Mixture modeling for genome-wide localization of transcription factors

Sündüz Keleş

Department of Statistics and Department of Biostatistics & Medical Informatics

1300 University Avenue, 1245B Medical Sciences Center, Madison, WI 53706

March 28, 2006

Summary. Chromatin immunoprecipitation followed by DNA microarray analysis (ChIP-chip methodology) is an efficient way of mapping genome-wide protein-DNA interactions. Data from tiling arrays encompass DNA-protein interaction measurements on thousands or millions of short oligonucleotides (probes) tiling a whole chromosome or a genome. We propose a new model-based method for analyzing ChIP-chip data. The proposed model is motivated by the widely used two-component multinomial mixture model of *de novo* motif finding. It utilizes a hierarchical gamma mixture model of binding intensities while incorporating inherent spatial structure of the data. In this model, genomic regions belong to either one of the following two general groups: regions with a local protein-DNA interaction (peak) and regions lacking this interaction. Individual probes within a genomic region are allowed to have different localization rates accommodating different binding affinities. A novel feature of this model is the incorporation of a distribution for the peak size derived from the experimental design and parameters. This leads to the relaxation of the fixed peak size assumption that is commonly employed when computing a test statistic for these types of spatial data. Moreover, the proposed framework promotes the control of the false discovery rate at the peak level rather than the probe level and this leads to less conservative analysis. Simulation studies and a real data application demonstrate good operating characteristics of the method including high sensitivity with small sample sizes when compared to available alternative methods.

Key words: Chromatin immunoprecipitation; False discovery rate; Hierarchical mixture model; Microarrays; Transcription factor; Tiling arrays.

email: keles@stat.wisc.edu

1. Introduction

Up until recently, most genomic research has focused on coding DNA. Microarray gene expression analysis has been widely used to identify subsets of target genomes that have similar transcription levels under various experimental conditions. What remains a challenge is to identify regulatory mechanisms governing similar expression patterns.

Sequence specific DNA binding proteins, also known as transcription factors (TFs), regulate transcription in eukaryotes. Each transcription factor, or group of related factors, recognizes a unique family of short sequence elements, usually between 5 and 20 base pairs (bps) in length. The non-coding DNA sequence surrounding a gene determines when and where the gene will be expressed, just as the coding sequence determines the gene's molecular function. Understanding precisely how this regulatory information is encoded in the genome is one of the major problems in current biology. A recent technological innovation, chromatin immunoprecipitation (ChIP) coupled with microarray (chip) analysis, hence the name ChIP-chip, has enabled researchers to identify regions of a given genome that are bound by specific DNA binding proteins. This genome-wide localization methodology is highly promising for annotating transcription factor binding sites on genome scale. It has been applied using both high density oligonucleotide arrays, i.e., tiling arrays, (Cawley et al., 2004) and spotted two color microarrays (Ren et al., 2000 and Weinmann et al., 2001). The experimental protocol depends on the array platform used. We refer to Buck and Lieb (2004) for an excellent overview of this technology and experimental details. The key point of ChIP-chip experiments is to detect differences between the intensity measurements of IP-enriched (treatment) and control hybridizations for each array element. It is expected that the array elements showing *significant* differences belong to genomic locations in the vicinity of binding sites for the TF of interest. Thus, analysis of ChIP-chip experiments consists of two steps: (1) identifying bound regions that are in the order of $\sim$ 1000 bps; (2) sequence analysis of bound regions to identify specific binding sites. The latter step is sensitive to the quality of identified bound regions. In this paper, we are concerned with developing a formal modeling framework for the first step.

A small number of available approaches for analyzing these data involve testing for each probe,

under a common null distribution, whether it is part of a bound DNA fragment. In particular, Cawley et al. (2004) and Keleş et al. (2004) have used various sliding window test statistics. Keleş et al. (2004) utilized a scan statistic which is an average of two sample t-statistics across certain number of probes whereas Cawley et al. (2004) used Wilcoxon rank sum test within a certain genomic distance in a sliding window fashion. Regardless of the test statistic used, one is left with a large multiple testing problem, i.e., over 100,000 tests per chromosome. One critical issue of the multiple testing framework, for controlling various error rates in this high dimensional testing, is the estimation of the joint null distribution of the test statistics. Although there are powerful resampling based approaches for estimating such a joint null distribution (Pollard and van der Laan, 2002), a typical ChIP-chip experiment might rarely have sufficient number of replicates for such a resampling. The lack of enough replicates typically leads to making assumptions such as identical test statistic marginal null distribution for probes. However, such an assumption might not always be valid since probes might have different binding affinities leading to different distributional characteristics for test statistics. Another shortcoming of sliding window testing approaches is that the sliding window size is kept constant even though experimental design and empirical results suggest the suitability of a variable window size (Keleş et al., 2004). More recently, Kim et al. (2005) used a double regression methodology fitting a triangle shaped regression function across a moving window of fixed size. One of the apparent limitations of this approach is the enforcement of a triangular shape. Although the underlying peak structure is expected to have a bell shape, in practice one only observes randomly sampled regions from this structure. On the other hand, Li et al. (2005) proposed a model-based method that relies on the use of external array data for estimating the distribution of probe intensities and utilizes a hidden Markov model. More recently, Ji and Wong (2005) developed a two step approach for analyzing tiling array data. This approach utilizes a hierarchical empirical Bayes model for computing probe specific test statistics with a robust variance estimate. Then, these test statistics are used either to compute moving averages as in Keleş et al., 2004 or to build a hidden Markov model which treats them as outcomes of bound and unbound hidden states.

Although some of above approaches consider explicit modeling of the tiling array data, none utilizes the apparent recent advancements in the modeling of the gene expression array data. The fact that the first step of ChIP-chip data analysis constitute microarray analysis with spatially structured data motivated our work. We develop a modeling framework for ChIP-chip data by extending the hierarchical mixture model of Newton et al. (2004) for gene expression. This extension is motivated by the spatial structure in the ChIP-chip data and its resemblance to the way one thinks about *de nova* motif finding using sequence data. In Section 2, we provide further motivation for our model and develop the model and establish inference procedures in Section 3. Sections 4 and 5 are dedicated to comparisons of our approach to sliding window testing approaches and TileMap of Ji and Wong (2005) both on simulated data and data from transcription factor cMyc.

2. Motivation for a hierarchical mixture model

Spatial data structure. There is a fundamental difference between ChIP-chip experiments and mRNA experiments in terms of what they are trying to measure. In mRNA microarray experiments, each array element measures the abundance of one type of mRNA. In contrast, each element in a ChIP-chip experiment measures the abundance of a population of fragments of DNA. The DNA fragments in this population are of various lengths due to sonication to an average length. These fragments are hybridized to arrays containing short oligonucleotides (25mer or longer probes) after being fragmented further into smaller pieces. As a result of this process, the set of probes that map to a target fragment of the transcription factor on the genome are hybridized by the corresponding fragment in the IP-enriched DNA solution. Therefore, enrichment due to a specific binding site will not be detected by a single array element but will be detected by a group of array elements mapping in the vicinity of the binding site on the genome (Buck and Lieb, 2004 and Keleş et al., 2004). This process 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 *peaks*.

Varying size of the sheared DNA fragments and its effect on the peak size. In the experiments of Cawley et al. (2004), IP-enriched DNA is sheared into fragments of average length 1 kbs and authors use a test statistic based on a fixed window size of 1 kbs, i.e., combining information across

probes that fall within a window size of 1 kbs. This corresponds to an average of ~ 28 probes within a peak. We investigated the Cawley et al. (2004) ChIP-chip data for 13 quantitative PCR verified p53 bound regions on chromosomes 21 and 22 and observed the following peak size distribution: $P(w \in (0,5]) = 2/13$, $P(w \in (5,10]) = 1/13$, $P(w \in (10,15]) = 6/13$, $P(w \in (15,20]) = 4/13$ where w represents the number of probes in a peak. Quite strikingly, all of the 13 verified binding regions contained much smaller number of probes (≤ 20) than the theoretical prediction of ~ 28 probes. Using a peak size that is much larger than the truth might lead to an increase in the false negative rate since for peaks with small number of probes, the peak quality measure will be down-weighted by probes not contributing to the peak but are within the window considered by the testing approach. Additionally, a study by Keleş et al. (2004) predicted the size of the DNA fragments hybridizing to the array to have a a left skewed distribution, generating much shorter sheared DNA fragments. Although, the distribution of the peak size is going to depend on the specific lab/sonication process, it is desirable to develop a framework that incorporates this variation into the analysis of ChIP-chip data.

Limited number of replicates. Finally, one other challenge regarding ChIP-chip data is the small number of replicates, e.g., typically one to three replicates due to monetary reasons. In the context of gene expression analysis, many (Newton et al. (2004), Gottardo et al. (2005), have illustrated that model-based approaches provide a powerful alternative to simple testing approaches especially when there are few number of replicates. This further motivates us to consider model-based approaches for the analysis of ChIP-chip data.

3. Model and inference

Genome tiling arrays usually interrogate an entire genome at a certain bps resolution. For example, Affymetrix uses a resolution of 35 bps (distance between mid points of two probes) and oligonucleotides of size 25 bps as probes, whereas NimbleGen technology utilizes longer oligonucleotides. However, tiling probes do not encompass repeat regions or regions with known unreliable hybridization characteristics. As a result, there might be large gaps (>Q bps) in genomic distance going from one probe to the adjacent one according to physical genomic location. Due to this gap structure,

it is reasonable to treat each sequence of probes that are more than Q bps away from each other separately. This forces the data to be partitioned into *genomic regions* of different lengths and each genomic region is at least Q base pairs away from the closest genomic region. Q is driven by the average length of the sheared DNA fragment, e.g. 1000 bps, and we discuss it further in the case study section. Data structure as a result of such partitioning is an important component of our framework and allows us to carry out the statistical calculations easily. Next, we present this new data structure and our model in detail.

We assume to have N genomic regions where each region i, $i = 1, \cdots, N$, has L_i probes. Let $(Y_j(i), X_j(i))$, $j = 1 \cdots, L_i$ $i = 1, \cdots N$, denote preprocessed treatment and control data on j-th probe of i-th genomic region. Each $Y_j(i) = (Y_{j1}(i), \cdots, Y_{jn}(i))$ is a vector of n replicates whereas each $X_j(i) = (X_{j1}(i), \cdots, X_{jm}(i))$ is a vector of m replicates. Let $\mu_{j1}(i)$ and $\mu_{j2}(i)$ denote the latent mean levels of $X_j(i)$ and $Y_j(i)$. We assume that conditional on these expectations $Y_j(i)$ form an independent and identically distributed random sample from an observation component $p_Y(. \mid \mu_{j2}(i))$ and likewise $X_j(i)$ is an independent random sample from $p_X(. \mid \mu_{j1}(i))$. Here, we are treating the means $\mu_{j1}(i)$ and $\mu_{j2}(i)$ as unobserved random variables and allowing probe specific intensity distributions.

In order to simplify model fitting, we will assume that there can be at most one peak in each genomic region. In the context of *de novo* motif finding, this corresponds to assuming that each sequence has 0 or 1 motif and several heuristics exist for extending this assumption (Bailey and Elkan, 1995; Keleş et al., 2003). In our context, this assumption easily holds unless a genomic region spans a large genomic distance and this might be avoided by partitioning genomic regions accordingly as will be discussed in the case study section. As a consequence of this assumption each genomic region can be in of the following two states: (1) *Peak state:* there is exactly one peak; (2) *Non-peak state:* there are no peaks. This translates into following probe specific hypothesis:

$$\begin{aligned} \text{Peak:} \quad & \mu_{j1}(i) \leq \mu_{j2}(i), \quad \text{j-th probe of genomic region } i \text{ is part of a peak,} \\ \text{Non-Peak:} \quad & \mu_{j1}(i) = \mu_{j2}(i), \quad \text{j-th probe of genomic region } i \text{ is not part of a peak.} \end{aligned}$$

Let R_i denote an unobserved random variable representing whether or not the i-th genomic region

has a peak. Let $\mathcal{W} = \{w_1, \cdots, w_M\}$ be the ordered set of possible peak sizes, i.e., w_1 represents the smallest peak size (# of probes within a peak) and w_M represents the maximum peak size. Furthermore, let $Z_i(z)$, $z = 1, \cdots, L_i - w_1 + 1$, be a set of indicator variables representing the start site of the peak in genomic region i. Note that we have $\sum_{z=1}^{L_i - w_1 + 1} Z_i(z) \in [0, 1]$, $\forall i$. We will denote the end position of the peak in i-th genomic region by indicator variables $V_i(v)$. Given $Z_i(z) = 1$ for a z, $V_i(v)$ can only take on values $\{z + w_1 - 1, \cdots, z + w_M - 1\}$[1].

The next assumption deals with the distributional forms of the intensities conditional on the latent means. A gamma observation component with constant shape parameter, hence constant coefficient of variation, for a single condition is shown to be well suited for microarray data by Kendziorski et al. (2003) and Newton et al. (2004). We also adopt this model and consider expectations $(\mu_{j1}(i), \mu_{j2}(i))$ to be a random pair from an unknown bivariate distribution. Unlike two-sample microarray gene expression data where f is typically a discrete mixture over three hypotheses - over-expressed, under-expressed and not differentially expressed, there are only two hypotheses of interest corresponding to the peak and non-peak states for the ChIP-chip data. Hence, we can write the joint distribution of the latent mean levels for a single probe j in the genomic region i as a two component mixture model:

$$f(\mu_{j1}(i), \mu_{j2}(i)) = P(R_i = 0) f_0(\mu_{j1}(i), \mu_{j2}(i)) + P(R_i = 1) f_1(\mu_{j1}(i), \mu_{j2}(i)), \tag{1}$$

where the densities f_0 and f_1 describe fluctuations of the means within each states. We will denote the mixing proportion $P(R_i = 0)$ by p_0. As in Newton et al. (2004), we will relate the joint distribution f to a one dimensional base distribution π to generate the probe and state specific latent means. This ensures closed form representations for both f_0, f_1 and also for the marginal Gamma observation component distributions. This base distribution relates to the mixture distribution f as follows: $f_0(\mu_{j1}(i), \mu_{j2}(i)) = \pi(\mu_{j1}(i)) I(\mu_{j1}(i) = \mu_{j2}(i))$ and $f_1(\mu_{j1}(i), \mu_{j2}(i)) = 2\pi(\mu_{j1}(i))\pi(\mu_{j2}(i)) I(\mu_{j1}(i) \leq \mu_{j2}(i))$ and this relation is best understood by considering the underlying data generation process for the i-th genomic region. Next, we describe this process and discuss the model components.

1. Draw the state indicator R_i from the $Bernoulli(p_0)$ distribution. $R_i = 0$ implies that there is

[1] v should be such that $v \leq L_i$ is ensured.

no peak in i-th genomic region while $R_i = 1$ implies the occurrence of a peak.

2. **Genomic region without a peak.** If $R_i = 0$, then for $j = 1, \cdots, L_i$:

   (a) Draw $\theta_{j1}(i)$, $\theta_{j2}(i)$ from a two-parameter Gamma distribution, with shape parameter a_0 and scale parameter $1/(a_0 x_0)$, which seems to be appealing for the microarray data and makes computations analytically tractable.

   (b) Set $\mu_{j2}(i) = \mu_{j1}(i) = 1/\theta_{j1}(i)$ and ignore $\theta_{j2}(i)$. Note that both $\mu_{j2}(i)$ and $\mu_{j1}(i)$ have Inverse Gamma distribution (base distribution π) which is conjugate to the Gamma observation component.

   (c) Randomly draw the control and treatment observations from the corresponding Gamma distributions:
   $Y_{jk}(i) \mid \mu_{j1}(i) \sim \mathcal{G}a\left(a_2, \frac{\mu_{j1}(i)}{a_2}\right), k = 1 \cdots n;\ X_{jk}(i) \mid \mu_{j1}(i) \sim \mathcal{G}a\left(a_1, \frac{\mu_{j1}(i)}{a_1}\right), k = 1 \cdots m.$

3. **Genomic region with a peak.** If $R_i = 1$:

   (a) Draw a peak start site position z from $\{1, \cdots, L_i - w_1 + 1\}$ under the assumption that all sites are equally likely to be a peak start site position. Set $Z_i(z) = 1$ and $Z_i(l) = 0$, for $l \in \{1, \cdots, L_i\} \backslash \{z\}$. Let $\rho(w)$, $w \in \mathcal{W}$ denote the discrete peak size distribution. Draw a peak size w from $\rho(w)$ and set $V_i(z + w - 1) = 1$ and the rest of the components of $V_i(.)$ to 0.

   (b) Draw $\theta_{j1}(i)$, $\theta_{j2}(i)$ from the base distribution π.

   (c) Randomly draw control and treatment observations as follows:

   For $j = \{Z_i, \cdots, V_i\}$, $k = 1 \cdots n$:
   $Y_{jk}(i) \mid \mu_{j2}(i) \sim \mathcal{G}a\left(a_2, \frac{\mu_{j2}(i)}{a_2}\right)$, where $\mu_{j2}(i) = 1/\min\left(\theta_{j1}(i), \theta_{j2}(i)\right)$,
   For $j = \{1, \cdots, L_i\} \ / \ \{Z_i, \cdots, V_i\}$, $k = 1 \cdots n$:
   $Y_{jk}(i) \mid \mu_{j1}(i) \sim \mathcal{G}a\left(a_2, \frac{\mu_{j1}(i)}{a_2}\right)$, where $\mu_{j1}(i) = 1/\max\left(\theta_{j1}(i), \theta_{j2}(i)\right)$,
   $X_{jk}(i) \mid \mu_{j1}(i) \sim \mathcal{G}a\left(a_1, \frac{\mu_{j1}(i)}{a_1}\right)$, $k = 1 \cdots m$, where $\mu_{j1}(i) = 1/\max\left(\theta_{j1}(i), \theta_{j2}(i)\right)$.

The above model allows different localization rates for each probe through Gamma distributions with different latent means (2.(c) and 3.(c)) and variable peak sizes (3.(a)). The overall parameter set for this hierarchical model consists of the mixing proportion p_0, i.e., proportion of genomic regions without a peak, latent mean base distribution parameters a_0, x_0, and observation component parameters a_1 and a_2. We estimate these parameters by maximizing the marginal likelihood of the data using the EM algorithm (Dempster et al. (1977)). Various marginals and conditional distributions induced by this hierarchical mixture model are easily obtained adapting the derivations in Newton et al. (2004) by accommodating the fact that we only have two hypotheses for each probe and additional unobserved random variables Z and V. Next, we present the complete data likelihood, outline the EM algorithm steps and discuss the posterior probabilities for inference.

3.1 *Full data likelihood*

The full data likelihood is a product over N genomic regions, where each genomic region i contributes likelihood $\mathcal{L}_i$ over L_i probes, given by

$$\begin{aligned}\mathcal{L}_i = P(Y(i), X(i), R_i, Z_i, V_i \mid a_0, x_0, a_1, a_2) &= \Big\{P(R_i=0)P(Y(i), X(i) \mid R_i = 0)\Big\}^{I(R_i=0)} \\ \times\Bigg[P(R_i=1) \prod_{z=1}^{L_i-w_1+1} \prod_{v=z+w_1-1}^{\min(L_i, z+w_M-1)} &\Bigg\{P(Y(i), X(i) \mid R_i=1, Z_i(z)=1, V_i(v)=1) \\ &\frac{\rho(v-z+1)}{L_i-w_1+1}\Bigg\}^{I(Z_i(z)=1, V_i(v)=1)}\Bigg]^{I(R_i=1)},\end{aligned}$$

where we substitute in $P(Z_i = v \mid R_i = 1) = 1/(L_i - w_1 + 1)$ and $P(V_i(v) = 1 \mid Z_i(z) = 1, R_i = 1) = \rho(v - z + 1)$. Furthermore, let $h(u_1, \cdots, u_k)$ represent a compound Gamma distribution (Hogg et al. (2005), p. 191) with density given in equation (A.1) of the Appendix. Then, we have

$$P(Y(i), X(i) \mid R_i = 0) = \prod_{j=1}^{L_i} P(Y_j(i), X_j(i) \mid R_i = 0) = \prod_{j=1}^{L_i} h(Y_j(i), X_j(i)).$$

Let $T_i^{z,v}$ denote all the probes outside the peak when the peak start and end positions are z and v, respectively. Then,

$$\begin{aligned}P(Y(i), X(i) \mid R_i = 1, Z_i(z) = 1, V_i = v) &= \prod_{j \in T_i^{z,v}} P(Y_j(i), X_j(i) \mid R_i = 1, Z_i(z) = 1, V_i(v) = 1) \\ &\times \prod_{j=z}^{v} P(Y_j(i), X_j(i) \mid R_i = 1, Z_i(z) = 1, V_i(v) = 1)\end{aligned}$$

$$= \prod_{j \in T_i^{z,v}} h(Y_j(i), X_j(i)) \prod_{j=z}^{v} 2h(Y_j(i))h(X_j(i))P(B > b),$$

where B is a Beta-distributed random variable with shapes $(a_0 + ma_1, a_0 + na_2)$ and

$$b = \frac{a_0 x_0 + a_1 \sum_{k=1}^{m} X_{jk}(i)}{2a_0 x_0 + a_1 \sum_{k=1}^{m} X_{jk}(i) + a_2 \sum_{k=1}^{n} Y_{jk}(i)}.$$

Here, probe specific conditional joint distributions $P(Y_j(i), X_j(i) \mid R_i = 0)$ and $P(Y_j(i), X_j(i) \mid R_i = 1, Z_i(z) = 1, V_i(v) = 1)$ are derived integrating over the latent means μ_{j1} and μ_{j2} as in Newton et al. (2004).

3.2 *Genomic region specific inference*

Fitting the hierarchical model readily provides several useful posterior probabilities that are by-products of the E-step of the EM algorithm. We use

$$\eta_i = P(R_i = 1 \mid Y(i), X(i)) = (1 - p_0)P(Y(i), X(i) \mid R_i = 1)/\, P(Y(i), X(i)), \quad \text{where} \tag{2}$$

$$P(Y(i), X(i) \mid R_i = 1) = \tag{3}$$
$$\sum_{z=1}^{L_i - w_1 + 1} \sum_{v=z+w_1-1}^{\min(z+w_M+1, L_i)} P(Y(i), X(i) \mid R_i = 1, Z_i(z) = 1, V_i(v) = 1) \frac{\rho(v - z + 1)}{(L_i - w_1 + 1)},$$
$$P(Y(i), X(i)) = p_0 P(Y(i), X(i) \mid R_i = 0) + (1 - p_0)P(Y(i), X(i) \mid R_i = 1),$$

for inferring the genomic regions that have a peak. Moreover, the following posterior probabilities are utilized to infer the most likely start and end positions of a peak.

$$\zeta_i(z, v) = P(Z_i(z) = 1, V_i(v) = 1, R_i = 1 \mid Y(i), X(i)) = \frac{P(Y(i), X(i) \mid R_i = 1, Z_i(z) = 1, V_i(v) = 1) \frac{\rho(v-z+1)}{L_i - w_1 + 1} (1 - p_0)}{P(Y(i), X(i))}. \tag{4}$$

In this modeling framework, M-step of the EM-algorithm does not have a closed form solution, hence `R` optimization function `optim()` is used. We have observed good convergence properties in this optimization step with both real and simulated data starting from different starting values. In order to provide starting values for the parameters, we evaluate the observed data likelihood for a large set of values and start the EM algorithm with the set giving the highest initial observed data likelihood. Due to space limitations, we do not discuss the estimation of the peak size distribution but refer the reader to an online technical report Keleş (2005).

4. Simulations

We carried out a simulation study to assess the proposed method and to compare it with the available approaches. The following parameters are varied in each simulation: (1) data generating mechanism, (2) number of replicates n and m, and (3) structure of the peaks, i.e., fixed peak size versus variable peak size. We focus on the following two data generating mechanisms.

- **Simulation model I.** The first simulation model follows the hierarchical Gamma mixture model with parameters a_0, x_0, a_1, a_2 as presented here. Instead of setting these parameters to arbitrary values in simulations, we set them to values derived from the fit of cMyc ChIP-chip data (Section 5) as $a_0 = 297$, $x_0 = 8$, $a_1 = 962$, and $a_2 = 2611$.

- **Simulation model II.** The second simulation model intends to evaluate performance when the Gamma observation component assumption is violated. The data are generated according to the following model for k-th replicate of the j-th probe in i-th genomic region:

$$\begin{aligned} X_{jk}(i) &= \theta_{j1}(k) + \epsilon_1(i), \\ Y_{jk}(i) &= \begin{cases} \theta_{j2}(k) + \epsilon_2(i), & \text{if } j\text{-th probe is within a peak,} \\ \theta_{j1}(k) + \epsilon_1(i), & \text{o.w.} \end{cases} \end{aligned}$$

  Here, $\theta_{j1}(k)$ and $\theta_{j2}(k)$ are two base means sampled from the set $\Theta = (8, 9, 10, 11, 12, 13, 14, 15)$ with probabilities $p_1 = (0.2, 0.6, 0.1, 0.05, 0.03, 0.01, 0.007, 0.003)$ and $p_2 = (0.05, 0.1, 0.05, 0.1, 0.2, 0.2, 0.2, 0.2)$, respectively. Here, p_2 puts higher probability on large values of the base mean. $\epsilon_1(i)$ and $\epsilon_2(i)$ are log-normal random variables with genomic region specific means μ_1^*, μ_2^* and variances σ_1^2, σ_2^2. For each genomic region, both means (μ_1^*, μ_2^*) are sampled uniformly between [0.1, 0.5] and σ_1^2 and σ_2^2 are set so that the constant coefficient of variation assumption is satisfied. These parameters were set as a result of a search across a wide range for generating the most disadvantageous (in terms of sensitivity, specificity, and FDR) case for our hierarchical model.

We compared the hierarchical Gamma mixture model (HGMM) with the scan statistic (SS) approach of Keleş et al. (2004), sliding window Wilcoxon rank sum test (WRS) approach of Cawley et al. (2004) and TileMap of Ji and Wong (2005). Since TileMap implements both a moving average

(TileMap-MA) and a HMM (TileMap-HMM) method, whenever possible we used both.

Results of simulation I, where the parametric model assumptions are met, are summarized in Figure 1. Here, we show the results for HGMM, SS, WRS and HMM of TileMap approaches where we apply false discovery rate (FDR) controlling procedures at level $\alpha = 0.05$, overall conclusions stayed the same when considering other FDR rates (data not shown). We used the *direct posterior probability* approach of Newton et al. (2004) to control FDR at the desired threshold using the posterior probabilities from the fitted hierarchical mixture model and posterior probabilities of the TileMap-HMM and the procedure of Benjamini and Hochberg (1995) for the sliding window testing approaches. TileMap-MA is not included in this comparison because setting the overall false discovery rate to reasonable values such as 0.05 and 0.1 did not provide any identified regions in a vast majority of the simulations. We obtained empirical estimates of sensitivity, specificity, and FDR for all the three approaches using $B = 100$ simulated datasets. Here, we note that for all approaches except for HGMM, FDR is controlled at the probe level and for HGMM it is controlled at the peak level since we are able to obtain region specific posterior probabilities. Sensitivity and specificity are computed at the peak level for all the approaches. For a peak to be declared a true positive for SS, WRS and TileMap, we required that at least one of the probes in the peak overlapped the true set of bound probes.

[Figure 1 about here.]

[Figure 2 about here.]

Here, only the mean estimates of the sensitivity, specificity, and false discovery rate are reported since the corresponding standard errors were small and comparable among the three methods. The first striking result is the poor performance of SS and WRS methods with small sample sizes. This result holds for other parameter sets not reported here (extensive simulations are available in Keleş, 2005). It is particularly interesting to note that when the observation component parameters are set to values obtained from real data, sliding window approaches have a maximum sensitivity of 0.38 with a sample size of 2 whereas HGMM has sensitivity and specificity values equal to 0.9818

and 0.983, respectively. Since the SS approach is based on constructing two-sample t-statistics and averaging these across a peak, it has poor performance with very small sample sizes, as expected. This was also pointed out by Ji and Wong (2005). Furthermore, with a small sample size of 2, the SS approach has elevated actual FDR levels since in this setting it identifies a few (1 or 2) bound regions a large portion of which are false positives. Another interesting observation is that with moderate sample sizes and variable peak structures, SS performs better than WRS in terms of sensitivity while achieving the targeted FDR rate. This is due to SS's capability to optimally choose a data-dependent peak size by cross-validation. Under this setting, although TileMap-HMM is performing better than SS and WRS, it has low sensitivity due to the conservative actual FDR levels. There might be two reasons for this. First, since the FDR is being controlled at the probe level rather than at the peak level, the FDR controlling procedure is dealing with many probe specific tests than that are relevant. Second, observation component distributions for TileMap are estimated by a method called unbalanced mixture subtraction which might be improved with a better tuning. In these simulations, estimates of a_0, a_1, a_2, and x_0 averaged over the 100 simulations were very close to the true values with little standard error. This indicates that optimization in the M-step did not have any adverse effects on the convergence of the parameter estimates.

In simulation setting II, we investigate the operating characteristics of our method under model misspecification. In this setting, the direct posterior probability approach for controlling the FDR does not provide correct inference since the data generating model is misspecified. Therefore, in order to make the comparisons meaningful we adjusted the nominal α levels of each procedure so that the actual false discovery rate of the approaches are set as desired. Panels (a), (b), (c) and (d) of Figure 2 summarize the results for HGMM, SS, WRS, and TileMap-HMM. For TileMap-HMM, we varied the threshold on the maximum posterior probability of bound regions to achieve various false discovery rates. We couldn't carry a similar scheme for TileMap-MA since it reported the same mean or minimum local FDR for all the detected regions. We observe that all the methods except SS have good operating characteristics in this scenario where the observations from different probes are independent but not identical. Overall, HGMM and HMM has the best performance with HGMM

slightly better in terms of specificity at varying peak sizes. Although this comparison is useful in this simulation setting, in practice we can't set the actual FDR. Therefore, we also compared the methods where each implemented a particular way of choosing a threshold. For TileMap-HMM and MA we used a 0.5 cut-off for maximum posterior probability and minimum local FDR of the regions, which are the defaults suggested by Ji and Wong (2005). For HGMM, we implemented the following general scheme of choosing a cut-off for the posterior probability of regions when the model assumptions are violated. We permuted the genomic regions $B = 100$ times and for each permuted dataset we computed the posterior probabilities using the estimated model parameters from the original fit. Then, by using the 95% quantile of the maximum of the posterior probabilities from each permuted dataset, we were able to obtain a threshold that provides less than one false positive. In fact, Cawley et al. (2004) also suggested a similar permutation based approach for choosing a threshold for the p-values obtained from the Wilcoxon Rank Sum test. Panel (e) of Figure 2 displays sensitivity and specificity for HGMM, WRS, HMM and MA under this setting with a variable peak size structure. It seems like cut-off determination based on permutation helps to get a more accurate inference from HGMM but provides a poorer sensitivity for WRS. HGMM identifies more regions with better accuracy than HMM and MA of TileMap though the performance of the latter might be improved with a different cut-off rule.

5. Case study: ChIP-chip data of cMyc

5.1 *cMyc ChIP-chip data*

Transcription factor cMyc is a growth-regulating transcription factor and plays an important role in apoptosis. It functions by binding to cMyc DNA recognition sequences and regulates transcription of growth-regulatory genes. Cawley et al. (2004) generated ChIP-chip data for cMyc on chromosomes 21 and 22 distributed across three chips with 6 replicates each. We used data from one of the chips encompassing 2/3 of chromosome 21. There are two types of controls available for this factor. We refer to control GST, control Input and the treatment conditions as C1, C2 and IP, respectively. Our current framework only allows two sample comparisons. However, to make comparisons with TileMap we used both controls separately and combined results from the two

analysis.

5.2 *Model diagnostics*

We checked whether constant coefficient of variation and Gamma observation component assumptions are reasonable using the diagnostic tools provided by Newton and Kendziorsky (2003). Figure 3 plots the coefficient of variation as a function of mean intensity for each probe on the chip of interest. Solid green line is the lowess fit to these data. This flat line indicates that the sample coefficient of variation does not have a systematic relation with the sample mean intensity. The data points above the dashed red line are points for which this assumption might be violated. For the control sample, these constitute only 0.35% of the data whereas for the IP-enriched sample they represent only 0.34% of the data. Having examined these points, we see that they correspond to probes with outlier data points, e.g., one observation is dramatically different than the rest of the 12 observations across 6 control and 6 IP-enriched intensities, but they don't seem to be located in the vicinity of potential peaks.

Furthermore, Figure 4 displays the quantiles of the fitted Gamma distribution versus empirical quantiles of the control hybridization data across probes with various mean levels. These quantile plots lie roughly on the 45 degree line indicating compliance with the Gamma observation component assumption. Similar quantile-quantile plots are also observed for the IP-enriched case (more plots involving other factors are available in Keleş, 2005).

[Figure 3 about here.]

[Figure 4 about here.]

5.3 *Implementation details and results*

We did not have an agarose gel image for these experiments to estimate the peak size distribution; therefore we used the cross-validated peak size of 10 identified by Keleş et al. (2004). Genomic regions were obtained by partitioning whenever there was a gap of $Q = 1000$ bps (average sheared DNA fragment size) between the physical genomic locations of two adjacent probes. Furthermore, genomic regions larger than 3000 bps were further partitioned. This generated a total of 7543

genomic regions. Below we present the results corresponding to this partition; however changing the partition parameters did not effect the results in this case study.

We compared HGMM to TileMap-HMM on the gold standard dataset for the transcription factor cMyc used by Ji and Wong (2005). These data included 18 arrays from chip A spanning 2/3 of chromosome 21. The gold standard dataset was obtained by applying HMM and gtrans (Affymetrix default software implementing WRS) using all the 6 replicates and then selecting the overlapping probes among the top 0.5% of the probes identified by the two methods. These probes were further grouped into binding regions generating 73 gold standard regions which are then treated as the true set of bound regions. Following the comparison of Ji and Wong (2005), we based our comparisons on the number of correctly identified regions among the top 100 regions declared as bound. Figure 5 displays the number of correctly identified gold standard regions as a function of total number of identified regions for various scenarios. Here, *s2r2* and *s3r2* refer to two ($IP > C1$) and three sample ($IP > C1$ and $IP > C2$) comparisons with 2 replicates and *s3r4* refers to three sample comparisons with 4 replicates. First, we see that under the composite hypothesis $s3r4$ scenario, HGMM is performing comparable to the other approaches which were utilized to build the gold standard regions. Here, although TileMap-HMM has a direct way of dealing with more than two sample comparisons, HGMM which combines two two-sample comparisons by replacing posterior probability of nonpeak for each region by the maximum of the two posterior probabilities from the two comparisons performs equally well. Interestingly, for the $s2r2$ case, HGMM starts to outperform the other approaches when the total number of rejections exceed 70. The reason why we don't see a uniform behavior over the range of 100 regions is that HGMM gives a different ordering of the peak regions with respect to their strength. Investigating the top ranking regions of HGMM, we observed that these are indeed identifiable when all the samples are utilized. When the composite hypothesis $IP > C1$ and $IP > C2$ is considered at a sample size of 2, TileMap-HMM is the best since it directly utilizes a test statistic using all the information.

[Figure 5 about here.]

Currently, a common practice with ChIP-chip experiments is to first perform genome-wide ChIP-

chip experiments with a sample size of one and then to carry out replicate experiments for the regions identified from the initial analysis. Additionally, when studying a new transcription factor, initial small scale experiments with different antibodies are typically required. Since both gtrans and HGMM are applicable with a sample size of one, we compared their reproducibility using the cMyc ChIP-chip data. We first identified peaks using gtrans and HGMM with all the six replicates and kept their highest scoring 100 peaks (referred to as full-analysis). Then, we carried out the analysis of each replicate separately. Panel (a) of Figure 6 displays the box-plots of the number of full-analysis peaks that are also identified by the six one-replicate analyses. Each white and gray box-plot correspond to comparisons at the levels of $10, 20, \cdots, 90, 100$ top scoring peaks for gtrans and HGMM, respectively. These results are for the control $C1$ and we obtained similar results for the other control and when we varied the number of peaks identified from the full-analysis. We observed that, for all the cutoffs on the number of peaks, HGMM identifies a larger proportion of the full-analysis peaks. One of the one-replicate analysis was able to identify all the 100 peaks of the full-analysis, this is reflected by the wide range of the whiskers of the box-plots in panel (a). We regenerated the same boxplots based on the remaining five one-replicate analyses in panel (b) of Figure 6 and observed that overall HGMM not only recovers more of the peaks of the full-analysis but also provides much less variability among different analyses.

[Figure 6 about here.]

6. Discussion

Although genome-wide localization data of transcription factors are becoming increasingly available, the number of replicates available is typically small due to the high cost of genome tiling arrays. None of the methods to date, except the recently developed TileMap of Ji and Wong (2005), present an explicit statistical modeling approach that allows information sharing across probes to accommodate small sample sizes. We presented a model-based approach that aims to accommodate the small sample size as well as other characteristics of ChIP-chip data. Our hierarchical model enables information sharing across probes, incorporates the inherent spatial structure of these data, and allows detection of variable size peaks. The current framework utilizes hierarchical Gamma

mixture model of Newton et al. (2004) for modeling at the probe level, however it is possible to use other statistical models such as that of Gottardo et al. (2005) at the probe level. A small simulation study illustrates that, when the model assumptions are met, such a model drastically improves the sensitivity when we have little information per probe. A case study of the transcription factor cMyc provided further support that this approach has the potential to detect more bound regions and provide a higher rate of reproducibility at varying sample sizes. This is especially important since biologists might often be interested in initially performing a genome-wide ChIP-chip experiment with a sample size of one. After analyzing the initial data, they might design the next generation tiling array to only incorporate candidate bound regions from the initial analysis and carry out replicate ChIP-chip experiments with the new array. Therefore, it is important to be able to create as many reliable candidates as possible with as small as one sample size.

Our current implementation relies on partitioning the genomic data into non-overlapping regions. This provides a nice framework where the overall FDR of the identified regions can be controlled accurately when the model assumptions for regarding the probe intensities are met. Typically, genomic gaps due to repeat masking present such partitions. Additionally, genomic annotations (exon-intron boundaries, CpG island markers) or pairwise alignments between related species might also be utilized. Another important issue in our current implementation is the one peak per genomic region assumption. It is possible to extend this by moving to a larger model; however the computational complexity of model fitting increases drastically. Therefore, heuristic approaches such as iterative identification of the peaks in a region while masking already identified peaks (Bailey and Elkan, 1995) or employing the zero or one peak model with varying partitions (Keleş et al., 2003) are reasonable alternatives. Our investigations with real data showed that the results are quite robust against the violation of this assumption (Isogai et al. (2005)). In particular, if a genomic region contains more than one peak, this region is almost certainly identified as bound and the largest posterior probability is observed at the peak location with the highest signal. Other potential peaks within a bound region can be revealed by a likelihood ratio test using the estimated model parameters and sliding over the genomic region. One important issue not addressed by our current

framework is the apparent correlation between intensities of nearby probes, both in the bound and unbound regions. Presumably, incorporation of such a correlation structure might further improve our framework. The proposed methodology is implemented in an `R` package and is available from the author's website at `www.stat.wisc.edu/~keles/software`.

Acknowledgements

This research has been supported by a WARF grant from UW Madison. The author thanks Yoh Isogai, Ron Stewart, and James Thomson for general discussions on ChIP-chip data, Hongkai Ji for providing gold standard datasets and acknowledges comments from Yoh Isogai and Jason Lieb regarding the use of agarose images and from Christina Kendziorski on the general hierarchical mixture model. Furthermore, the author thanks Deniz D. Yavuz for a careful reading of the manuscript, the associate editor and a referee for their constructive comments that led to a much improved version of the manuscript.

References

Bailey, T. L. and Elkan, C. (1995). Unsupervised learning of multiple motifs in biopolymers using EM. *Machine Learning* **21**, 51–80.

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.

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.

Cawley, S., Bekiranov, S., Ng, H., Kapranov, P., Sekinger, E., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A., Wheeler, R., Wong, B. ., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K. and Gingeras, T. (2004). Unbiased mapping of transcription factor binding sites along h uman chromosomes 21 and 22 points to widespread regulation of non-coding RNAs. *Cell* **116**, 499–511.

Dempster, A. P., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *JRSSB* **39**, 1–38.

Gottardo, R., Raftery, A. E., Yeung, K. Y. and Bumgarner, R. (2005). Bayesian robust inference for differential gene expression in cdna microarrays with multiple samples. *Biometrics* In press.

Hogg, R. V., McKean, J. W. and Craig, A. T. (2005). *introduction to Mathematical Statistics*. Prentice-Hall.

Isogai, Y., Takada, S., Tjian, R. and Keleş, S. (2005). Genome-wide requirement of TRF1/BRF complex in drosophila Pol III transcription. Technical report, Departments of Statistics and Biostatistics & Medical Informatics. In preparation.

Ji, H. and Wong, W. H. (2005). TileMap: Create chromosomal map of tiling array hybridizations. *Bioinformatics* **21**, 3629–3636.

Keleş, S. (2005). Mixture modeling for genome-wide localization of transcription factors. Technical report, University of Wisconsin, Madison. `www.stat.wisc.edu/~keles/ggcc.v1.pdf`, July 2005.

Keleş, S., van der Laan, M. J., Dudoit, S. and Cawley, S. E. (2004). Multiple testing methods for ChIP-Chip high density oligonucleotide array data. Technical report, UC Berkeley, Division of Biostatistics. `http://www.bepress.com/ucbbiostat/paper147/`. To appear in the *Journal of Computational Biology*.

Keleş, S., van der Laan, M. J., Dudoit, S., Xing, B. and Eisen, M. B. (2003). Supervised detection of regulatory motifs in DNA sequences. *Statistical Applications in Genetics and Molecular Biology* **2**. Article 5.

Kendziorski, C. M., Newton, M. A., Lan, H. and Gould, M. N. (2003). On parametric empirical Bayes methods for compairing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**, 3899–3914.

Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M., Richmond, T., Wu, Y., Green, R. D. and Ren, B. (2005). A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880.

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**, i244–i282.

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

Newton, M. A. and Kendziorsky, C. M. (2003). Parametric empirical Bayes methods for microarrays. In Parmigiani, G., Garrett, E. S., Irizarry, R. and Zeger, S. L., editors, *The analysis of gene expression data: methods and software*. Springer Verlag, New York.

Pollard, K. and van der Laan, M. (2002). Resampling-based multiple testing: Asymptotic control of type i error and applications to gene expression data. *Journal of Statistical Planning and Inference* **125**, 85–100.

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

Weinmann, A. S., Yan, P., Oberley, M., Huang, T. and Farnham, P. (2001). Use of chromatin immunoprecipitation to clone novel E2F target promoters. *Molecular and Cellular Biology* **21**, 6820–6832.

Appendix A

Compound Gamma distribution (Hogg et al. 2005)

Let θ be a Gamma random variable with scale parameter a_0 and shape parameter $1/(a_0 x_0)$. Given θ, let $U_1, \cdots, U_k$ be conditionally independent Gamma-distributed variables with $U_i \sim \mathcal{G}\left(a_i, \frac{1}{a_i \theta}\right)$ with $1/\theta$ as the common conditional mean of the U_is. The compound Gamma distribution for the joint distribution of U_is is known as:

$$h(u_1, \cdots, u_k) = \frac{x_0 \Gamma(\sum_{i=1}^k a_i)}{\left(\sum_{i=0}^k a_0 u_i\right)^{\sum_{i=1}^k a_i}} \prod_{i=0}^k \left[\frac{a_i^{a_i} u_i^{a_i - 1}}{\Gamma(a_i)}\right]. \tag{A.1}$$

List of Figures

Figure 1. *Simulation model I.* Empirical estimates of sensitivity (open circle), specificity (filled circle), and FDR (filled triangle) are plotted over 100 simulated datasets. Sample sizes of $n = 2$ and $n = 6$ and fixed and variable peak sizes are considered. Nominal α is set to 0.05.

Figure 2. *Simulation model II.* Empirical estimates of sensitivity (open circle) and specificity (closed circle) are plotted over 100 simulated datasets. Fixed and variable peak sizes are considered at a sample size of $n = 2$. In panels (a)-(b) and (c)-(d), nominal α is set such that the actual α levels of the methods are 0.05 and 0.1, respectively. Panel (e) implements permutation based thresholding rule for HGMM and WRS, and maximum posterior probability and minimum local FDR based thresholds for TileMap-HMM and TileMap-MA.

Figure 3. *Hierarchical Gamma mixture model diagnostic of cMyc ChIP-chip data: checking the constant coefficient of variation assumption.* **Left panels:** Plotted is the coefficient of variation versus sample mean of the intensities across 6 replicates for each probe on the chip. Solid green line is the lowess fit to these data. Data points above the dashed red line are points for which this assumption might be violated. For the control sample, these constitute only 0.35% of the data whereas for the IP-enriched sample they represent only 0.34% of the data. Having examined these points, we see that they correspond to probes with outlier data points, e.g., one observation is dramatically different than the rest of the 12 observations across 6 control and 6 IP-enriched intensities, but they don't exhibit peak characteristics. **Right panels:** Zooming into low mean region of the plots on the left panel.

Figure 4. *Hierarchical Gamma mixture model diagnostic for control hybridizations of cMyc ChIP-chip data: Gamma quantile-quantile plot.* For various mean levels, quantiles of the fitted Gamma distribution versus empirical quantiles of the control hybridization data across ~ 600 probes with that mean level are plotted.

Figure 5. *Comparison of HGMM with the HMM of TileMap and gtrans. s3r2* and *s3r4* refer to three sample comparisons with sample sizes of 2 and 4, respectively. *s2r2* is two-sample comparison with a sample size of 2. Gold standard dataset is based on TileMap-HMM and gtrans applications

based on the whole dataset.

Figure 6. *Comparison of HGMM (gray) and WRS (white) based on one-replicate analysis.* For each method, top 100 peaks are identified using all the replicates. Box-plots at each top scoring peak number of $10, \cdots, 100$ provide the number of full-analysis peaks identified by the one-replicate analysis. (a) Box-plots from six one-replicate analyses. (b) Box-plots from five one-replicate analyses.

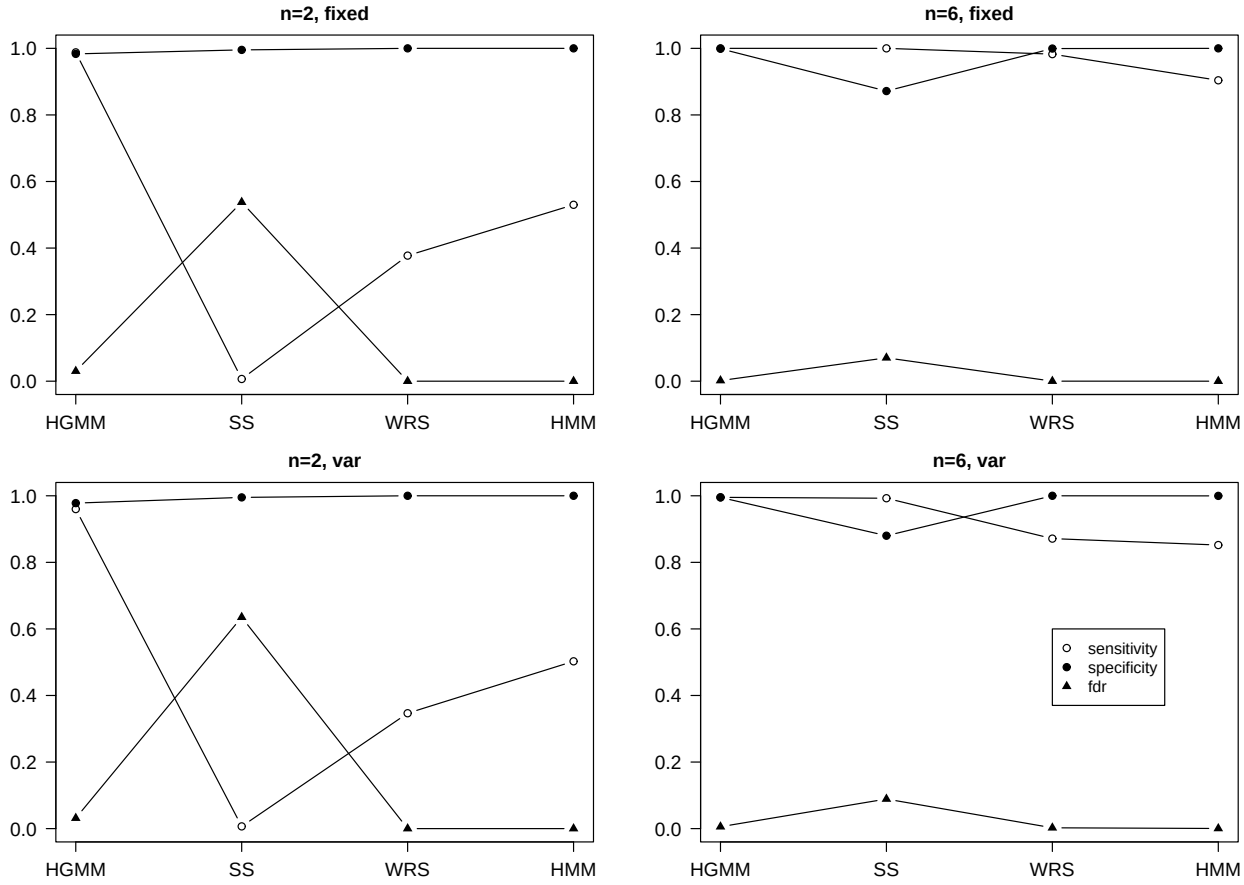
n=2, fixed
n=6, fixed
n=2, var
n=6, var
1.0
0.8
0.6
0.4
0.2
0.0
HGMM
SS
WRS
HMM
sensitivity
specificity
fdr

Figure 1.

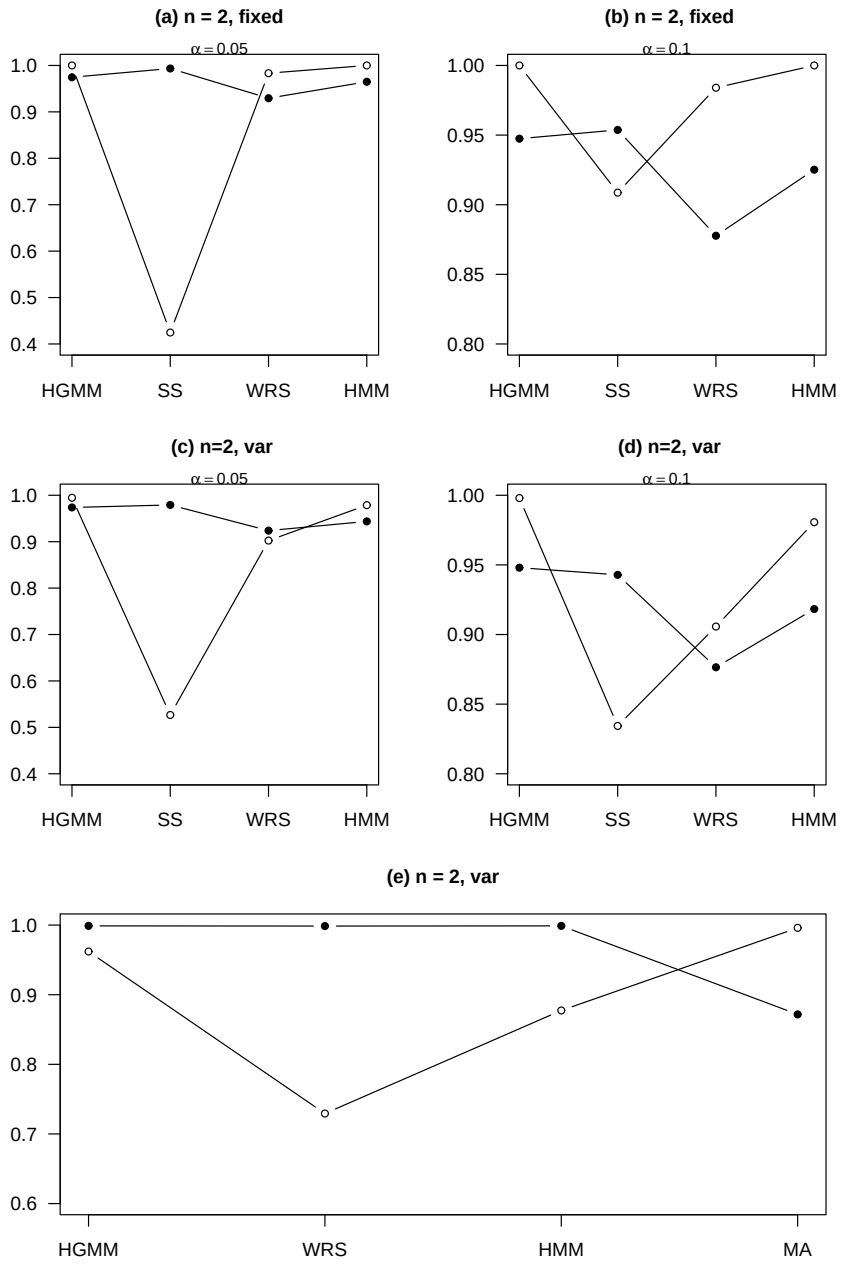
(a) n = 2, fixed
α = 0.05
1.0
0.9
0.8
0.7
0.6
0.5
0.4
HGMM
SS
WRS
HMM
(b) n = 2, fixed
α = 0.1
1.00
0.95
0.90
0.85
0.80
HGMM
SS
WRS
HMM
(c) n=2, var
α = 0.05
1.0
0.9
0.8
0.7
0.6
0.5
0.4
HGMM
SS
WRS
HMM
(d) n=2, var
α = 0.1
1.00
0.95
0.90
0.85
0.80
HGMM
SS
WRS
HMM
(e) n = 2, var
1.0
0.9
0.8
0.7
0.6
HGMM
WRS
HMM
MA

Figure 2.

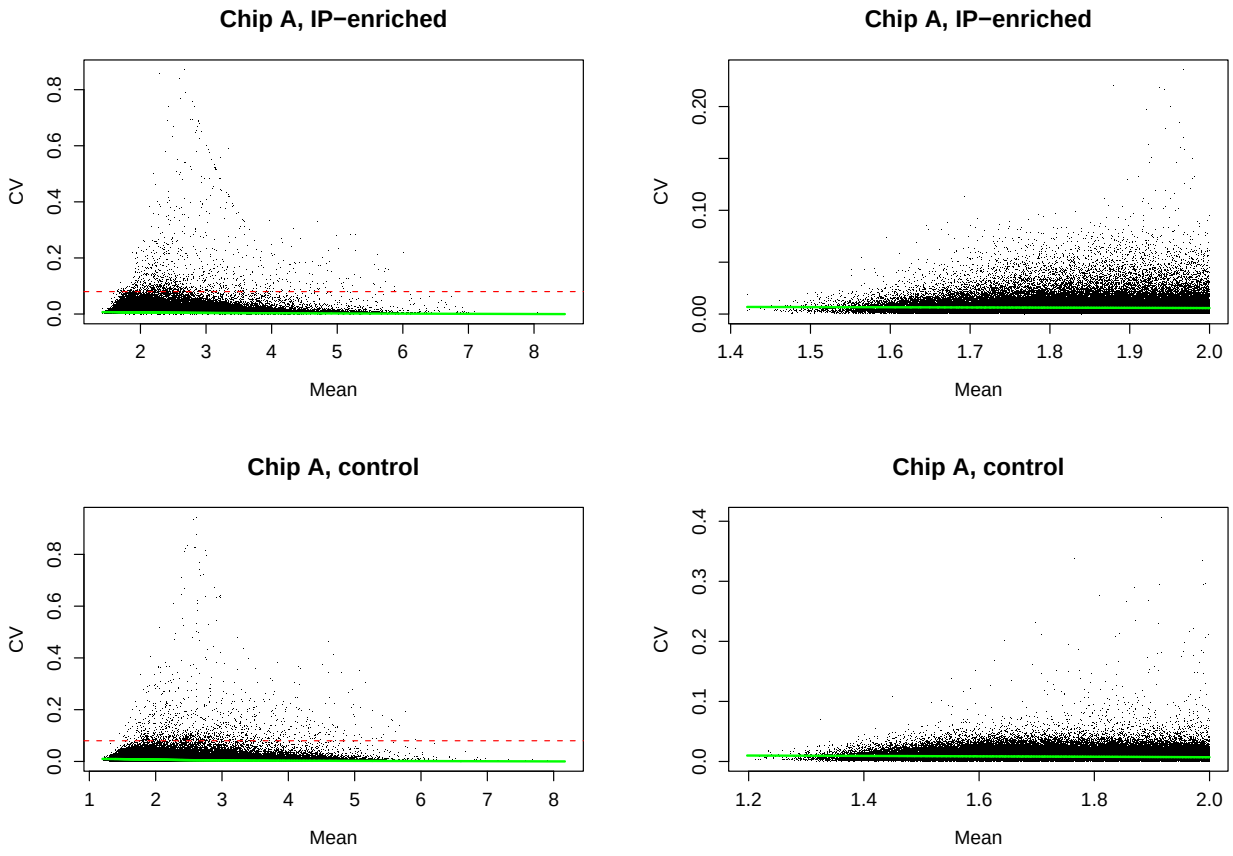
Chip A, IP−enriched
CV
Mean
Chip A, IP−enriched
CV
Mean
Chip A, control
CV
Mean
Chip A, control
CV
Mean

Figure 3.

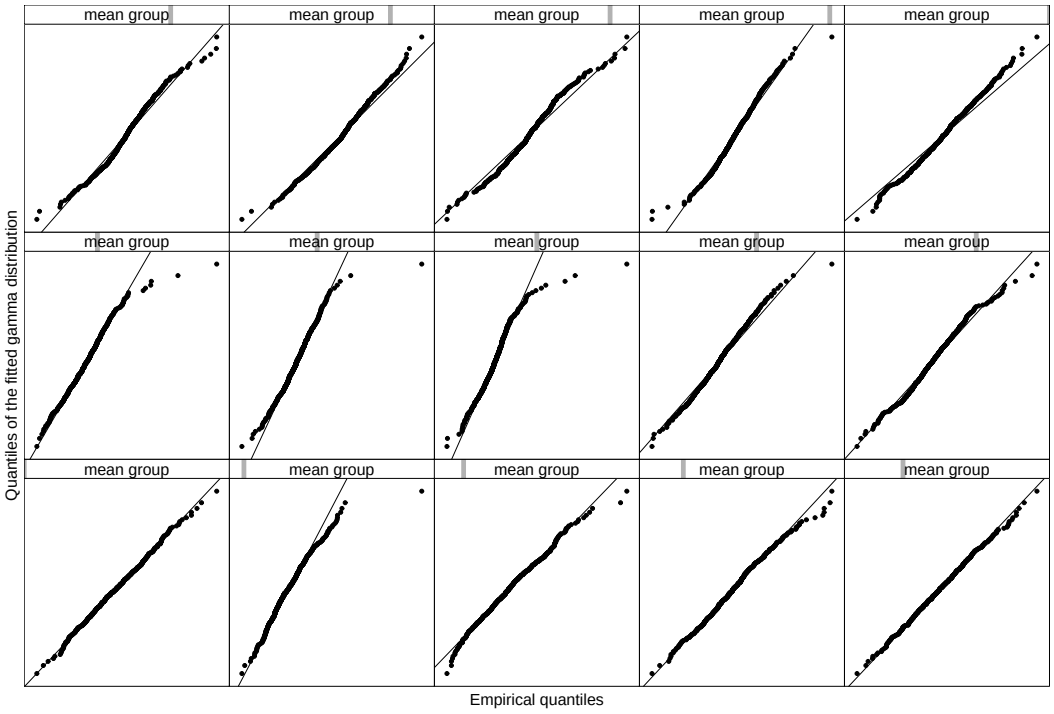
mean group
Quantiles of the fitted gamma distribution
Empirical quantiles

Figure 4.

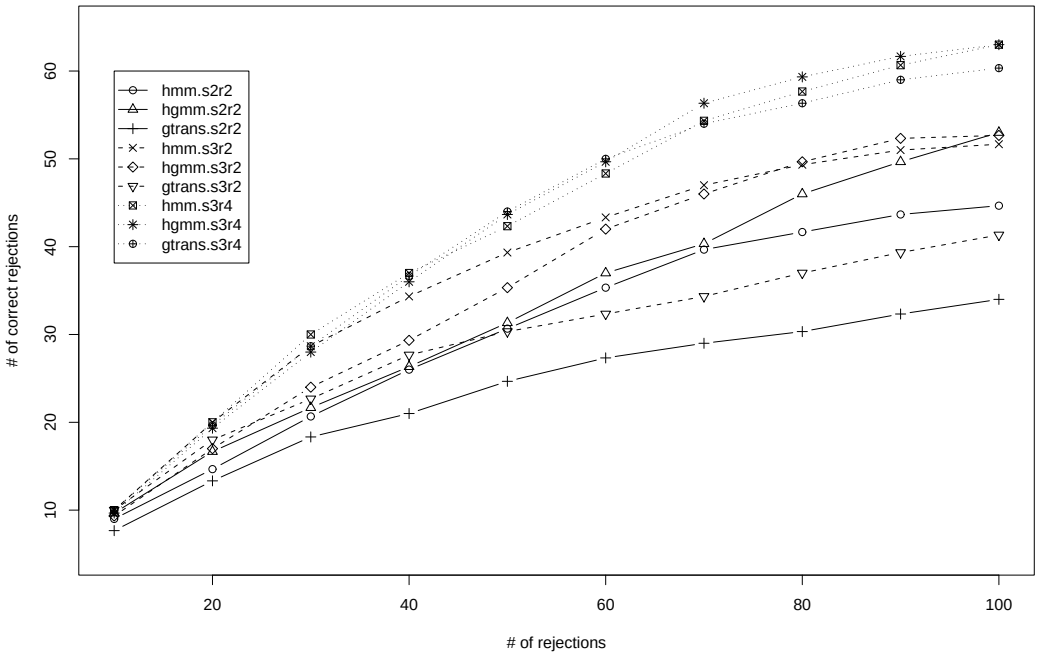
hmm.s2r2
hgmm.s2r2
gtrans.s2r2
hmm.s3r2
hgmm.s3r2
gtrans.s3r2
hmm.s3r4
hgmm.s3r4
gtrans.s3r4
# of correct rejections
10
20
30
40
50
60
20
40
60
80
100
# of rejections

Figure 5.

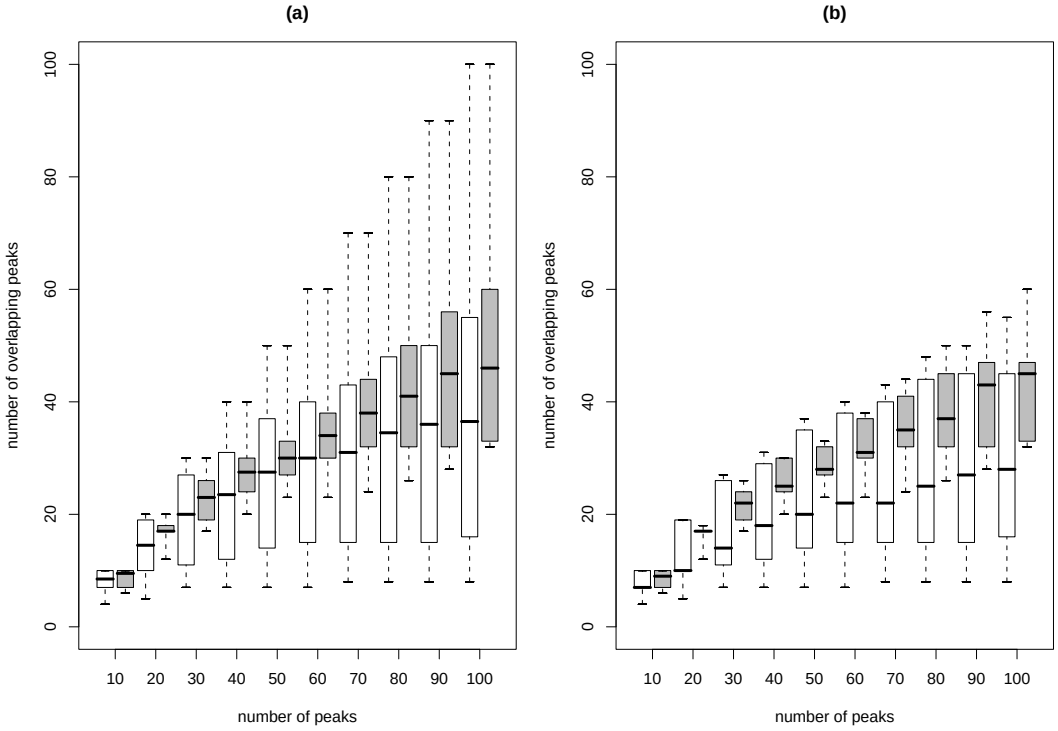
(a)
(b)
number of overlapping peaks
number of peaks
0
20
40
60
80
100
10
20
30
40
50
60
70
80
90
100

Figure 6.