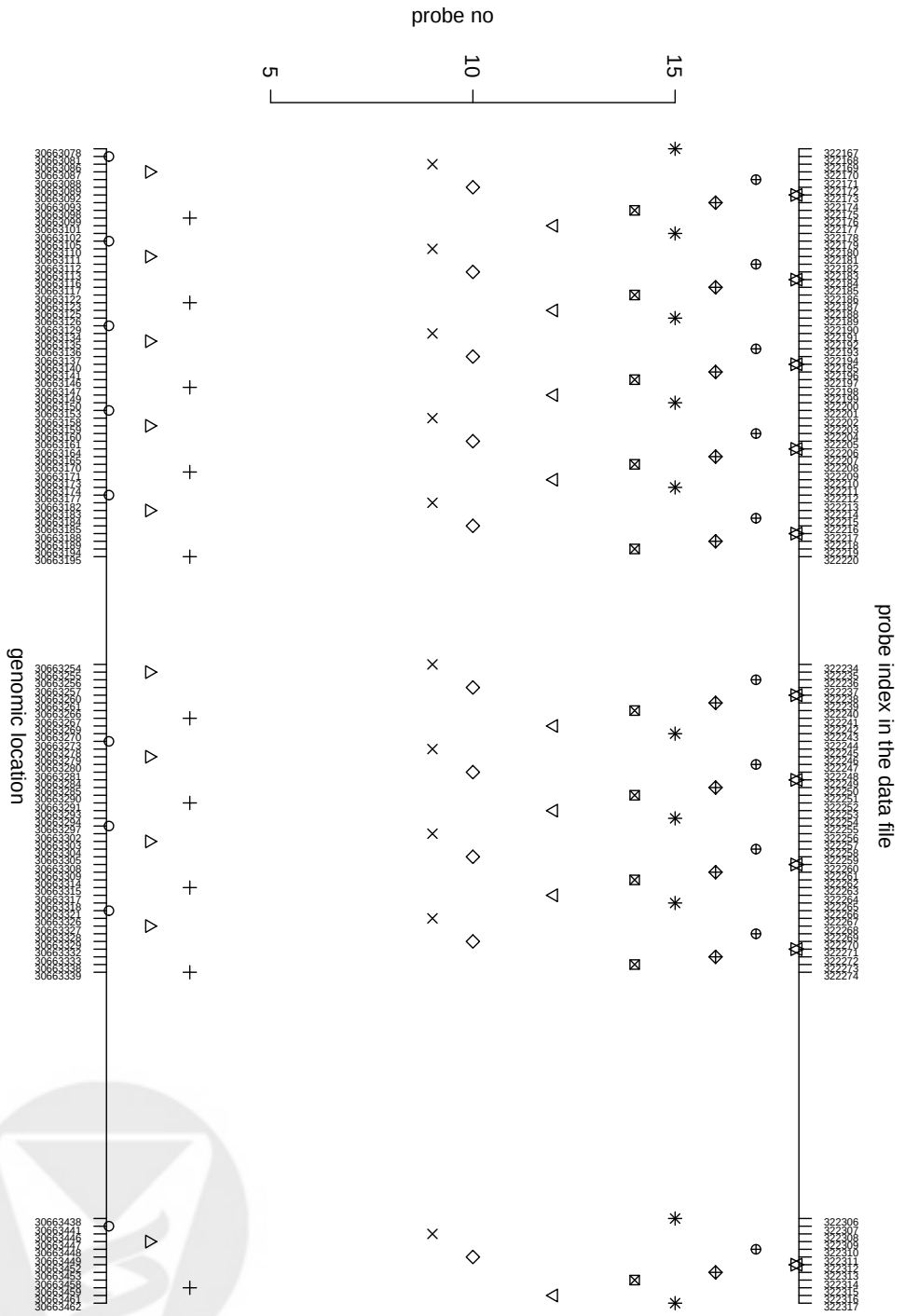

Figure 5: *Display of representative repeat regions on chromosome 21 from chip A*. The x-axis represents 11 unique 25mers (each represented by a unique plotting symbol) and the y-axis on the left represents the genomic locations that these 25mers match to, whereas the y-axis on the right represents the locations of the corresponding occurrences of the 25mers in the data file, i.e., their row numbers. We see that rows 322234 to 322274 of the data file have a total of 41 measurements; however these correspond to only eleven unique measurements on eleven 25mers.

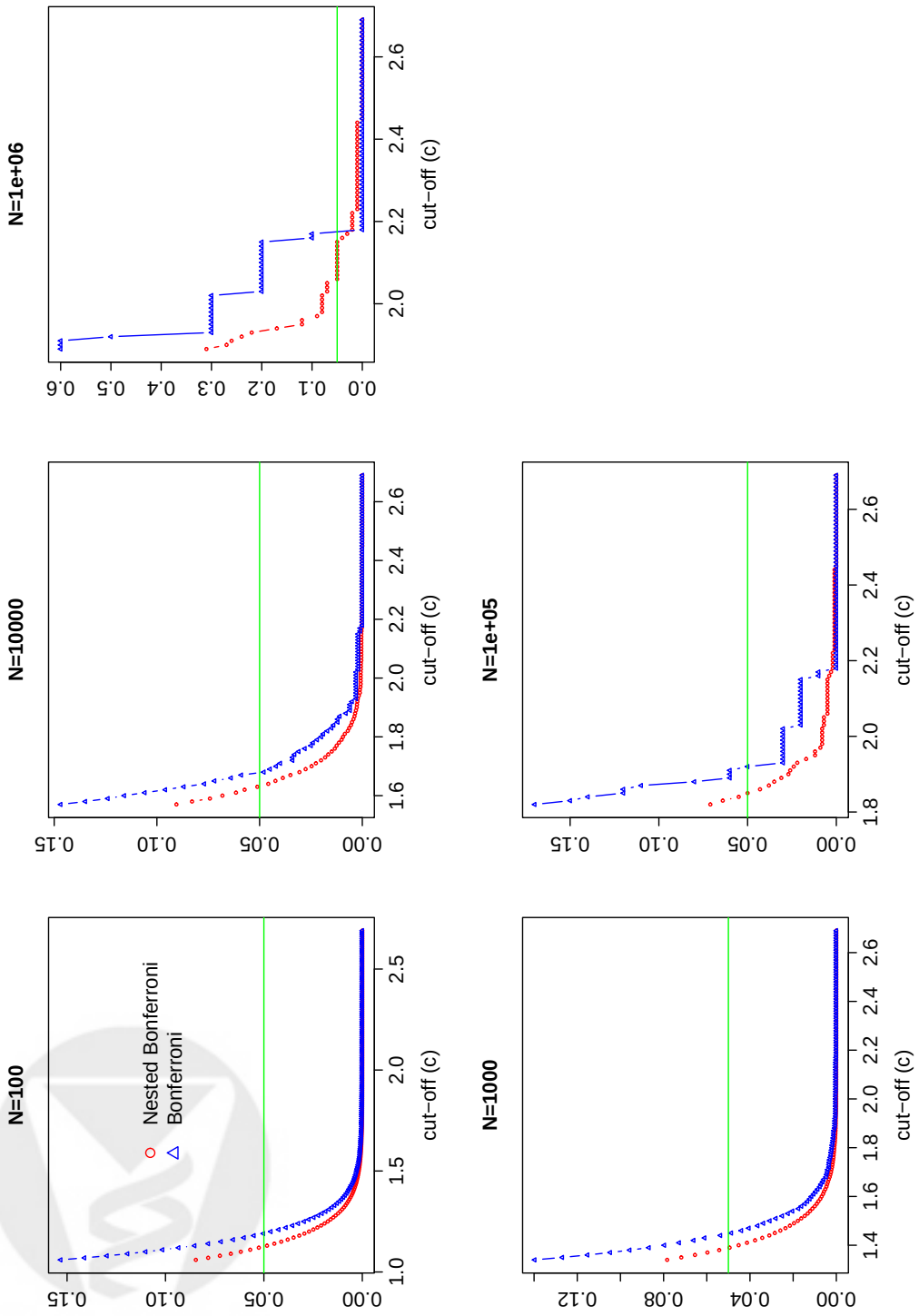

Figure 6: *Comparison of the Bonferroni and nested-Bonferroni adjustments.* Plot of nominal Type I error rate versus cut-off for the test statistic Z. $T_{i,n}$, $i \in \{1, \cdots, N\}$, are two-sample Welch t-statistics based on $n_1 = 6$ and $n_2 = 6$ independent observations from a normal distribution with mean 0 and standard deviation 1; $T^*_{i,n}$, $i \in \{1, \cdots, N - w + 1\}$, are dependent scan statistics and the assumed blip size is set to $w = 10$. The distribution functions $\mathcal{F}_0$ and $\mathcal{G}_0$ of $T^*_{i,n}$ and $Z_{k,n}$ are estimated by their empirical distributions based on $B = 10,000,000$ simulated observations using $w = 10$.

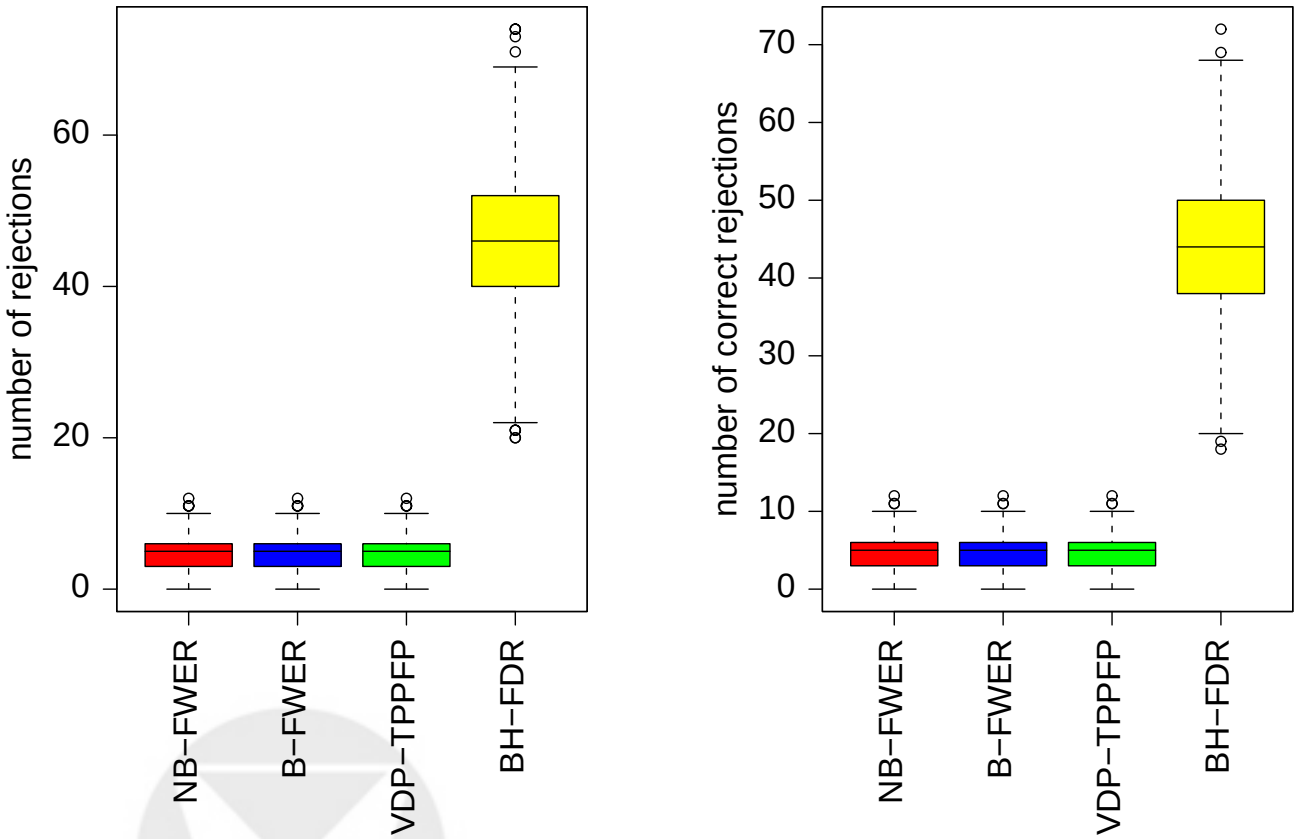

Figure 7: *Simulation 0. Boxplots of the numbers of probe-level rejections and correct rejections with an assumed blip size of* $w = 1$ *for the procedures NB-FWER, B-FWER, VDP-TPPFP, and BH-FDR.* The nominal Type I error rate is $\alpha = 0.05$ or $q = 0.05$ for each procedure and there are a total of ~ 2000 tests.

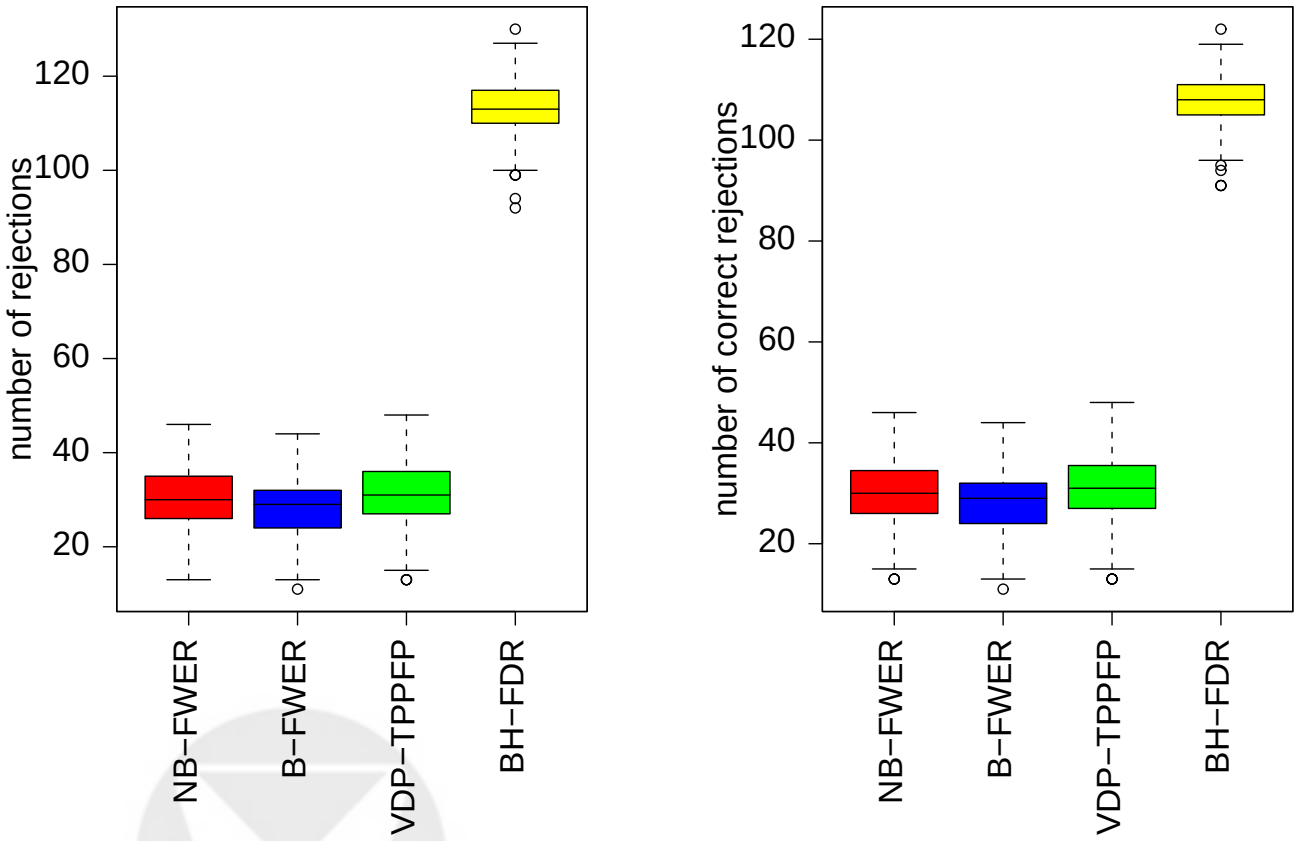

Figure 8: *Simulation 0. Boxplots of the numbers of probe-level rejections and correct rejections with an assumed blip size of* $w = 2$ *for the procedures NB-FWER, B-FWER, VDP-TPPFP, and BH-FDR.* The nominal Type I error rate is $\alpha = 0.05$ or $q = 0.05$ for each procedure and there are a total of ~ 2000 tests.

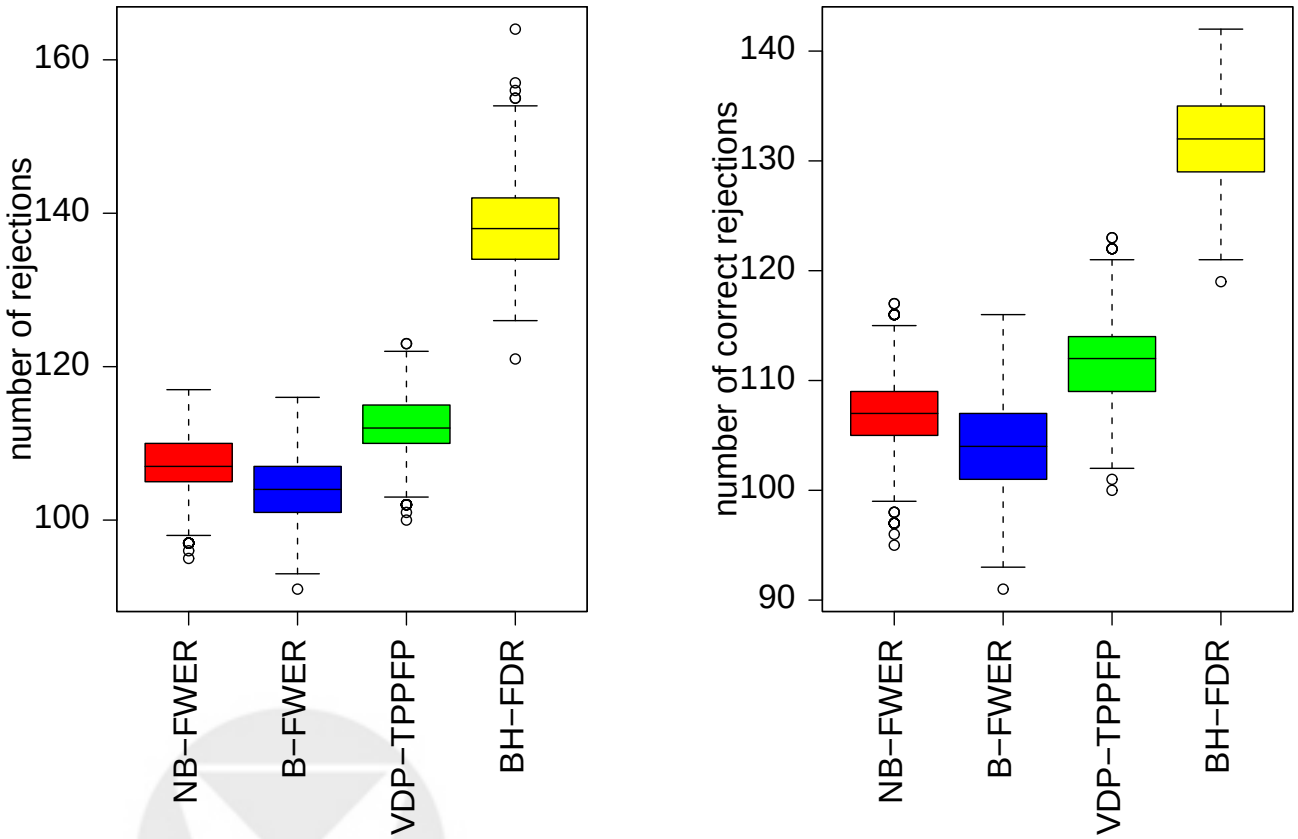

Figure 9: *Simulation 0. Boxplots of the numbers of probe-level rejections and correct rejections with an assumed blip size of* $w = 5$ *for the procedures NB-FWER, B-FWER, VDP-TPPFP, and BH-FDR.* The nominal Type I error rate is $\alpha = 0.05$ or $q = 0.05$ for each procedure and there are a total of ~ 2000 tests.

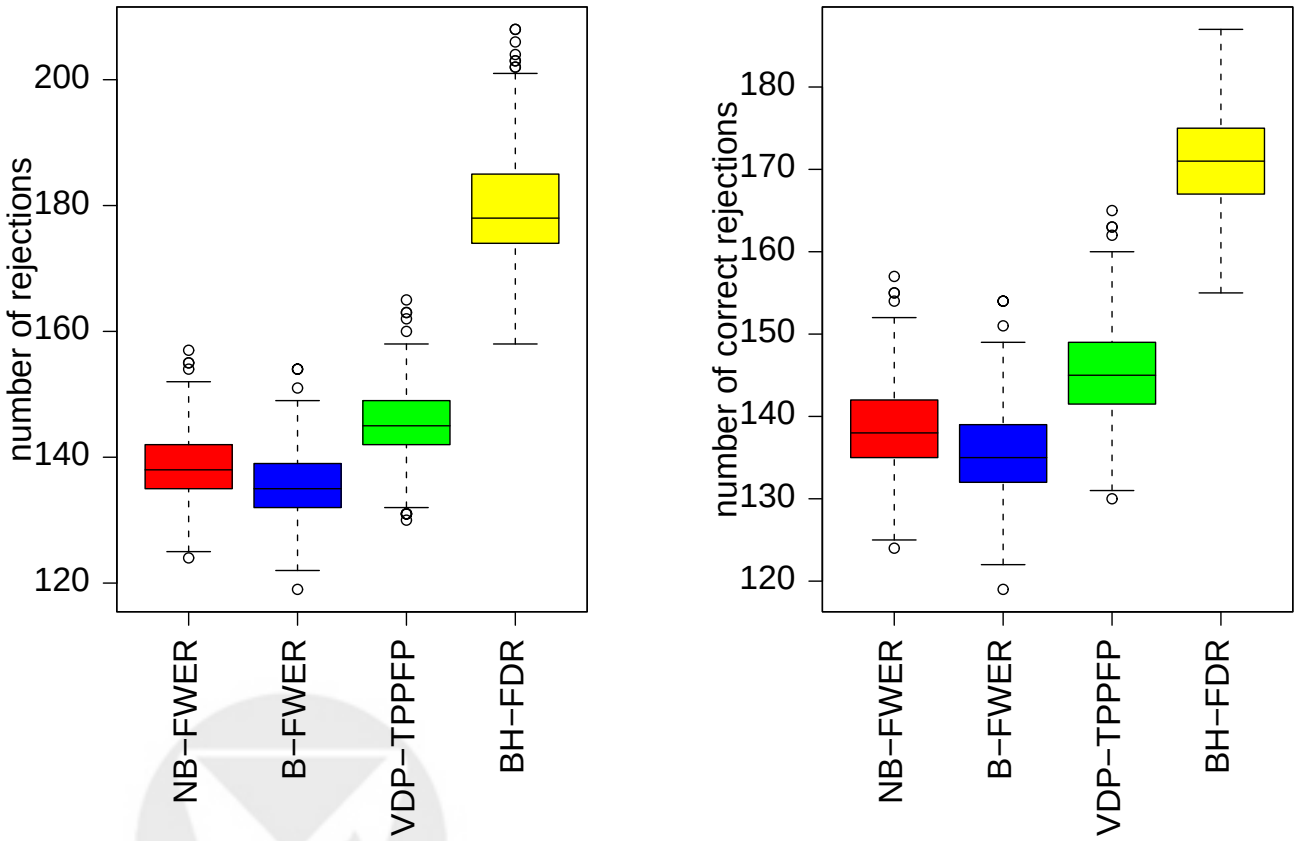

Figure 10: *Simulation 0. Boxplots of the numbers of probe-level rejections and correct rejections with an assumed blip size of* $w = 10$ *for the procedures NB-FWER, B-FWER, VDP-TPPFP, and BH-FDR.* The nominal Type I error rate is $\alpha = 0.05$ for each procedure and there are a total of ~ 2000 tests.

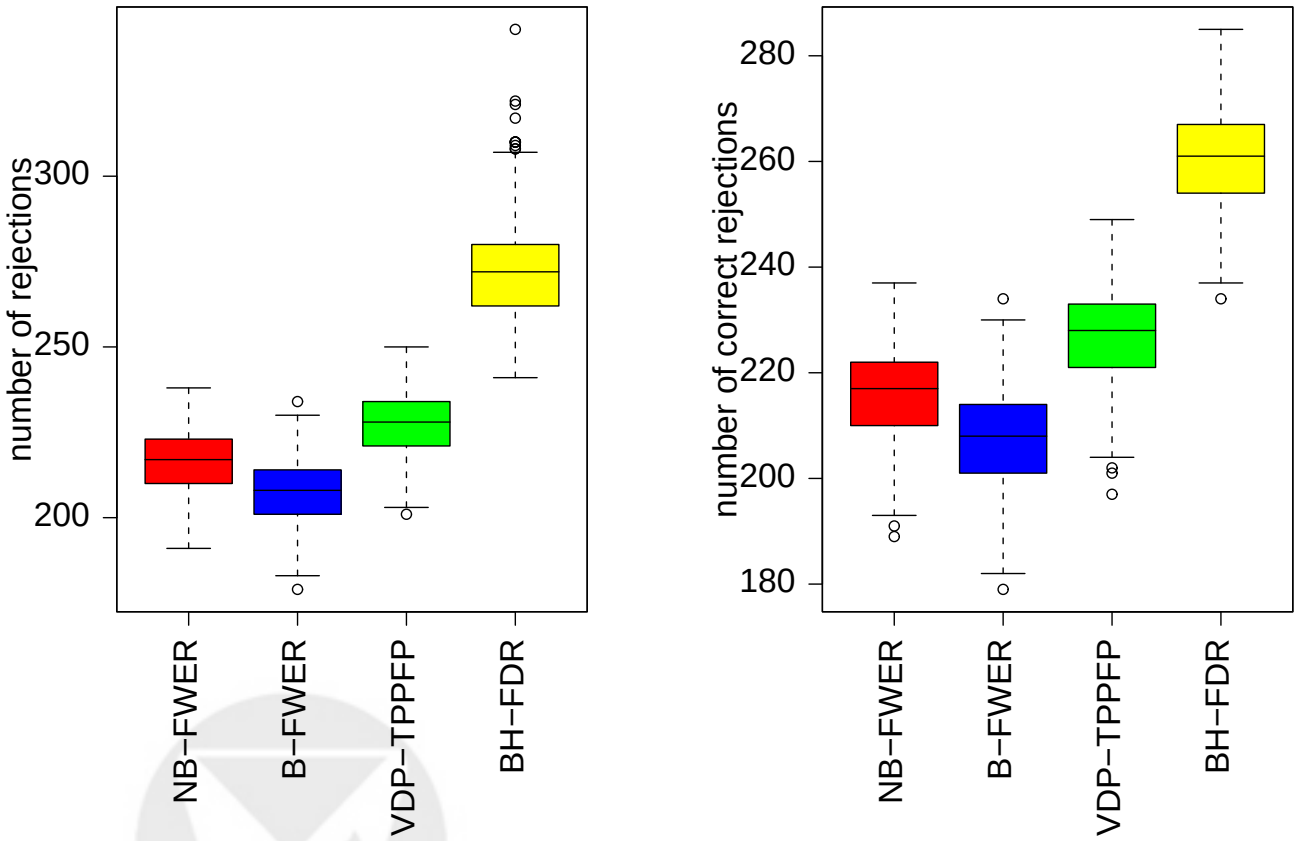

Figure 11: *Simulation 0. Boxplots of the number of rejections and number of correct rejections with an assumed blip size of* $w = 20$ *for the procedures NB-FWER, B-FWER, VDP-TPPFP, and BH-FDR.* The nominal Type I error rate is $\alpha = 0.05$ for each procedure and there are a total of ~ 2000 tests.

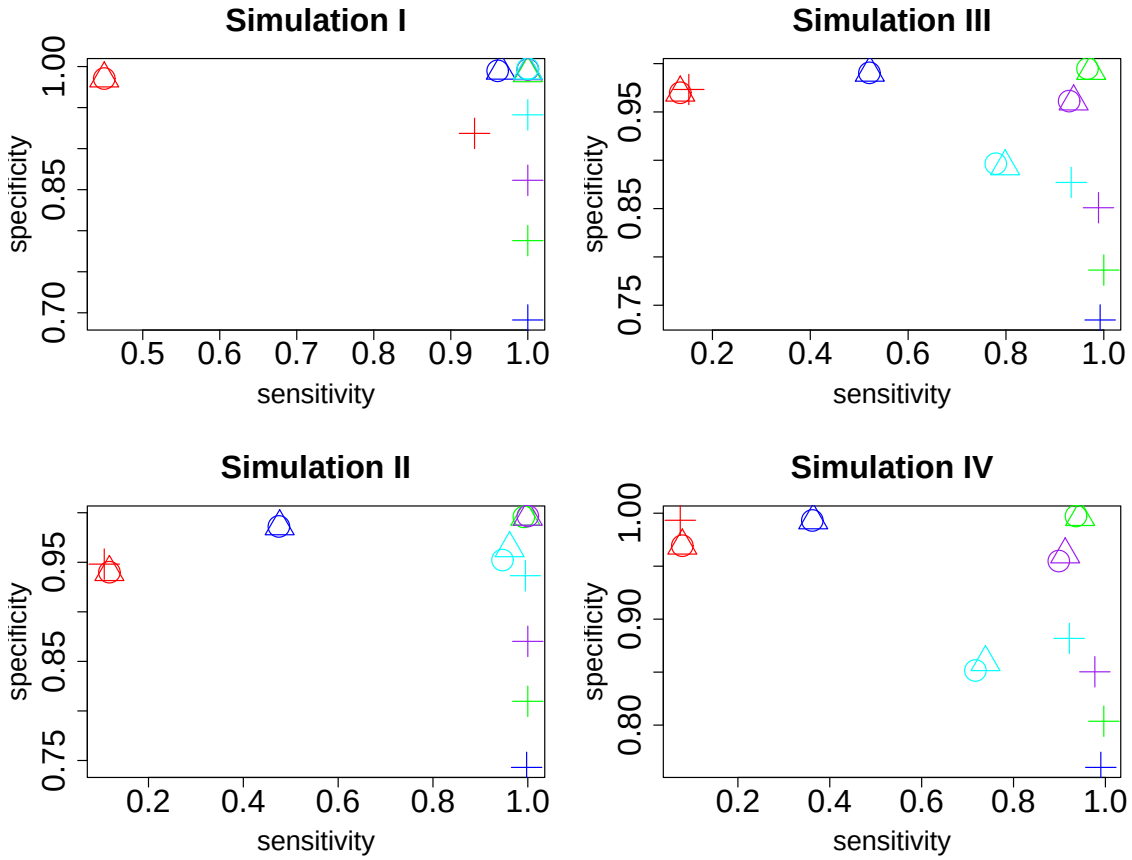

Figure 12: *Simulations I, II, III, and IV. Specificity versus sensitivity plots.* Specificity and sensitivity are computed at the blip-level using averages over 100 independently simulated datasets and are defined as the ratio of *correctly identified* blips to the total number of *identified* blips and total number of *true* blips, respectively. Plotting symbols are ◯: NB-FWER, △:VDP-TPPFP, and +: BH-FDR. Different colors represent different assumed blip sizes: $w = 1$ in `red`, $w = 2$ in `blue`, $w = 5$ in `green`, $w = 10$ in `purple`, and $w = 20$ in `cyan`. The true blip size equals $w^* = 10$ probes in Simulations I and II, and for Simulations III and IV the true blip size is variable with a mean of 10.5 probes.

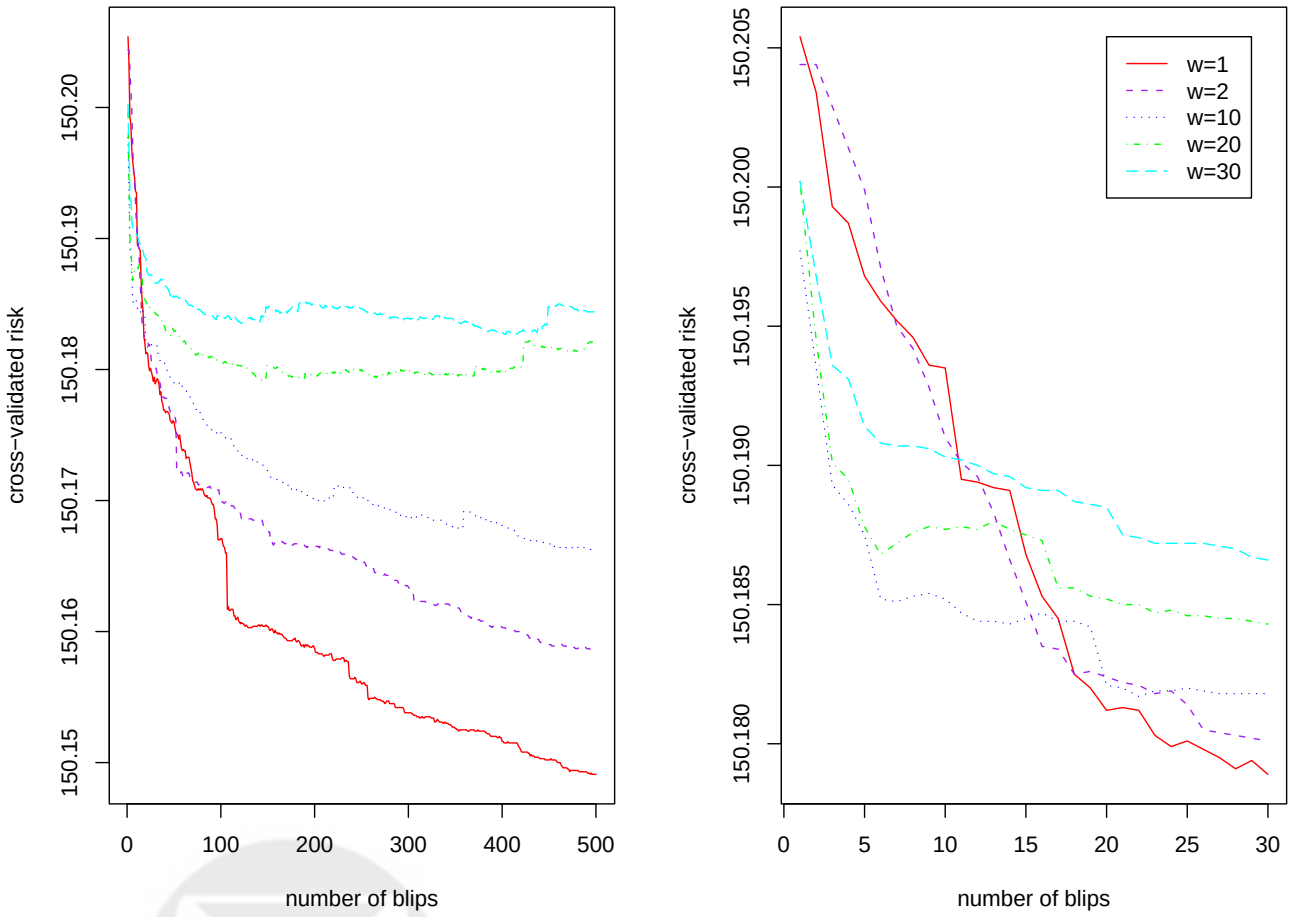

Figure 13: *p53 ChIP-Chip data. Cross-validated risk as a function of assumed blip size w and number of blips.* Monte-Carlo cross-validation is performed using 6 control and 6 treatment replicates for $\sim 300,000$ probe-pairs. A total of 9 different cross-validation steps were performed, where each validation set included one technical hybridization replicate for each of the IP replicates. The panel on the right zooms into the first 30 blips for each assumed blip size.

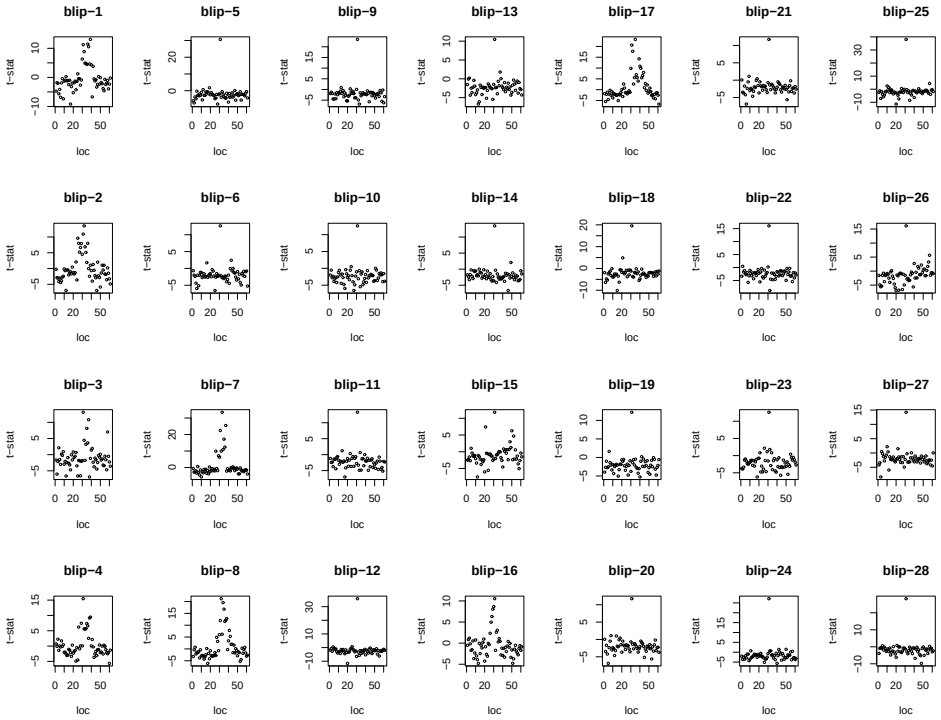

Figure 14: *p53 ChIP-Chip data. The 28 blips identified on chip A (chromosome 21) using the NB-FWER multiple testing procedure with an assumed blip size of* $w = 1$. Blips displayed here are identified by controlling the FWER at nominal level $\alpha = 0.05$ using the NB-FWER procedure. Control of the FDR using the BH-FDR procedure at nominal level $q = 0.05$ identified 48 blips and only 10 of these blips resembled *real blips*.

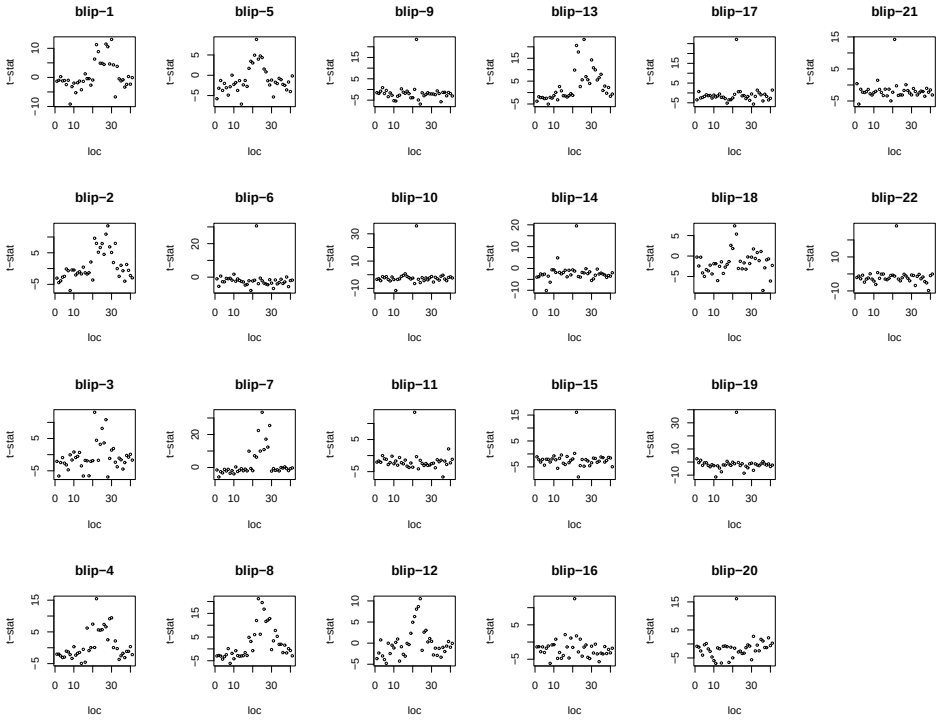

Figure 15: *p53 ChIP-Chip data. The 22 blips identified on chip A (chromosome 21) using the NB-FWER multiple testing procedure with an assumed blip size of* $w = 2$. Blips displayed here are identified by controlling the FWER at nominal level $\alpha = 0.05$ using the NB-FWER procedure. Control of the FDR using the BH-FDR procedure at nominal level $q = 0.05$ identified 41 blips and only 9 of these blips resembled *real blips*.

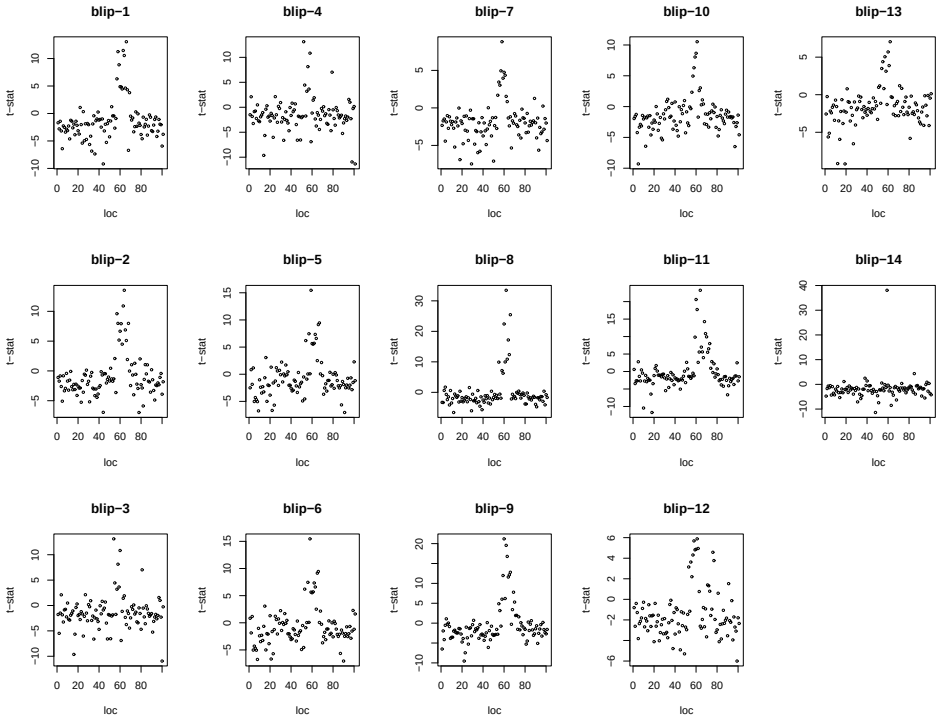

Figure 16: *p53 ChIP-Chip data. The 14 blips identified on chip A (chromosome 21) using the NB-FWER multiple testing procedures with an assumed blip size of* $w = 10$. Blips displayed here are identified by controlling the FWER at nominal level $\alpha = 0.05$ using the NB-FWER procedure. Control of the FDR using the BH-FDR procedure at nominal level $q = 0.05$ identified 23 blips and only 13 of these blips resembled *real blips*.

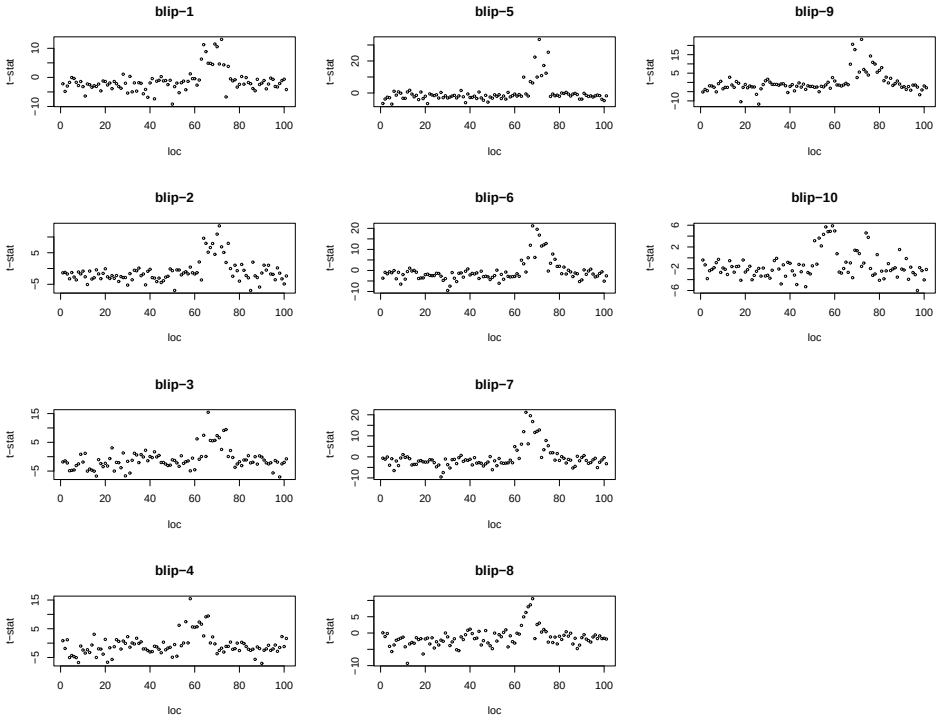

Figure 17: *p53 ChIP-Chip data. The 10 blips identified on chip A (chromosome) using the NB-FWER multiple testing procedures with an assumed blip size of* $w = 20$. Blips displayed here are identified by controlling the FWER at nominal level $\alpha = 0.05$ using the NB-FWER procedure. Control of the FDR using the BH-FDR procedure at nominal level $q = 0.05$ identified 11 blips and all of these blips resembled *real blips*.

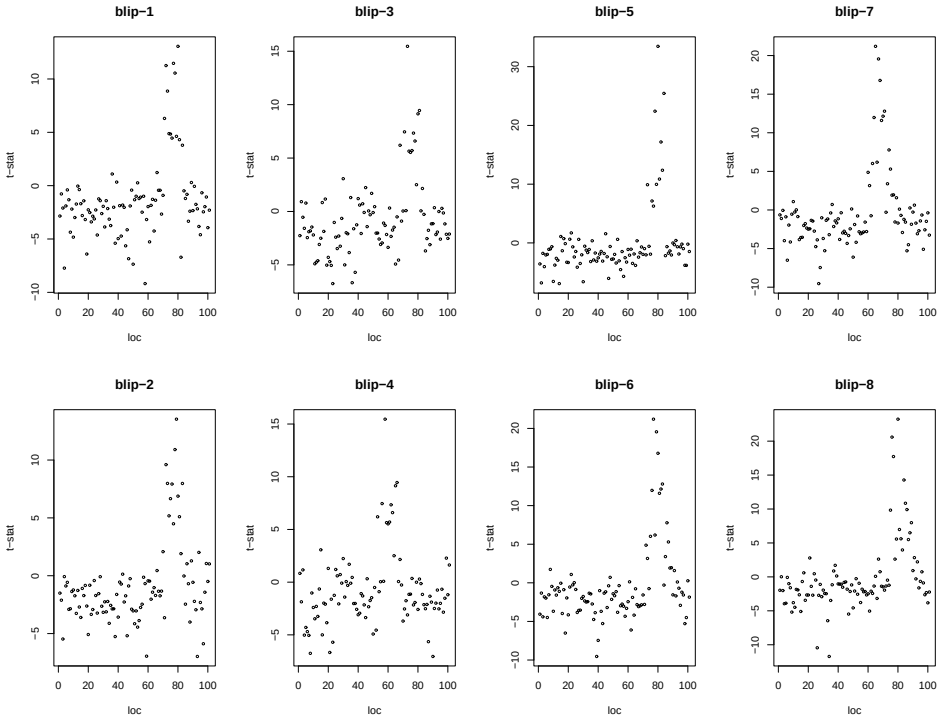

Figure 18: *p53 ChIP-Chip data. The 8 blips identified on chip A (chromosome) using the NB-FWER multiple testing procedures with an assumed blip size of* $w = 30$. Blips displayed here are identified by controlling the FWER at nominal level $\alpha = 0.05$ using the NB-FWER procedure. Control of the FDR using the BH-FDR procedure at nominal level $q = 0.05$ identified 9 blips and all of these blips resembled *real blips*.

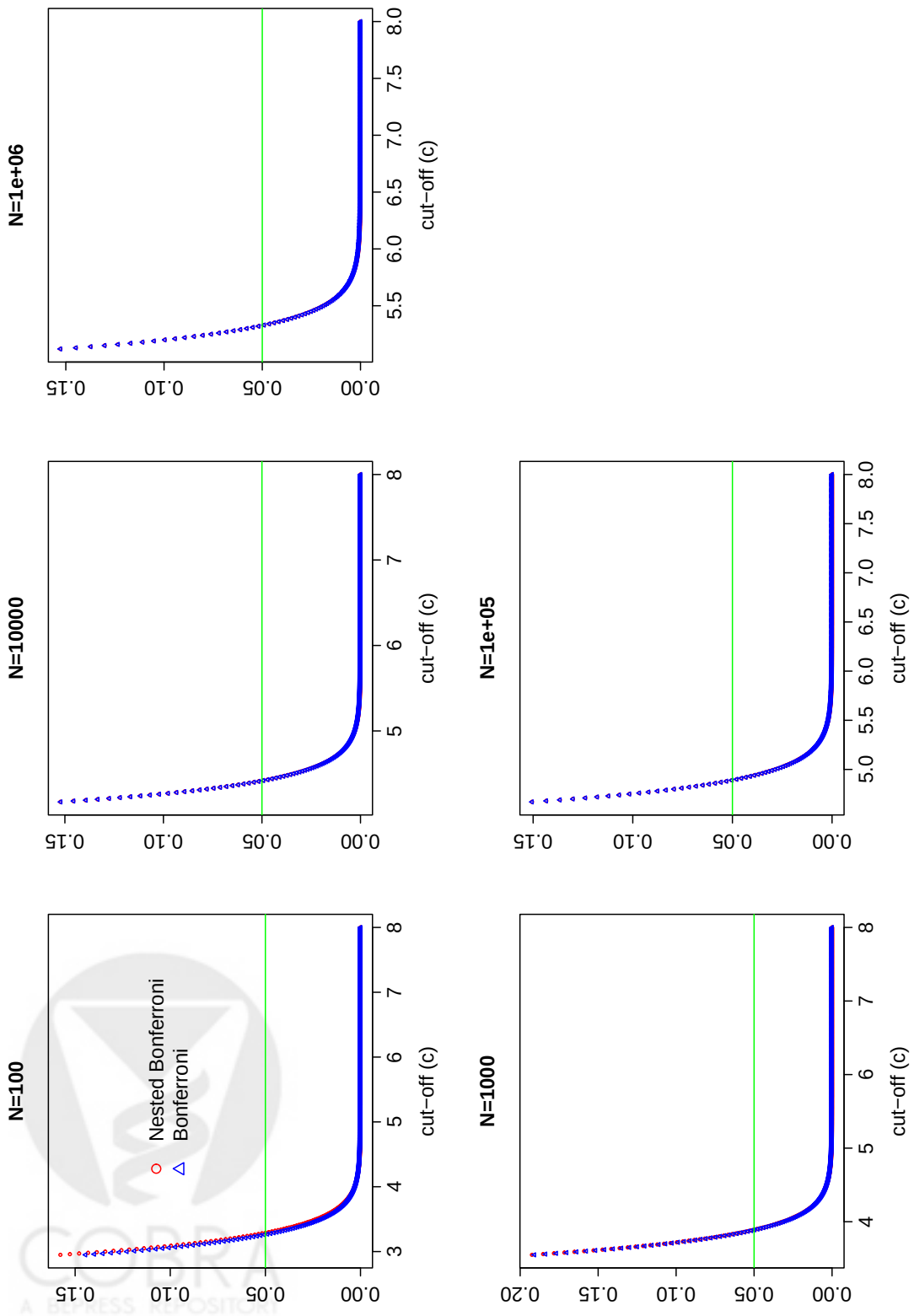

Figure 19: *Comparison of the Bonferroni and nested-Bonferroni adjustments.* $T^*_{i,n}$, $i \in \{1, \cdots, N-w+1\}$, are independently generated from $\mathcal{N}(0,1)$ and the blip size is set to $w = 10$. The null distribution $\mathcal{F}_0$ of $T^*_{i,n}$ is set to $\mathcal{N}(0,1)$.

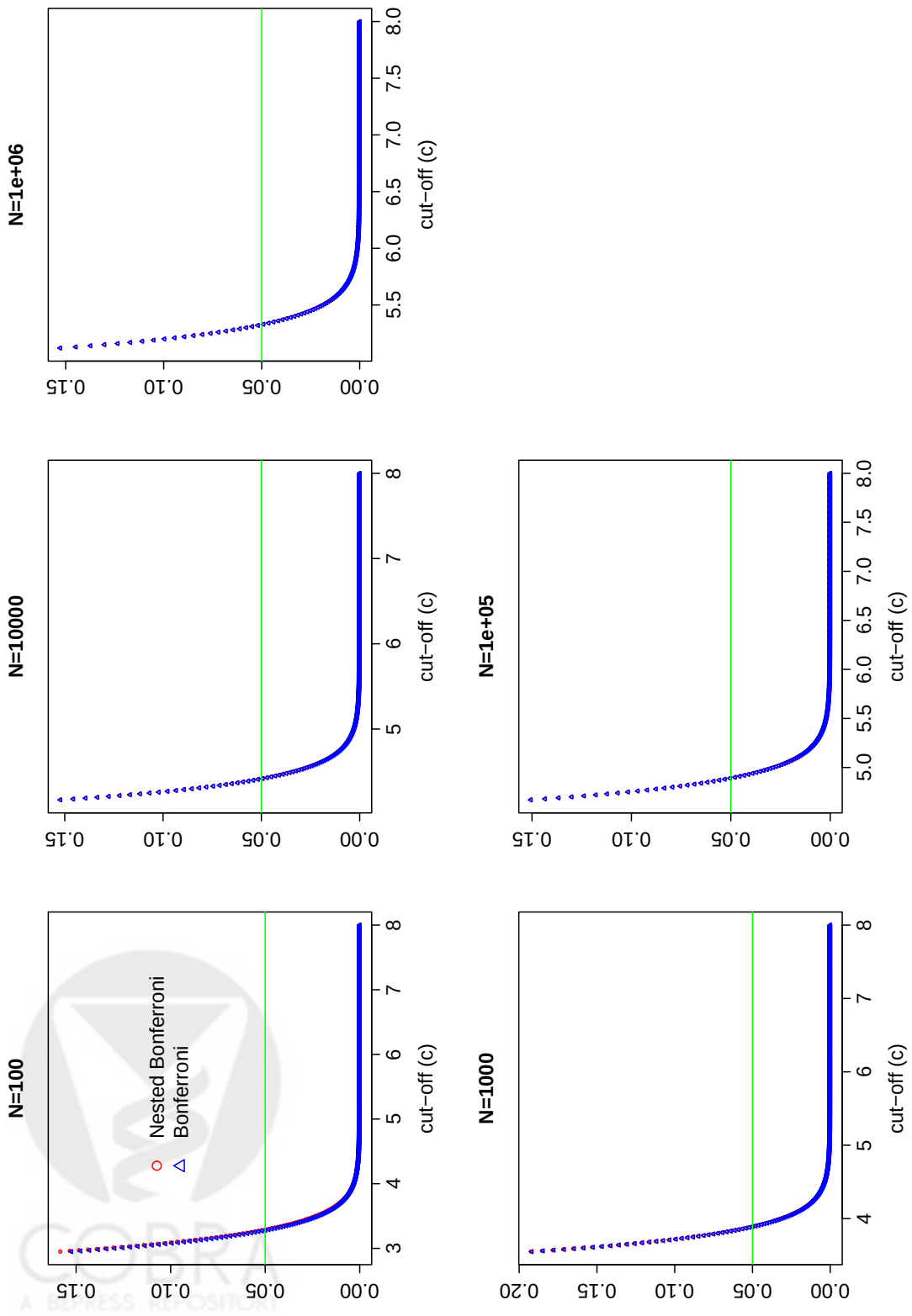

Figure 20: *Comparison of the Bonferroni and nested-Bonferroni adjustments.* $T^*_{i,n}$, $i \in \{1, \cdots, N-w+1\}$, are independently generated from $\mathcal{N}(0,1)$ and the blip size is set to $w = 5$. The null distribution $\mathcal{F}_0$ of $T^*_{i,n}$ is set to $\mathcal{N}(0,1)$.

List of Tables

25mer			genomic location					PM			MM		
GCACACGGTGTTGTGTCAGCATCGC	f	chr21	43673223	901	411	901	412	3757	815	25	2305	529	25
CACACGGTGTTGTGTCAGCATCGCC	f	chr21	43673224	332	231	332	232	4139	879	25	2258	392	25
GCACACGGTGTTGTGTCAGCATCGC	f	chr21	43673270	901	411	901	412	3757	815	25	2305	529	25
CACACGGTGTTGTGTCAGCATCGCC	f	chr21	43673271	332	231	332	232	4139	879	25	2258	392	25
GCACACGGTGTTGTGTCAGCATCGC	f	chr21	43673317	901	411	901	412	3757	815	25	2305	529	25
CACACGGTGTTGTGTCAGCATCGCC	f	chr21	43673318	332	231	332	232	4139	879	25	2258	392	25
GCACACGGTGTTGTGTCAGCATCGC	f	chr21	43673364	901	411	901	412	3757	815	25	2305	529	25
CACACGGTGTTGTGTCAGCATCGCC	f	chr21	43673365	332	231	332	232	4139	879	25	2258	392	25
GCACACGGTGTTGTGTCAGCATCGC	f	chr21	43673411	901	411	901	412	3757	815	25	2305	529	25

Table 1: *p53 ChIP-Chip data. Example of a local repeat region.* 10 consecutive probe-pairs on chip B, i.e., data points correspond to rows 194671 to 194680 in the data file. We note that these ten data points contain only two unique measurements.

25mer			genomic location					PM			MM		
AAAAAGTCCTGCAAATGTTCCTCAT	f	chr21	11348557	375	189	375	190	558	91	25	506	84	25
AAAAAGTCCTGCAAATGTTCCTCAT	f	chr21	25973833	375	189	375	190	558	91	25	506	84	25
AAAAAGTCCTGCAAATGTTCCTCAT	f	chr21	11348557	376	189	376	190	573	81	25	462	73	25
AAAAAGTCCTGCAAATGTTCCTCAT	f	chr21	25973833	376	189	376	190	573	81	25	462	73	25

Table 2: *p53 ChIP-Chip data. Example of a distant repeat region.* Probe-pairs in rows 8733, 245036, 8734, and 245037 of the data file corresponding to chip A.

	$N = 100$	$N = 1,000$	$N = 10,000$	$N = 100,000$	$N = 1,000,000$
c_B	1.1908	1.4466	1.6771	1.9103	2.1688
c_{NB}	1.1262	1.3888	1.6317	1.8488	2.0564

Table 3: *Cut-offs for the Bonferroni and nested-Bonferroni adjustments at nominal FWER* $\alpha = 0.05$. $T_{i,n}$, $i \in \{1, \cdots, N\}$, are two-sample Welch t-statistics based on $n_1 = 6$ and $n_2 = 6$ independent observations from a normal distribution with mean 0 and standard deviation 1; $T^*_{i,n}$, $i \in \{1, \cdots, N - w + 1\}$, are dependent scan statistics and the assumed blip size is set to $w = 10$. The distribution functions $\mathcal{F}_0$ and $\mathcal{G}_0$ of $T^*_{i,n}$ and $Z_{k,n}$ are estimated by their empirical distributions based on $B = 10,000,000$ simulated observations using $w = 10$.

w	Method	NB-FWER	B-FWER	VDP-TPPFP	BH-FDR
$w = 1$	B	0.042	0.042	0.042	0.0440
	N		0.042		0.0451
$w = 2$	B	0.032	0.028	0.002	0.0476
	N		0.326		0.0719
$w = 5$	B	0.05	0.036	0.00	0.0459
	N		0.124		0.0559
$w = 10$	B	0.04	0.024	0.002	0.0449
	N		0.054		0.0498
$w = 20$	B	0.034	0.014	0.004	0.0415
	N		0.026		0.0449

Table 4: *Simulation 0. Comparison of the multiple testing procedures NB-FWER, B-FWER, VDP-TPPFP, and BH-FDR in terms of their actual Type I error rates.* B: Bootstrap, N: Normal approximation for null distributions $\mathcal{G}_0$ and $\mathcal{F}_0$ of Z and T^*, respectively. The targeted nominal Type I error rate is set to $\alpha = 0.05$. The true blip size is $w^* = 10$. The results are averages over 500 independent simulated datasets.

$w = 1$

	R	CR	IB	CIB
NB-FWER	7.51(3.59)	7.42(3.56)	5.48(2.16)	5.4(2.16)
VDP-TPPFP	7.51(3.59)	7.42(3.56)	5.48(2.16)	5.4(2.16)
BH-FDR	38.08(13.97)	36.94(13.32)	12.16(2.04)	11.17(1.39)

$w = 2$

NB-FWER	47.64(9.98)	47.54(9.89)	11.59(0.79)	11.53(0.72)
VDP-TPPFP	49.67(10.5)	49.57(10.41)	11.64(0.79)	11.58(0.71)
BH-FDR	122.05(5.46)	114.42(3.47)	17.36(2.72)	12(0)

$w = 5$

NB-FWER	112.9(3.87)	112.78(3.82)	12.04(0.2)	12(0)
VDP-TPPFP	118.33(4.28)	118.09(4.18)	12.1(0.3)	12(0)
BH-FDR	143.53(5.57)	137.15(3.55)	15.23(1.68)	12(0)

$w = w^* = 10$

NB-FWER	150.68(5.4)	150.56(5.32)	12.04(0.2)	12(0)
VDP-TPPFP	158.14(5.83)	157.88(5.71)	12.07(0.26)	12(0)
BH-FDR	187.36(8.3)	181.12(5.68)	13.93(1.47)	12(0)

$w = 20$

NB-FWER	238.29(8.58)	238.22(8.58)	12.03(0.17)	12(0)
VDP-TPPFP	250.37(9.03)	250.03(9.12)	12.06(0.28)	12(0)
BH-FDR	288.64(12.14)	280.64(9.04)	12.75(0.87)	12(0)

Table 5: *Simulation I. Fixed blip size, high separation.* The model parameters are set as follows: $\mu_1 = 0, \sigma_1 = 1$, $\mu_2 = 2, \sigma_2 = 0.75$. There are a total of 12 blips of size $w^* = 10$ (true blip size). R: # rejected hypotheses; CR: # correctly rejected hypotheses; IB: # identified blips; CIB: # correctly identified blips. Averages over $B = 100$ independent simulated datasets are reported and the standard errors of each of these estimates are given in parentheses.

	R	CR	IB	CIB
	$w = 1$			
NB-FWER	1.6(1.52)	1.51(1.44)	1.5(1.41)	1.41(1.32)
VDP-TPPFP	1.6(1.52)	1.51(1.44)	1.5(1.41)	1.41(1.32)
BH-FDR	1.46(1.72)	1.39(1.61)	1.35(1.51)	1.28(1.41)
	$w = 2$			
NB-FWER	10.88(4.9)	10.79(4.88)	5.78(1.95)	5.7(1.91)
VDP-TPPFP	10.96(5.06)	10.87(5.05)	5.8(1.98)	5.72(1.94)
BH-FDR	81.95(9.22)	76.32(8.17)	16.11(2.33)	11.97(0.17)
	$w = 5$			
NB-FWER	78.26(7.95)	78.14(7.89)	11.95(0.44)	11.9(0.36)
VDP-TPPFP	81.96(8.34)	81.83(8.27)	11.99(0.36)	11.94(0.28)
BH-FDR	123.7(6.32)	118.33(4.71)	14.82(1.7)	12(0)
	$w = w^* = 10$			
NB-FWER	114.23(7.17)	114.11(7.06)	12.04(0.2)	12(0)
VDP-TPPFP	119.71(7.6)	119.54(7.43)	12.05(0.22)	12(0)
BH-FDR	161.31(9.31)	155.73(7.2)	13.79(1.41)	12(0)
	$w = 20$			
NB-FWER	184.11(14.22)	184.04(14.23)	11.93(0.36)	11.36(0.79)
VDP-TPPFP	193.27(14.92)	193.1(14.9)	11.97(0.36)	11.54(0.72)
BH-FDR	253.27(13.75)	246.3(11.14)	12.75(0.83)	11.94(0.24)

Table 6: *Simulation II. Fixed blip size, low separation.* The model parameters are set as follows: $\mu_1 = 0, \sigma_1 = 1.5$, $\mu_2 = 1.5, \sigma_2 = 1$. There are a total of 12 blips of size $w^* = 10$. R: # rejected hypotheses; CR: # correctly rejected hypotheses; IB: # identified blips; CIB: # correctly identified blips. Averages over $B = 100$ independent simulated datasets are reported and the standard errors of each of these estimates are given in parentheses.

	R	CR	IB	CIB
	$w = 1$			
NB-FWER	1.8(1.6)	1.75(1.57)	1.66(1.44)	1.61(1.41)
VDP-TPPFP	1.8(1.6)	1.75(1.57)	1.66(1.44)	1.61(1.41)
BH-FDR	2.19(2.5)	2.15(2.45)	1.87(1.92)	1.82(1.85)
	$w = 2$			
NB-FWER	12.43(5.14)	12.39(5.14)	6.31(1.9)	6.25(1.89)
VDP-TPPFP	12.52(5.33)	12.48(5.33)	6.32(1.92)	6.26(1.92)
BH-FDR	87.58(13.47)	81.89(12.36)	16.21(2.11)	11.91(0.29)
	$w = 5$			
NB-FWER	83.38(12.89)	83.31(12.88)	11.66(0.64)	11.6(0.6)
VDP-TPPFP	87.28(13.62)	87.17(13.65)	11.78(0.56)	11.69(0.51)
BH-FDR	132.01(15.89)	126.08(14.53)	15.26(2.01)	12(0)
	$w = 10$			
NB-FWER	118.47(15.77)	118.44(15.76)	11.6(0.53)	11.15(0.81)
VDP-TPPFP	124.2(16.64)	124.08(16.56)	11.72(0.55)	11.26(0.76)
BH-FDR	170.2(18.3)	162.97(15.96)	13.95(1.36)	11.87(0.37)
	$w = 20$			
NB-FWER	170.29(27.13)	170.15(27)	10.43(1.24)	9.35(1.46)
VDP-TPPFP	178.82(28.55)	178.59(28.37)	10.72(1.13)	9.58(1.44)
BH-FDR	254.36(26.5)	244.64(23.37)	12.77(1.07)	11.2(0.84)

Table 7: *Simulation III. Variable blip size, low separation.* The model parameters are set as follows: $\mu_0 = 0, \sigma_0 = 1.5$, $\mu_1 = 1.5, \sigma_1 = 1$. There are a total of 12 blips and the blip sizes are generated from a uniform distribution with the support $[5, 16]$. R: # rejected hypotheses; CR: # correctly rejected hypotheses; IB: # identified blips; CIB: # correctly identified blips. Averages over $B = 100$ independent simulated datasets are reported and the standard errors of each of these estimates are given in parentheses.

$w = 1$

	R	CR	IB	CIB
NB-FWER	1.71(1.37)	1.66(1.36)	1.63(1.26)	1.58(1.25)
VDP-TPPFP	1.71(1.37)	1.66(1.36)	1.63(1.26)	1.58(1.25)
BH-FDR	1.6(2.26)	1.59(2.24)	1.5(2.04)	1.49(2.02)
$w = 2$				
NB-FWER	12.36(5.54)	12.33(5.56)	7.29(2.35)	7.24(2.34)
VDP-TPPFP	12.49(5.79)	12.46(5.81)	7.33(2.42)	7.28(2.41)
BH-FDR	127.03(18.74)	118.9(17.2)	26.05(3.03)	19.8(0.4)
$w = 5$				
NB-FWER	113.1(17.06)	113.08(17.07)	18.77(0.97)	18.72(1)
VDP-TPPFP	118.57(17.97)	118.53(17.95)	18.97(0.99)	18.9(1)
BH-FDR	199.23(18.89)	189.88(17.22)	24.8(2.2)	19.93(0.29)
$w = 10$				
NB-FWER	170.38(19.55)	170.33(19.51)	18.82(1.18)	17.97(1.45)
VDP-TPPFP	178.94(20.54)	178.86(20.51)	18.99(1.15)	18.25(1.36)
BH-FDR	261.05(21.97)	249.92(19.56)	22.98(1.9)	19.54(0.69)
$w = 20$				
NB-FWER	246.13(34.68)	246.01(34.7)	16.84(1.57)	14.34(1.99)
VDP-TPPFP	258.57(36.5)	258.41(36.5)	17.2(1.56)	14.78(2)
BH-FDR	398.69(34.94)	384.7(31.07)	20.9(1.41)	18.43(1.17)
$w = 30$				
NB-FWER	293.57(50.62)	293.52(50.63)	15.16(1.89)	11.3(2.26)
VDP-TPPFP	308.52(53.3)	308.34(53.29)	15.48(1.84)	11.83(2.27)
BH-FDR	528(56.13)	511.07(50.95)	19.19(1.32)	17.27(1.56)

Table 8: *Simulation IV: Variable blip size, low separation.* The model parameters are set as follows: $\mu_0 = 0, \sigma_0 = 1.5$, $\mu_1 = 1.5, \sigma_1 = 1$. There are a total of 20 blips. The blip sizes are generated from a truncated gamma distribution with mean and standard deviation of 10 probes. R: # rejected hypotheses; CR: # correctly rejected hypotheses; IB: # identified blips; CIB: # correctly identified blips. Averages over $B = 100$ independent simulated datasets are reported and the standard errors of each of these estimates are given in parentheses.

	$w = 1$	$w = 2$	$w = 10$	$w = 20$	$w = 30$
#blips identified	28	22	14	10	8
# *real blips*	8	10	13	10	8

Table 9: *p53 ChIP-Chip data. Multiple testing procedures applied to Chip A.* Number of *real blips* identified by visual inspection. A *real blip* refers to a small cluster of probes (> 1 probes) that have test statistics greater than their surroundings.

$w = 1$

	1kb of 5'	3kb of 5'	1kb of CpG	3kb of CpG	WCR	WE	Total
NB-FWER	1	3	6	13	37	6	128
VDP-TPPFP	1	3	6	13	39	7	134
BH-FDR	14	29	31	75	195	18	553

$w = 10$

	1kb of 5'	3kb of 5'	1kb of CpG	3kb of CpG	WCR	WE	Total
NB-FWER	6	14	17	39	87	1	254
VDP-TPPFP	6	14	22	45	93	1	269
BH-FDR	21	47	86	162	231	15	719

$w = 20$

	1kb of 5'	3kb of 5'	1kb of CpG	3kb of CpG	WCR	WE	Total
NB-FWER	5	11	13	27	55	2	188
VDP-TPPFP	6	11	13	28	60	2	208
BH-FDR	9	23	32	68	112	4	355

$w = 30$

	1kb of 5'	3kb of 5'	1kb of CpG	3kb of CpG	WCR	WE	Total
NB-FWER	2	4	7	23	33	0	145
VDP-TPPFP	2	4	7	23	34	0	149
BH-FDR	3	7	15	38	63	1	225

Table 10: *p53 ChIP-Chip data. Annotation information for the identified blips (bound probes) according to their genomic locations, ignoring the gap structure. "1kb of 5'"* and *"3kb of 5'"* refer to upstream regions within 1kb and 3kb of a known gene's transcription start site. *"1kb CpG"* and *"3kb CpG"* refer to regions within 1kb and 3kb of a known CpG island. *"WCR"* refers to blips that fall between the codon start site and the codon end site of a known gene (cdsStart and cdsEnd attributes of UCSC Human Genome Browser). *"WE"* refers to blips that are among *"WCR"* and actually fall within an exon. *"Total"* refers to the total number of blips identified.

$w = 1$

	1kb of 5'	3kb of 5'	1kb of CpG	3kb of CpG	WCR	WE	Total	Match
NB-FWER	1	2	6	12	36	7	123	121
VDP-TPPFP	1	2	6	12	37	7	128	127
BH-FDR	14	28	31	73	190	17	531	531

$w = 10$

	1kb of 5'	3kb of 5'	1kb of CpG	3kb of CpG	WCR	WE	Total	Match
NB-FWER	5	11	15	36	80	1	230	230
VDP-TPPFP	5	11	21	42	84	2	243	241
BH-FDR	19	40	82	150	209	14	651	651

$w = 20$

	1kb of 5'	3kb of 5'	1kb of CpG	3kb of CpG	WCR	WE	Total	Match
NB-FWER	4	7	10	22	45	2	154	154
VDP-TPPFP	5	7	10	22	49	2	171	171
BH-FDR	8	19	28	60	100	5	315	306

$w = 30$

	1kb of 5'	3kb of 5'	1kb of CpG	3kb of CpG	WCR	WE	Total	Match
NB-FWER	2	3	6	19	22	0	112	112
VDP-TPPFP	2	3	6	19	22	0	115	114
BH-FDR	5	8	15	33	50	1	185	178

Table 11: *p53 ChIP-Chip data. Annotation information for the identified blips (bound probes) according to their genomic locations, taking into account the gap structure. "1kb of 5'"* and *"3kb of 5'"* refer to upstream regions within 1kb and 3kb of a known gene's transcription start site. *"1kb CpG"* and *"3kb CpG"* refer to regions within 1kb and 3kb of a known CpG island. *"WCR"* refers to blips that fall between the codon start site and the codon end site of a known gene (cdsStart and cdsEnd attributes of UCSC Human Genome Browser). *"WE"* refers to blips that are among *"WCR"* and actually fall within an exon. *"Total"* refers to the total number of blips identified. The last column *"Match"* refers to the number of blips that are common to both Table 10 and this table.

	$w = 1$	$w = 10$	$w = 20$	$w = 30$
NB-FWER	54, 74	95, 159	56, 132	44, 101
VDP-TPPFP	58, 76	100, 169	60, 148	45, 104
BH-FDR	261, 292	272, 447	119, 236	60, 165

Table 12: *p53 ChIP-Chip data. Chromosomal distribution of the identified blips, ignoring the gap structure.* The first number of each pair represents the number of blips identified on chromosome 21 and the second number is the number of blips identified on chromosome 22.

start	stop	chr	dist 5'UTR	5'UTR gene	dist 3'UTR	3'UTR gene	dist CpG	orf?	which orf?
7777422	7777846	21	24196	-	42276	-	23703	n	-
11699915	11700701	21	551954	-	177774	-	1652	n	-
11728556	11728603	21	524052	-	149872	-	4380	n	-
15702897	15704236	21	53578	-	99446	-	53043	n	-
24057629	24058064	21	16393	-	337003	-	62690	y	BC004369, Y00264
24341029	24341731	21	219084	-	75796	-	182140	n	-
31925451	31926087	21	59971	-	85478	-	59849	n	-
33140027	33140363	21	150900	-	473814	-	163366	n	-
36853023	36853579	21	318795	-	80239	-	80772	n	-
36942786	36943309	21	383647	-	29020	-	170535	n	-
39501974	39502603	21	47845	-	96843	-	47251	n	-
40032889	40033225	21	54648	-	40069	-	41135	n	-
40169396	40169695	21	17362	-	110259	-	17043	n	-
40486104	40486430	21	21849	-	70539	-	3483	n	-
42373539	42374277	21	12097	-	722	AB001535	10970	n	-
42421540	42422134	21	24820	-	16112	-	7372	n	-
42478191	42479410	21	209058	-	29695	-	41582	y	BC021197
42504978	42505783	21	1428	BC021197	56482	-	68369	n	-
42546767	42547290	21	43217	-	98271	-	89456	n	-
42639110	42639958	21	135560	-	59193	-	414	n	-
43015275	43015544	21	12510	-	108748	-	10166	n	-
43262556	43263686	21	27745	-	42424	-	4603	n	-
43560884	43561881	21	88855	-	117566	-	69	n	-
13285986	13286389	22	740796	-	1255195	-	32703	n	-
14322904	14323825	22	160520	-	217759	-	108845	n	-
14772647	14772864	22	127825	-	182227	-	20673	n	-
15343145	15343665	22	17340	-	190826	-	78042	n	-
16007301	16007747	22	51337	-	53945	-	41565	y	D79985
16796612	16796929	22	13923	-	6367	-	13375	n	-
17236485	17236726	22	40523	-	5373	-	8554	y	AB051440
18170292	18170383	22	84750	-	88478	-	10609	n	-
18452171	18452964	22	17421	-	9711	-	13297	n	-
18587642	18587925	22	17169	-	10413	-	14000	n	-
19700988	19701235	22	26648	-	11932	-	15086	y	BC012876, BC020233, BC022098, BC030984
20274302	20274891	22	26420	-	51184	-	29022	y	X14675, X14676, Y00661
20356176	20356716	22	173279	-	3488	AF487522	51939	n	-
20395575	20396729	22	212678	-	35504	-	32224	n	-
20863694	20865084	22	30118	-	41048	-	11552	y	AK024025, U04847
21723591	21723733	22	42238	-	2495	L20493	27494	n	-
23585526	23585972	22	9624	-	31865	-	5313	y	AL050258
24185824	24186148	22	475734	-	463093	-	336592	n	-
24811133	24811596	22	272956	-	28443	-	40793	n	-
26074855	26075749	22	63154	-	75006	-	46858	y	AB051436
27315866	27316439	22	21802	-	16049	-	19339	n	-
33370077	33370416	22	4091	L29141	63195	-	8012	n	-
33541227	33541316	22	1500	BC014912	33612	-	1056	n	-
36107533	36107825	22	224785	-	9255	-	47728	n	-
42030704	42031721	22	15199	-	148537	-	6640	y	BC012187
42103497	42104704	22	62592	-	113667	-	61801	n	-

Table 13: *p53 ChIP-Chip data. 49 additional potential p53 target regions.* These regions show higher hybridization signal than the minimum hybridization signal of the 13 quantitative PCR verified regions of Cawley *et al.* (2004). Column descriptions are as follows. `start:` Start position of the blip in bp; `stop:` Stop position of the blip in bp; `chr:` Number of the chromosome that the blip is located on; `dist 5'UTR:` Distance of the blip to the closest transcription start site that is located downstream of it; `5'UTR gene:` The closest gene whose 5' upstream region contains the blip; `dist 3'UTR:` Distance of the blip to the closest 3' end of a gene; `3'UTR gene:` The closest gene whose 3' downstream contains the blip; `dist CpG:` Distance of the blip to the closest CpG island; `orf?:` Whether or not the blip falls between the transcription start and end sites of a gene; `which orf?:` If applicable, the gene for which the blip is located between the transcription start and end sites.

	$\triangleright\triangleleft - \triangleright\triangleleft$	$\triangleright\triangleleft\triangleright$	$\triangleleft\triangleright\triangleleft$	$\triangleright\triangleleft -\triangleleft$	$\triangleright\triangleleft -\triangleright$	$\triangleleft - \triangleright\triangleleft$	$\triangleright - \triangleright\triangleleft$	U	$\triangleright\triangleleft$	Total
All	4	17	17	72	86	86	72	86	118	221
Verified	0	2	2	7	7	7	7	7	8	13
Filtered	1	6	7	21	21	21	21	21	33	49

Table 14: *p53 ChIP-Chip data. Occurrences of various arrangements of the 5mer RRRCW among all the 221 blips identified by the NB-FWER approach (All), the 49 filtered blips (Filtered), and the 13 experimentally verified blips of Cawley* et al. *(2004) (Verified).* The symbol $\triangleright$ represents `RRRCW`, the symbol $\triangleleft$ represents its reverse complement `WGYYY`, and $-$ is a spacer of length 0 to 15bps. Column U is the union of the last four columns before it.

	All	Verified	Filtered
0 mismatch	4 (2.25)	0 (-)	1 (0)
1 mismatch	31 (5.4)	4 (0.25)	8 (4.25)
2 mismatch	106 (6.2)	3 (2)	26 (5.34)
`CATG` in **either** p4-p7 **or** p14-p17	45	3	14
`CATG` in **both** p4-p7 **and** p14-p17	12	3	3
`C` in p4 **and** `G` in p7	53	0	11
`C` in p14 **and** `G` in p17	75	7	23
Total	221	13	49

Table 15: *p53 ChIP-Chip data. Scanning the 221 blips identified by the NB-FWER approach (All), the 49 filtered blips (Filtered), and the 13 experimentally verified blips of Cawley* et al. *(2004) (Verified) for occurrences of the p53 consensus sequence* `RRRCWWGYYY{0-15}RRRCWWGYYY`. The first three rows report the number of blips that have matches to the consensus sequence, allowing 0-2 mismatches. Numbers in parentheses are the average spacer distances among that many mismatch occurrences. p* refers to *-th position in the consensus.