Biometrics 64, 468–478
June 2008

DOI: 10.1111/j.1541-0420.2007.00899.x

A Flexible and Powerful Bayesian Hierarchical Model for ChIP–Chip Experiments

Raphael Gottardo,[1,*] Wei Li,[2] W. Evan Johnson,[2] and X. Shirley Liu[2]

[1]Department of Statistics, University of British Columbia, Vancouver, Canada
[2]Dana Farber Cancer Institute, Harvard University, Boston, Massachusetts U.S.A.
[*]*email:* raph@stat.ubc.ca

Summary. Chromatin-immunoprecipitation microarrays (ChIP–chip) that enable researchers to identify regions of a given genome that are bound by specific DNA-binding proteins present new challenges for statistical analysis due to the large number of probes, the high noise-to-signal ratio, and the spatial dependence between probes. We propose a method called BAC (Bayesian analysis of ChIP–chip) to detect transcription factor bound regions, which incorporate the dependence between probes while making little assumptions about the bound regions (e.g., length). BAC is robust to probe outliers with an exchangeable prior for the variances, which allows different variances for the probes but still shrink extreme empirical variances. Parameter estimation is carried out using Markov chain Monte Carlo and inference is based on the joint distribution of the parameters. Bound regions are detected using posterior probabilities computed from the joint posterior distribution of neighboring probes. We show that these posterior probabilities are well calibrated and can be used to obtain an estimate of the false discovery rate. The method is illustrated using two publicly available ChIP–chip data sets containing 18 experimentally validated regions. We compare our method to four other baseline and commonly used techniques, namely, the Wilcoxon's rank sum test, TileMap, HGMM, and MAT. We found BAC and HGMM to perform best at detecting validated regions. However, HGMM appears to be very sensitive to probe outliers compared to BAC. In addition, we present a simulation study, which shows that BAC is more powerful than the other four techniques under various simulation scenarios while being robust to model misspecification.

Key words: Affymetrix tiling arrays; Bayesian hierarchical model; Empirical Bayes; Heteroscedasticity; Markov chain Monte Carlo; Mixture distribution; Multiple testing; Outlier; Spatial statistics.

1. Introduction

The advent of microarray technology (Lockhart et al., 1996) has enabled biomedical researchers to monitor changes in the expression levels of thousands of genes. Until recently, however, the mechanisms driving these changes have been harder to study in a similarly high-throughput level. A recent technological innovation, chromatin immunoprecipitation (ChIP) coupled with microarray (chip) analysis, hence the name ChIP–chip (Lee et al., 2002; Cawley et al., 2004), now makes it possible for researchers to identify regions of a given genome that are bound by specific DNA-binding proteins (transcription factors TF). Affymetrix developed the high-density oligonucleotide arrays that tile all nonrepetitive sequences of the human genome (Krapanov et al., 2002). These arrays coupled with ChIP permits the unbiased mapping of in vivo TF-binding sequences. Annotation of the TF-binding sites in a given genome is essential for building genome-wide regulatory networks, which can then be used in health research to better understand diseases and identify new targets for drugs, etc. However, the large amount of data (in the order of one million measurements for one chromosome) and the small number of replicates available is very challenging for any statistical analysis.

Similar to oligonucleotide gene expression arrays (Lockhart et al., 1996), Affymetrixtiling arrays (Affymetrix, Inc., Santa Clara) query each sequence of interest with a perfect match (PM) and a mismatch (MM) probe, where the MM probe is complementary to the sequence of interest except at the central base, which is replaced with its complementary base. The difference is that the probes used on tiling arrays do not necessarily belong to genes. This platform coupled with ChIP permits the unbiased mapping of in vivo TF-binding sequences (Cawley et al., 2004; Carroll et al., 2005). The experimental protocol using tiling arrays is described in Figure 1.This procedure generates an immunoprecipitation (IP)-enriched DNA fragment population and measures the enrichment of each PM and MM in this population. In general, a control sample is also generated and there are various ways of obtaining control populations and we refer the reader to Buck and Lieb (2004) for an overview. Currently available Affymetrix tiling arrays contain oligonucleotides of average length of 25 base pairs (bps) spanning the nonrepetitive regions of the human genome at an average resolution of 35 bps. Because the original genomic DNA is sheared into segments of length 1 kbps (Figure 1 (2)), one would expect a bound region to be of average length 20–30 probes with intensities forming a peak-like

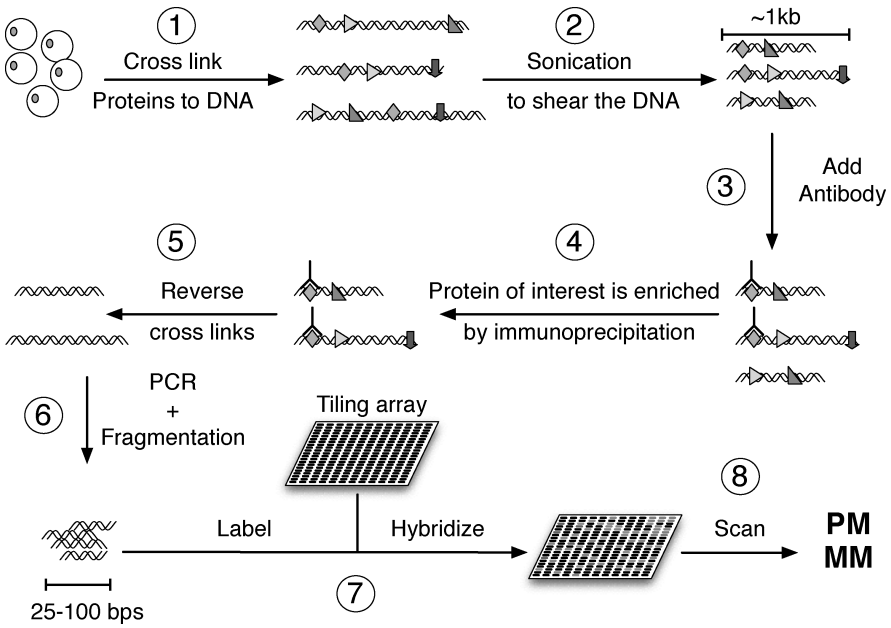

Figure 1. Details of a ChIP–chip experiment on tiling arrays. A transcription factor is cross-linked to its genomic DNA targets in vivo and the chromatin (a complex of DNA and protein) is isolated (1). The DNA with the bound TFs is sheared by sonication into small fragments of average length, 1 kbs (2). DNA fragments cross-linked to the protein of interest are enriched by immunoprecipitation with a protein-specific antibody (3–4). After the immunoprecipitation step, the DNA is separated from the protein (5) and the resulting solution (IP-enriched DNA) is amplified with polymerase chain reaction (PCR) and fragmented further into segments of size 50–100 bps (6). Then, IP-enriched DNA is fluorescently labeled and hybridized to a chip (7). After hybridization, scanning, and image processing, an intensity measurement is obtained for each PM and MM measurement (8).

structure, where the center of the peaks corresponds to probes closest to the binding site. However, empirical studies suggest that bound regions can be of variable length (Cawley et al., 2004; Keles, 2007). The analysis of ChIP–chip data consists of two steps: (a) identifying bound regions that are about 1kbps long, and (b) sequence analysis of bound regions to identify the actual binding sites and locations. Here we only deal with (a) and our ultimate goal is to identify bound regions, which can be seen as collections of adjacent probes with intensity significantly higher than the background.

A small number of approaches are available for analyzing ChIP–chip data. A common approach is to test a hypothesis for each probe and then try to correct for multiple testing (Keles, Van der Laan, and Cawley, 2004; Buck, Nobel, and Lieb, 2005). Most of the statistics used are variants of t-statistics computed for each probe using a sliding window. Keles et al. (2004) used a scan statistic, which is an average of t-statistics across a certain number of probes and Cawley et al. (2004) used Wilcoxon's rank sum (WRS) test within a certain genomic distance sliding window. A difficulty with sliding window approaches is that the resulting p-values (or t-statistics) are not independent due to the fact that each test uses information from neighboring probes, and it is challenging to devise powerful multiple adjustment procedures. Another problem with sliding window approaches is that the window size is fixed and has to be determined in advance.

Li, Meyer, and Liu (2005) used a hidden Markov model (HMM) for the identification of bound regions where the model parameters are estimated in an ad hoc way using previous results on Affymetrix SNPs arrays. Ji and Wong (2005) proposed a two-stage approach to detect bound regions. In the first step, a test statistic is computed for each statistic based on a hierarchical empirical Bayes model. In the second step, neighboring probes are combined through a moving average method (MA) or HMM.

Bayesian hierarchical models have become increasingly popular in the analysis of gene expression data (Newton et al., 2001; Lönnstedt and Speed, 2002; Parmigiani et al., 2002; Gottardo et al., 2006); they can make the best of available prior information while borrowing strength from the data when estimating the quantities of interest. Using such models, inference is usually based on the posterior distribution of the parameters. To date, there has been only one (empirical) Bayesian treatment of ChIP–chip data (Keles, 2007). The author uses a hierarchical gamma–gamma (GG) model, which is an extension of the model used in Newton et al. (2001). Even though the model is appealing by modeling the spatial structure using peaks of variable length and borrowing strength from all the probes, it has several limitations. It uses a GG hierarchical model with constant coefficient of variation, and this can have an undesired effect in the presence of probe outliers. Finally, in order to use this approach one needs to divide the data into genomic regions containing at most one peak (bound region) but such information is, in general, not available.

In this article, we introduce a flexible hierarchical Bayesian model that overcomes these limitations. Our model is built on previous approaches used in gene expression analysis (Newton

et al., 2001; Parmigiani et al., 2002) and uses mixtures to identify probes that have an intensity that is significantly different from the background. However, we take into account the spatial dependence between probes by allowing the weights of the mixture to be correlated for neighboring probes on a chromosome. A similar approach was taken in the context of array comparative genomic hybridization (CGH) (Broet and Richardson, 2006).Our model also includes an exchangeable prior for the variances, allowing each probe to have a different variance while still achieving some shrinkage. This allows us to regularize empirical variance estimates, which can be very noisy due to the small number of replicates. Finally, as we know that bound regions are made of several consecutive probes, we use the joint posterior distribution of neighboring probes to detect such regions. This, combined to the fact that each probe has its own variance, makes Bayesian analysis of ChIP–chip (BAC) very robust to probe outliers.

The article is organized as follows. Section 2 introduces the data structure and the notation. In Section 3, we present the Bayesian hierarchical model and show how we use it to detect bound regions. In Section 4, we apply our method to experimental data and compare it to four other techniques. Section 5 presents the results of a simulation study comparing our approach to the same four techniques. Finally, in Section 6 we discuss our results and possible extensions.

2. Data

We use two publicly available data sets that have already been analyzed by several research groups. Cawley et al. (2004) mapped the binding sites of three human TF, Sp1, cMyc, and p53 on chromosomes 21–22; here we focus on the p53-FL experiment. Similarly, Carroll et al. (2005) mapped the association of the estrogen receptor (ER) on chromosomes 21–22. These data contain two conditions (control and IP enriched) with three replicates each. Several binding sites have already been identified and experimentally validated, and we will use this information to compare the different methods used in this article. Both Cawley et al. (2004) and Carroll et al. (2005) used three tiling arrays, named A, B, and C, to tile all of chromosomes 21 and 22. Here we only use chip A, which represents 2/3 of chromosome 21 and contains 2 and 16 validated regions for the p53 and the ER data, respectively.

Following the idea that MM intensities are poor measures of nonspecific hybridization (Irizarry et al., 2003; Keles et al., 2004; Keles, 2007), we only used the PM intensity. The PM measurements were normalized using MAT (model-based analysis of tiling arrays) developed by Johnson et al. (2006). MAT uses the probe sequence information and copy number on each array to perform background adjustment and normalization. Such normalization is necessary to diminish probe sequence biases and to allow us to model the residual background as normal random effects. We refer the reader to Johnson et al. (2006) for further details about MAT. After normalization, the data take the form y_{cpr}, $c = 1, 2$; $p = 1, \ldots, P$; $r = 1, \ldots, R_c$, where y_{cpr} is the preprocessed intensity of probe p in condition c from replicate r.

3. Hierarchical Bayesian Modeling

In this section, we introduce the Bayesian hierarchical model used to detect bound regions. From now on, $\mathcal{G}a(a, b)$ denotes a gamma distribution with mean a/b and variance a/b^2, N (a, b) a Gaussian distribution with mean a and variance b, TN (a, b) a truncated Gaussian distribution at zero with parameters a and b, and $(x|y)$ means the conditional distribution of x given y.

3.1 *Model and Priors*

We model probe measurements as follows:

$$y_{1pr} = \mu_p + \epsilon_{1pr} \quad \text{and} \quad y_{2pr} = \mu_p + \gamma_p + \epsilon_{2pr},$$
$$\epsilon_{cpr} \sim \mathrm{N}\left(0, \lambda_{cp}^{-1}\right), \tag{1}$$

where $c = 1, 2$ denotes the treatment label equal to 1 for control and 2 for IP enriched. In (1), μ_p is the probe background intensity, and γ_p is the probe enrichment effect, which we expect to be large if probe p is part of a bound region. We model the background as a random effect with Gaussian distribution N $(0, \psi^{-1})$, where the variance ψ^{-1} is constant across probes. Even though we have used MAT to normalize the probe intensities for sequence-specific effects, we believe that it is still necessary to include probe-specific effects for two main reasons: (1) the MAT sequence normalization model is not perfect and some unexplained residual effects are likely to remain, and (2) some of the probe-to-probe variation might be due to other (nonsequence specific) factors.

To model the fact that enrichment effects can be exactly zero, we use the following prior:

$$\gamma_p \sim (1 - w_p)\delta_0 + w_p \mathrm{TN}(\xi, \tau^{-1}), \tag{2}$$

which is a mixture of a point mass at zero and a Gaussian distribution with mean ξ and variance τ^{-1} truncated at zero, where w_p is the mixing weight representing the a priori probability that probe p has positive enrichment effect. Such mixture priors have been widely used in the analysis of gene expression data (Lönnstedt and Speed, 2002; Gottardo et al., 2003, 2006). Here we use a truncated normal at zero as enrichment effects should be positive. Note also that we allow the mixing weights to be probe specific and to spatially vary with the genomic location. Similar to Broet and Richardson (2006) in the context of CGH arrays, we model the probe dependence and borrow strength from neighboring probes by relating the weights, the w_p's, to a latent Markov random field prior $\boldsymbol{\theta} = \{\theta_p, 1 \le p \le P\}$'s, by a logistic transformation $w_p = \exp(\theta_p)/(1 + \exp(\theta_p))$. We use a Gaussian intrinsic autoregression model (Besag and Kooperberg, 1995) for $\boldsymbol{\theta}$ as follows:

$$(\theta_p \,|\, \theta_{\partial p}) \sim \mathrm{N}\left(\frac{\sum_{p' \in \partial p} \theta_{p'}}{n_p}, \frac{n}{n_p \kappa}\right), \tag{3}$$

where ∂p corresponds to the probes p' immediately adjacent to p, n is the number of neighboring probes used, $n_p \le n$ is the cardinality of ∂p, and κ is a smoothing parameter. Basically, n_p is n for all probes except the ones at the two extremities for which n_p will vary between $n/2$ and n. The conditional distributions given by (3) correspond to a valid, but improper joint distribution, given by

$$\pi(\theta_1,\ldots,\theta_P) \propto \exp\left(-\frac{\kappa}{2n}\sum_{p}\sum_{p'\in\partial_p,p'>p}(\theta_p-\theta_{p'})^2\right), \quad (4)$$

see Besag and Kooperberg (1995) for details. Intuitively, this joint distribution is improper as the overall level is not fixed; adding a constant to all of the θ_p's does not change (4). A common solution is to impose an identifiability constraint, for example, $\sum_p \theta_p = 0$ or fix one of the θ_p's, such that the resulting $(P-1)$-dimensional density becomes proper. In the context of ChIP–chip and tiling arrays, the very first probe of a given chromosome should not be enriched and a reasonable solution would be to fix the corresponding θ to a small value such that the corresponding weight is virtually zero. Here we chose $\theta_1 = -5$, which leads to a value of w_1 less than 0.001. Given the large number of probes, the exact value has little influence on the posterior. We have also tried the constraint $\sum_p \theta_p = 0$ and obtained essentially the same results. However, we prefer to fix one of the θ_p's as it leads to a simpler, unconstrained, Markov chain Monte Carlo (MCMC) algorithm (see Web Appendix A).

The prior given by (3) will induce similar mixing weights across neighboring probes, thus encouraging neighboring probes to be of the same class (enriched or not enriched). In our application we use $n = 10$, based on empirical studies suggesting that bound regions can contain as few as 10 probes; however, the exact value is not crucial. We have experimented with values from 2 to 20 and observed little difference in the estimated parameters. Formulations (2) and (3) were chosen both for their flexibility and computational convenience; they should be seen as an approximation to the true biological/experimental process inducing enrichment in probe intensities. This said, we will see later that our model provides good results when applied to both experimental and synthetic data.

We regularize noisy variance estimates by borrowing strength from all the probes with an exchangeable prior for the probe precisions (Parmigiani et al., 2002; Gottardo et al., 2006; Lewin et al., 2006), defined by $(\lambda_{cp} \,|\, \alpha_c, \beta_c) \sim \mathcal{G}a(\alpha_c^2/\beta_c, \alpha_c/\beta_c)$, that is, a gamma distribution with mean α_c and variance β_c.

Finally, we use the following priors for the hyperparameters: α_c and β_c are taken uniform over [0,1000], τ, ψ, and κ are assumed to come from an exponential distribution with mean 1000, and ξ is taken to be uniform between 0 and 15. All these priors are vague but proper and will not have much influence in the posterior because the parameters are shared across probes, and so there is plenty of information in the data.

3.2 *Parameter Estimation*

Realizations were generated from the posterior distribution via MCMC algorithms (Gelfand and Smith, 1990); see Web Appendix A for details. We used four parallel chains started from different values, each run for 10,000 iterations after discarding the first 1000. This allowed us to check for convergence issues and obtain more stable estimates by combining the four chains. Trace and autocorrelation plots did not reveal any convergence problems; see Web Figure 1 in supplementary material. An `R` software package called `BAC` implementing the method is available from Bioconductor at `www.bioconductor.org`.

3.3 *Inference and Detection of Bound Regions*

Our ultimate goal is to identify bound regions, and this can be done using parameter estimates from our model. Let us define $z_p \equiv \mathbf{1}(\gamma_p > 0)$, that is, z_p is equal to 1 if the associated enrichment effect is strictly positive. By definition of (2) it is straightforward to compute estimates of the marginal posterior probabilities of enrichment, defined as $\varrho_p(\boldsymbol{y}) \equiv \Pr(z_p = 1 \,|\, \boldsymbol{y})$, from the MCMC output. Here, we expect bound regions to be constituted of several consecutive probes with positive enrichment effects. Thus, to detect bound regions we propose to look at the joint distribution of neighboring probes, in particular the joint distributions of the z_p's. This is similar to the joint modeling approach of Keles (2007). We call a probe p part of a bound region if, in a window of size $2w+1$ probes centered at p, at least m probes have positive enrichment effects. We define the associated joint posterior probabilities as

$$v_p(w,m,\boldsymbol{y}) \equiv \Pr\left(\sum_{k=p-w}^{p+w} z_k \geq m \,\middle|\, \boldsymbol{y}\right), \quad (5)$$

where w is the predetermined window size and m is the minimum number of probes with positive enrichment effect tolerated in the window. The rationale is that for a fixed (large enough) window size, if only few isolated (noisy) probes have large enrichment effects, the corresponding joint probability would still be small. Similarly, if within a bound region, a few isolated probes have small enrichment effects, the overall joint probability would still be large. We found the values $w = 5$ and $m = 6$ to work well in practice. The value of w is consistent with the value of $n = 10$ used in (3) and the window size used by Keles (2007), whereas $m = 6$ was chosen for robustness to outlying probes and to account for the fact that probes at the extremities of a bound region only have half of their neighboring probes with positive enrichment effect due to the peak-like structure. As we will see, this would allow for more accurate identification of a bound region's endpoints. Note that the estimation of $v_p(w,m,\boldsymbol{y})$ is trivial using MCMC because one obtains samples from the full posterior distribution. Probes belonging to bound regions can be selected by applying a joint posterior probability cutoff. Here, we investigate two such cutoffs: a 0.5 cutoff, corresponding to the usual 0–1 loss and a false discovery rate (FDR) cutoff. The FDR cutoff can be selected using a direct posterior probability calculation as described in Newton et al. (2004). Finally, we follow the approach taken by Cawley et al. (2004) and merge resulting regions separated by 500 bps or less.

In order to look for groups of probes with large enrichment effects, one could be tempted to average the marginal posterior probabilities over a sliding window of size $2w+1$, which can be related to the expected number of enriched probes, $n_p(w,\boldsymbol{y})$, within the same window, via

$$n_p(w,\boldsymbol{y}) \equiv \mathbb{E}\left[\sum_{k=p-w}^{p+w} z_k \,\middle|\, \boldsymbol{y}\right] = \sum_{k=p-w}^{p+w} \varrho_p(\boldsymbol{y}). \quad (6)$$

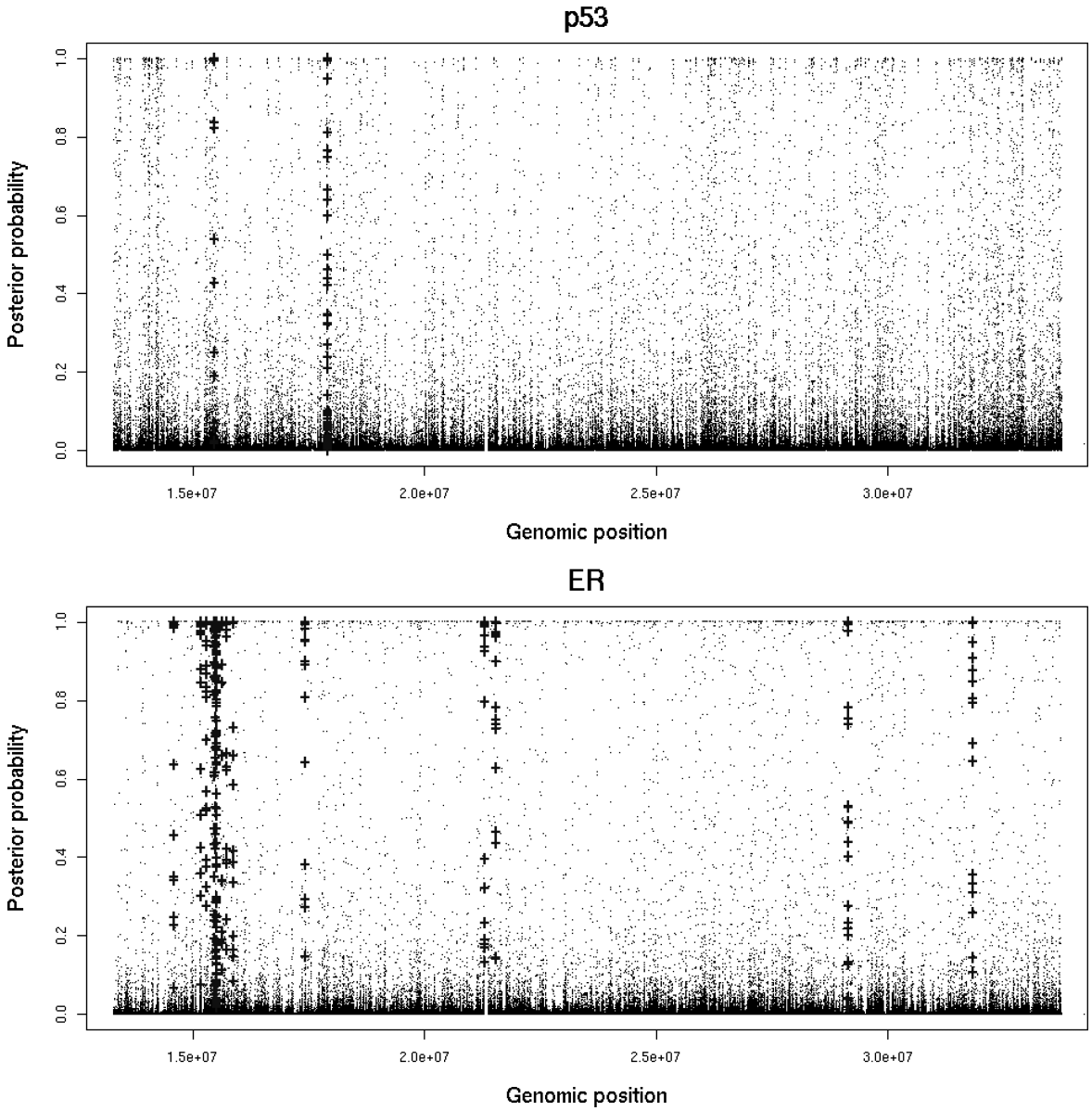

Figure 2. Marginal posterior probabilities of enrichment versus genomic positions for the p53 data (top) and the ER data (bottom). All validated regions (shown with red plus signs in the electronic version of this article) contain probes with probabilities close to 1. Many isolated probes have posterior probabilities close to 1.

However, the information contained in $v_p(w, m, \boldsymbol{y})$ is much greater than the information contained in $n_p(w, \boldsymbol{y})$. In fact the two can be related via Markov's inequality,

$$v_p(w,m,\boldsymbol{y}) = \Pr\left(\sum_{k=p-w}^{p+w} z_k \geq m \,\middle|\, \boldsymbol{y}\right) \leq \frac{\mathbb{E}\left[\sum_{k=p-w}^{p+w} z_k \,\middle|\, \boldsymbol{y}\right]}{m} = \frac{n_p(w,\boldsymbol{y})}{m}, \quad (7)$$

and if $v_p(w, m, \boldsymbol{y})$ is large then $[n_p(w, \boldsymbol{y}) \geq m]$ will be likely to be true, but the converse is not true! This shows that using a sliding window approach based on the marginal posterior probabilities would be suboptimal. Of course, the posterior probabilities (both marginal and joint) described here depend on model assumptions and provide only an approximation of the reality. However, as we will see in the next section with experimental data, the joint posterior probabilities can lead to good detection of validated regions.

4. Application to Experimental Data

4.1 *Illustration of BAC on the p53 and ER Data*

We have applied BAC to both the p53 and ER data. Figure 2 shows the marginal posterior probabilities of enrichment versus genomic positions for both data sets. Overall the p53 data exhibit more activity than the ER data in which more probes have posterior probabilities close to 1. As expected, most of the probes have small posterior probabilities

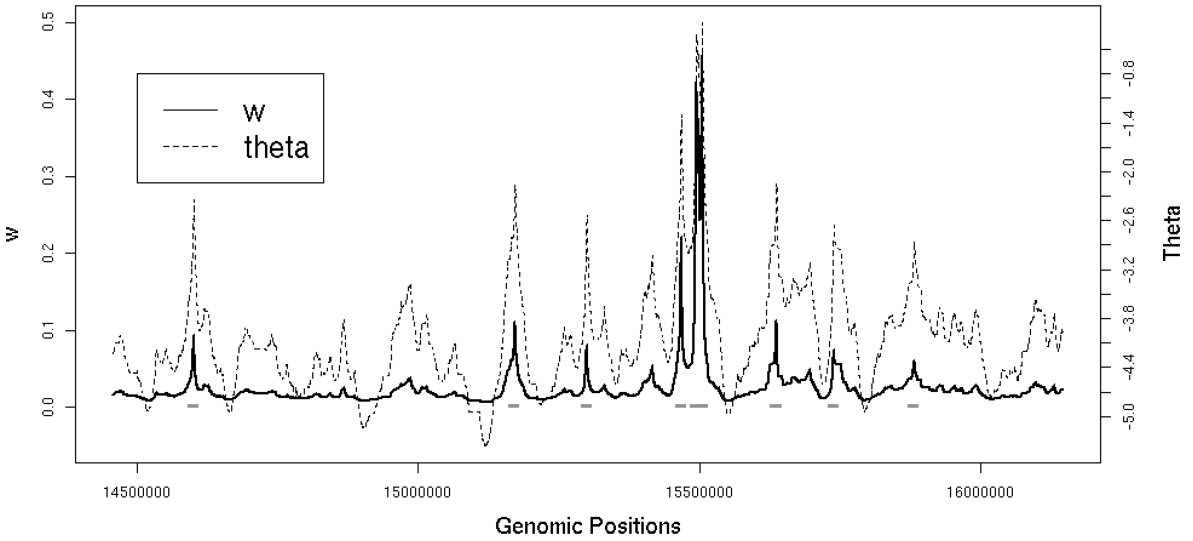

Figure 3. Posterior means of the mixing weights, the w_p's, and corresponding smoothing parameters, the θ's, versus genomic positions for the ER data. For clarity, only part of the data are shown. Validated regions are shown with gray marks along the axis $w = 0$. As expected, validated regions have large mixing weights.

and probes part of validated regions have, for the most part, large probabilities. In order to demonstrate the smoothing effect of the Markov random field on the estimated weights, we have plotted the posterior means of the w_p's as a function of the genomic positions for part of the ER data (Figure 3). The smoothing clearly helps to estimate probe-specific weights by borrowing strength across neighboring probes while allowing enriched regions to get larger probabilities.

Figure 2 also shows that many isolated probes have large marginal posterior probabilities, and averaging the marginal posterior probabilities over windows of size 11 does not help (Web Figure 2 in supplementary material). However, as we have said earlier, we are interested in groups of probes with positive enrichment effects. Figure 4 shows that the joint posterior probabilities, $v_p(5, 6, \boldsymbol{y})$, seem to be better calibrated and allow for much better separation of the validated regions from the noise.

Some of the validated regions used here were initially detected using MA methods with fixed width (Cawley et al., 2004; Johnson et al., 2006); thus some probes could belong to "validated" regions even though they might not be within the true bound regions. Web Figure 3 in supplementary material shows that BAC get better resolutions of bound regions than MA approaches with fixed width, which could significantly improve further detection of binding sites by sequence analysis.

4.2 *Comparisons*

We now compare BAC to WRS (Cawley et al., 2004), TileMap (Ji and Wong, 2005), MAT (Johnson et al., 2006), and hierarchical gamma minture model (HGMM; Keles, 2007) on both the p53 and ER data. We briefly review the different approaches:

WRS: Cawley et al. (2004) used a WRS with a ±500 bps sliding window approach to detect bound regions. Genomic positions belonging to transcription factor binding site (TFBS) were defined by applying a p-value cutoff of 10^{-5}, resultant positions separated by 500 bps are merged to form a predicted TFBS.

TileMap: Ji and Wong (2005) described an approach where neighboring probes are combined through an MA method or an HMM. Unbalanced mixture subtraction is proposed to provide estimates of local false discovery rate for MA and model parameters for HMM.

MAT: MAT can also detect bound regions using a sliding window approach based on a trimmed mean statistic combined to an FDR estimation procedure (Johnson et al., 2006).

HGMM: Keles (2007) proposed a hierarchical gamma mixture model of binding intensities while incorporating inherent spatial structure of data. Parameters are estimated by maximizing the marginal likelihood using the EM algorithm. Inference is based on the posterior probabilities.

To ease comparison, we applied all methods (except HGMM) on the MAT-normalized data. HGMM could not be applied to the MAT-normalized data because it uses a gamma model and requires the measurements to be positive. Therefore HGMM was used on the quantile–quantile normalized log-transformed PM measurements as suggested by Keles (2007). For each method, we used the default cutoff values and (if needed) a few others that are comparable to the cutoffs used in BAC. For the MA method of TileMap, we used a local FDR of 0.5 whereas for the HMM model we used a cutoff of 0.5 for the posterior probabilities; by definition of the FDR these two cutoffs are comparable (Efron, 2004). For WRS we used the p-value cutoff of 10^{-5} used in Cawley et al. (2004) as well as another cutoff to control the FDR at 0.1 using the method proposed by Benjamini and Hochberg (1995). When detecting regions with MAT, we fixed the FDR at 0.1. For both BAC and HGMM, we used a 0.5 posterior probability cutoff and an FDR cutoff of 0.1 using a direct posterior probability approach (Newton et al., 2004). The results are

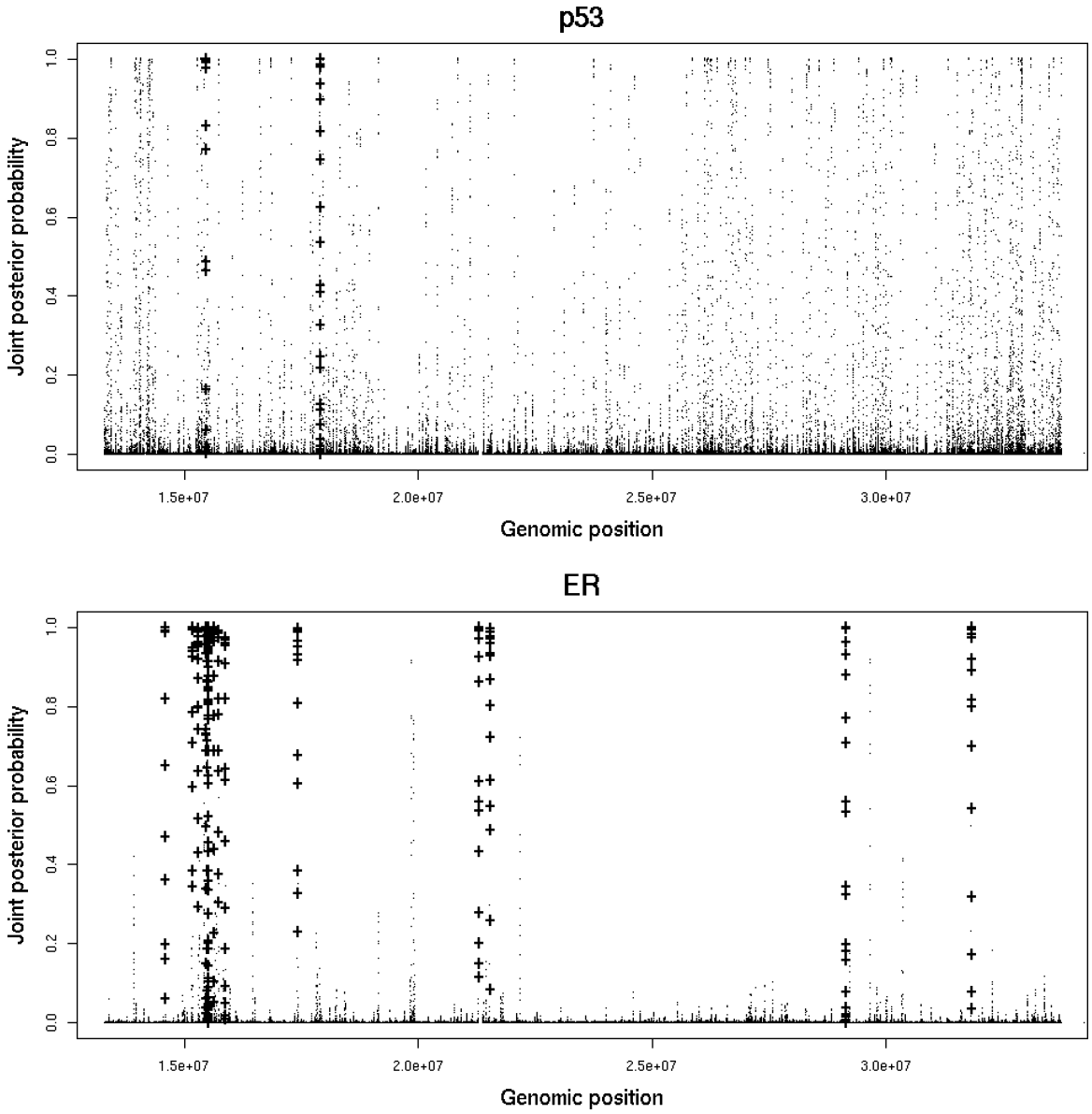

Figure 4. Joint posterior probabilities versus genomic positions for the p53 data (top) and the ER data (bottom). All validated regions (shown with red plus signs in the electronic version of this paper) contain probes with probabilities close to 1. The joint posterior probabilities reflect the knowledge that bound regions are made of consecutive probes with positive enrichment, and no isolated probes have such large probabilities.

summarized in Table 1. Both HGMM and BAC detect all validated regions and seem to perform better than TileMap, WRS, and MAT, which fail to detect validated regions at fixed FDR or at posterior probability thresholds. In order to give a better comparison in terms of ranking performance, we have also looked at how many of the qPCR-validated regions were detected in the top 50, top 20, and top 10 for all methods. The results, summarized in Table 1, show that TileMap MA, MAT, and BAC always find the most validated regions possible in each of the top 10, 20, or 50 regions. In comparison, all others fail to detect some of the validated regions even when increasing the number of regions to 50.

Note that the two p53-validated regions, not detected here, were originally found with WRS by pooling two different samples to overcome the small sample size; see Cawley et al. (2004) for more details. Overall, HGMM detects more regions than BAC but a visual inspection of the observed intensities for the regions detected by HGMM but not by BAC shows that many of them contain probe outliers; see Web Figure 4 in supplementary material. This is due to the fact that HGMM assumes a constant coefficient of variation. Similar observations have been made with the original GG model used by Newton et al. (2001) in the context of differential gene expression; see Gottardo et al. (2006).

Table 1

Number of regions (total and validated) detected by each method on the p53 and ER data. Only HGMM and BAC detect all of the validated regions at fixed FDR/posterior probability thresholds. Best results in terms of detection of validated regions are highlighted in bold.

		p-53		ER	
Method	Cutoff	qPCR validated (2)	Total	qPCR validated (16)	Total
WRS	10^{-5}	1	4	5	5
	0.1 FDR	1	6	7	10
	Top 50	**2**	50	13	50
	Top 20	**2**	20	8	20
	Top 10	1	10	7	10
TileMap HMM	0.5 pp	**2**	82	9	9
	Top 50	**2**	50	14	50
	Top 20	**2**	20	13	20
	Top 10	**2**	10	10	10
TileMap MA	0.5 fdr	**2**	10	0	0
	Top 50	**2**	50	**16**	50
	Top 20	**2**	20	**16**	20
	Top 10	**2**	10	**10**	10
MAT	0.1 FDR	1	3	**16**	26
	Top 50	**2**	50	**16**	50
	Top 20	**2**	20	**16**	20
	Top 10	**2**	10	**10**	10
HGMM	0.5	2	242	16	38
	0.1 FDR	**2**	132	**16**	42
	Top 50	1	50	**16**	50
	Top 20	1	20	14	20
	Top 10	1	10	9	10
BAC	0.5 pp	**2**	209	**16**	24
	0.1 FDR	**2**	116	**16**	27
	Top 50	**2**	50	**16**	50
	Top 20	**2**	20	**16**	20
	Top 10	**2**	10	**10**	10

Finally, we have also compared various simplifications of our model to see which aspects of it are important. We have looked at the following simplifications: (i) Set the probe-specific background effects to 0 ($\mu_p \equiv 0$), (ii) Assume that the mixing weights are constant across probes ($w_p \equiv w$), and (iii) Assume that the variance is constant across probes ($\lambda_{cp} \equiv \lambda_c$). Web Table 1 in supplementary material shows that all three features are important and that removing them significantly affects the detection of the validated regions.

5. Simulation Studies

We now use a series of simulations to study the performance of BAC under various model specifications compared to the four methods described previously. In order to do so, we generated data sets both from a Gaussian hierarchical model satisfying (1) and from the hierarchical GG model of Keles (2007). In each case, we generated 100 data sets with 50,000 probes and three replicates in both control and treatment conditions. In order to form enriched regions, we also need probe-genomic coordinates, and we used the first 50,000 genomic positions of chromosome 21 for that.

For the Gaussian hierarchical model, enriched regions were assumed to describe a peak with intensity function given by $A \exp\{-4(x_p - C)^2/B^2\}$, where A is the amplitude of the peak, B controls the width of the peak, C represents the center of the peak, and x_p is the genomic position of probe p. Under this parameterization, the peak intensity is approximately zero as soon as x_p is $B/2$ bps away from C. For each data set, we fixed the number of enriched regions to 50, whose centers (the C's) were randomly generated across the set of possible coordinates while imposing a separation of at least 3 kbps between peaks. The width of the peaks (B parameters) were generated from a uniform distribution between 600 bps and 1 kbps. Similarly, the amplitude of the peaks was randomly generated from a uniform distribution between 0.5 and 4. Probe enrichment effects, the γ_p's in (1), were set to their corresponding values given by $A \exp\{-4(x_p - C)^2/B^2\}$ if the probes were within $B/2$ of a peak center and zero otherwise. Finally, the probe-specific backgrounds (μ_p's) were generated from a Gaussian distribution with mean 0 and variance ψ^{-1}, whereas the probe-specific precisions in each condition were generated from a Gamma distribution with mean α_c and variance β_c. The parameters ψ, α_c, and β_c were set to the BAC

posterior mean estimates from the ER data. In order to assess model robustness, we have also tried variants of this where we fixed all probe backgrounds to zero and constrained the precision (and thus the variance) to be constant across probes, which we set to α_c (the mean of the precision distribution). In summary, we have four combinations: [probe-specific background (PB) or no background (NB)] × [probe-specific variance (PV) or constant variance (CV)].

For the hierarchical GG model, data were generated as described in Keles (2007) with the setting described in Section 4 and the unknown parameters fixed to the values estimated by HGMM on the ER data.

Data generated from the Gaussian model were first exponentiated before using HGMM, as measurements need to be positive. Similarly, data generated from the GG model were log transformed before using TileMap, BAC, WRS, and MAT to make the data more normally distributed. In each case, we summarized the results using two plots: a receiving operating characteristic (ROC) curve, which shows the number of true positive regions against the number of false positive regions found (averaged across the 100 data sets) when varying the cutoff for each method, and a plot of the nominal FDR against the true FDR (averaged across the 100 data sets). For the latter, we only show the methods that can control the FDR, namely, HGMM, WRS, MAT, and BAC.

Let us first look at the results from the hierarchical Gaussian model (Figure 5). In terms of ROC curves, most methods perform better when the variance is constant, whereas the addition of probe-specific background does not affect the result much. The latter is not surprising as all methods either work on the difference between the IP and control conditions (thus removing the background) or account for probe-specific signals (HGMM and BAC). The fact that all methods (including BAC) perform better when the variance is constant is mainly due to the gamma prior for the probe precisions, which tend to generate noisier data when the variance is indeed not constant. Overall, BAC performs very well on all variants of the Gaussian hierarchical model with the best ROC curves. TileMap performs second best overall and slightly better than TileMap HMM, especially when the variance is not constant. In comparison, MAT and WRS are not as good, this is particularly true of MAT. MAT uses a trimmed mean, which probably removes some of the true signal (e.g., highest values of the peak). HGMM performs the worst with an extremely large false positive rate, which is why the curve cannot be seen in Figure 5 (a–d).In terms of FDR, BAC has the closest curve to the expected line $y = x$, whereas other methods tend to underestimate the true FDR. Again, HGMM performs the worst with a nominal FDR way below the true FDR.

Looking at the results from the hierarchical GG model, HGMM performs the best with the highest ROC curve and an FDR curve almost perfectly aligned with the line $y = x$. TileMap MA has good performance again. BAC still performs relatively well and third best after HGMM and TileMap MA, whereas the performance of MAT and WRS clearly deteriorated. In addition, BAC still gives an FDR curve that is not too far from the actual line $y = x$, at least for values between 0 and 0.3, which contains the range of FDR values used in practice.

6. Conclusion

We have developed a framework, named BAC, for detecting bound regions with ChIP–chip experiments in a way that is robust to outlying measurements and is powerful even with a small number of replicates. In two ChIP–chip experiments on Affymetrix tiling arrays, we compared BAC to four other baseline and commonly used methods, and it performed better, at least in terms of the number of validated regions detected and robustness to probe outliers. In addition, we have performed a simulation study, which showed that BAC is robust to model misspecification and can outperform other methods in a wide variety of settings. Our model requires more computing (roughly 10 hours for a data set with 300,000 probes on a personal computer) than some other methods because it involves MCMC, and users would need to decide whether the improved results are worth the additional computing time.

In this article, we considered a homogeneous spatial structure to induce smoothing in the weights of the mixtures. For areas of the genome where there are few binding sites, one might fear that this could lead to oversmoothing. However, as seen with the two data sets used here, bound regions are wide enough and this is not a problem. Based on the assumption that probes are roughly equally spaced, we have opted for a Markov random field prior that did not explicitly use the genomic distance between them. However, if needed, a spatial prior that uses the genomic distance could easily be used instead (Ripley, 2004).

We assumed that normalization was done as a preprocessing step using MAT (Johnson et al., 2006); we found this normalization step to be necessary in order for the background effect to be approximately normal. Although one could incorporate the normalization and background correction into our model, this would severely increase the complexity of the model and does not seem worthwhile.

In order to compare our method with others, we identified bound regions if the joint posterior probability was greater than a given threshold. In practice, though, we would often not use a cutoff, but instead would report the posterior probabilities themselves. A biologist could then choose to do further research on a number of the most likely regions, taking account of resource constraints, or to study regions whose posterior probability exceeds a prespecified threshold (e.g., FDR).

In this article, we have compared our model with four alternatives, but there are other methods for detecting bound regions with ChIP–chip data. We chose these four because they are either obvious baseline methods or widely used; they are also representative of other methods. For example, there are several other sliding window approaches that we could have used (Keles et al., 2004; Buck et al., 2005).

Given the complexity of our hierarchical model it is not possible to use standard diagnostic tools to check model assumptions. However, it is possible to look at posterior quantities from our model to check some assumptions; see Web Figure 5–7 in supplementary material. Finally, we would like to say that even though our model was illustrated with Affymetrix arrays, it could easily be used with other oligonucleotide type of arrays, such as Nimblegen or Agilent. Though, the model may need to be modified slightly.

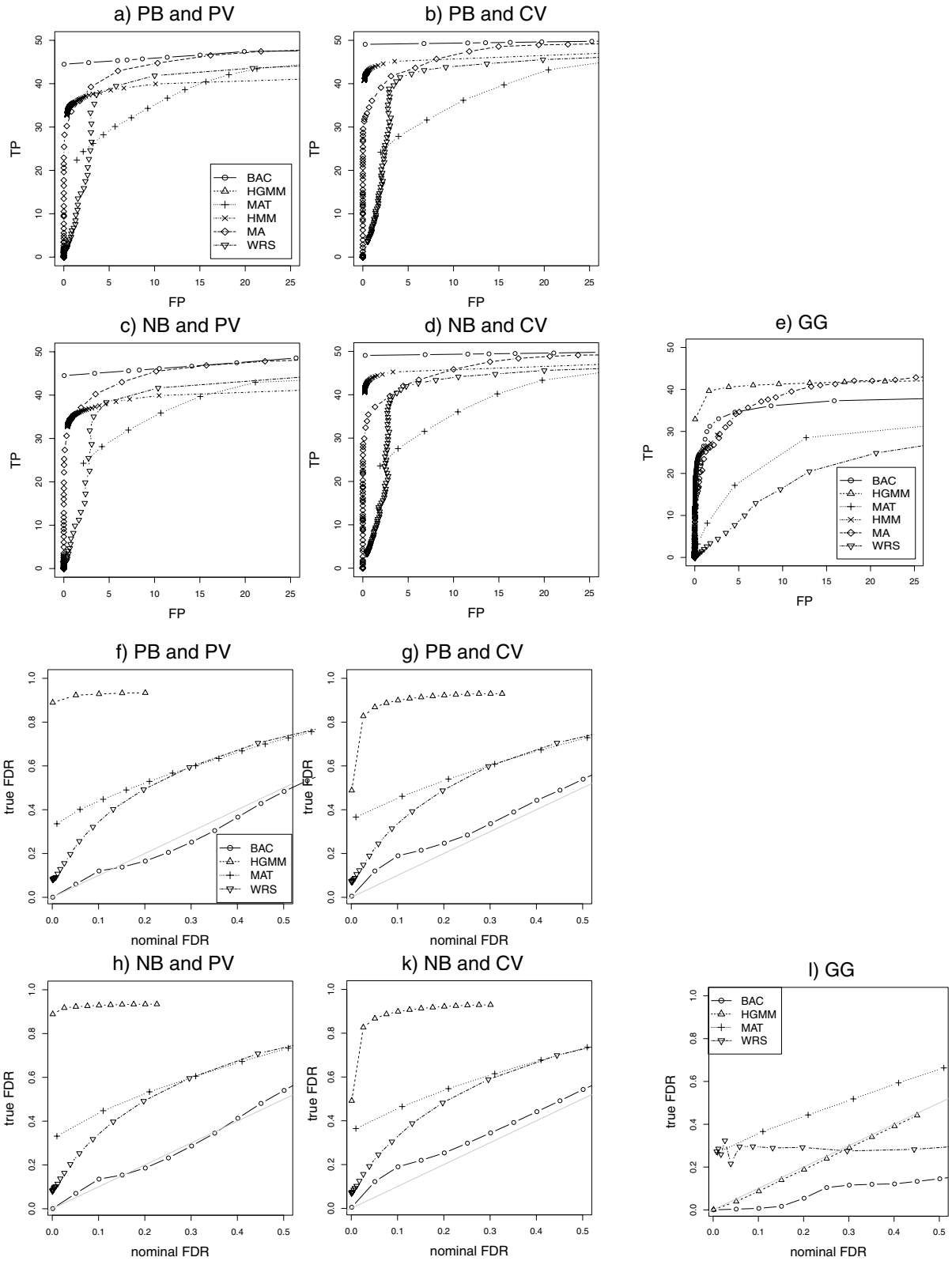

Figure 5. Receiving operating characteristics (a–e) and FDR curves (f–l) for all methods applied to data generated from variants of the Gaussian hierarchical model (a–d; f–k) and the hierarchical gamma–gamma model of Keles (e; l). Variants of the Gaussian hierarchical models considered are the four possible combinations of probe-specific background (PB), no background (NB), probe-specific variance (PV), and constant variance (CV). For the FDR plots the solid gray line corresponds to the expected line "nominal FDR = true FDR."

7. Supplementary Material

Web Appendices, Tables, and Figures referenced in Sections 3, 4, and 6 are available under the Paper Information link at the Biometrics website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors thank Jenny Bryan for helpful comments, Sunduz Keles for helpful discussion about HGMM and the associated `R` package, and three anonymous referees, the editor, and associate editor for suggestions that clearly improved an earlier draft of the article. X.S.L and W.L. were supported by NIH grant 1R01 HG004069-01.

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.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82,** 733–746.

Broet, P. and Richardson, S. (2006). Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics* **22,** 911–918.

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

Buck, M. J., Nobel, A. B., and Lieb, J. D. (2005). ChIPOTIe: A user-friendly tool for the analysis of ChIP-chip data. *Genome Biology* **6,** R97.

Carroll, J. S., Liu, X. S., Brodsky, A. S., et al. (2005). Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein foxa1. *Cell* **122,** 33–43.

Cawley, S., Bekiranov, S., Ng, H., et al. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of non-coding RNAs. *Cell* **116,** 499–511.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of American Statistical Association* **99,** 96–104.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association* **85,** 398–409.

Gottardo, R., Pannucci, J. A., Kuske, C. R., and Brettin, T. (2003). Statistical analysis of microarray data: A Bayesian approach. *Biostatistics* **4,** 597–620.

Gottardo, R., Raftery, A. E., Yeung, K. Y., and Bumgarner, R. (2006). Bayesian robust inference for differential gene expression in cDNA microarrays with multiple samples. *Biometrics* **62,** 10–18.

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

Ji, H. and Wong, W. (2005). Tilemap: Create chromosomal map of tiling array hybridizations. *Bioinformatics* **18,** 3629–3636.

Johnson, E. W., Li, W., Meyer, C., Gottardo, R., Carroll, J., Brown, M., and Liu, S. (2006). Model-based analysis of tiling-arrays for chip-chip. *Proceedings of the National Academy of Sciences* **103,** 12457–12462.

Keles, S. (2007). Mixture modeling for genome-wide localization of transcription factors. *Biometrics* **63,** 10–21.

Keles, S., Van der Laan, M. J., and Cawley, S. E. (2004). *Multiple testing methods for ChIP-chip high density oligonucleotide array data.* Technical report, Biostatistics, Berkeley, CA.

Krapanov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P., and Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296,** 916–919.

Lee, T. I., Rinaldi, N. J., Robert, F., et al. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science* **298,** 799–804.

Lewin, A., Richardson, S., Marshall, C., Glazier, A., and Aitman, T. (2006). Bayesian modelling of differential gene expression. *Biometrics* **62,** 1–10.

Li, W., Meyer, C. A., and Liu, X. S. (2005). A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* **21,** i275–i282.

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14,** 1675–1680.

Lönnstedt, I. and Speed, T. P. (2002). Replicated microarray data. *Statistica Sinica* **12,** 31–46.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8,** 37–52.

Newton, M., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5,** 155–176.

Parmigiani, G., Garrett, E. S., Anbazhagan, R., and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **64,** 717–736.

Ripley, B. D. (2004). *Spatial Statistics.* New York: Wiley.

Received July 2006. *Revised May* 2007.
Accepted July 2007.