A Estimating model parameters

This section presents algorithms for estimating the model parameters $\mathbf{\Sigma}$, $\mathbf{m}$, $\mathbf{\Psi}$, π and $\mathbf{q}$.

A.1 Estimating $\mathbf{\Sigma}$

If a dataset has replicates, the variances $\mathbf{\Sigma}$ are estimated using the replicate samples in that dataset. If a dataset does not have replicate samples, then the probe variances cannot be estimated from that dataset alone. In this scenario, we use information from other datasets to assist the variance estimation. Details for estimating $\mathbf{\Sigma}$ are described below.

1. If dataset d has replicates and the replicate samples are paired (e.g., in Agilent arrays, IP and control intensities are obtained from two channels Cy5 and Cy3 of the same hybridization), then the probe variances σ_{id}^2 are estimated as follows. Let X_{idjk} be the normalized and log2 transformed probe intensity of probe i in the k-th replicate under the condition j ($j = 1$: IP; $j = 0$: control) of dataset d. Let $Y_{idk} = X_{id1k} - X_{id0k}$ be the log2 IP/control ratio in replicate k, K_d be the number of replicate samples of dataset d, I denote the total number of probes, and Y_{id} be the sample mean of probe i's log2 ratios Y_{idk} in dataset d. We first compute

$$SSW_{id} = \sum_k (Y_{idk} - Y_{id})^2 \tag{A.1}$$

The variances σ_{id}^2 are then estimated by

$$\begin{aligned} v_d &= K_d - 1 \\ S_{id}^2 &= \frac{1}{K_d}\frac{SSW_{id}}{v_d} \\ \overline{S_d^2} &= \frac{\sum_i S_{id}^2}{I} \\ B_d &= min\left[1, \frac{2}{v_d+2}\frac{I-1}{I} + \frac{2}{v_d+2}\frac{(I-1)(\overline{S_d^2})^2}{\sum_i (S_{id}^2 - \overline{S_d^2})^2}\right] \\ \hat{\sigma}_{id}^2 &= (1-B_d)S_{id}^2 + B_d\overline{S_d^2} \end{aligned} \tag{A.2}$$

Here $\overline{S_d^2}$ is the sample average of S_{id}^2s across all probes, and B_d is a shrinkage factor computed using the method described in Ji and Wong (2005).

2. If dataset d has replicates (i.e., the dataset has either more than one IP sample, or more than one control sample, or both) but the replicate samples are not paired (e.g., the typical

Affymetrix data), then the probe variances σ_{id}^2 are estimated as follows. Let K_{dj} be the number of replicate samples under the condition j of dataset d, and $\bar{X}_{idj.}$ be the sample mean of X_{idjk} under the condition j and dataset d. We first compute

$$SSW_{idj} = \sum_k (X_{idjk} - \bar{X}_{idj.})^2 \tag{A.3}$$

The variances σ_{id}^2 are then estimated by

$$\begin{aligned}
v_d &= \sum_j (K_{dj} - 1) \\
S_{id}^2 &= (\frac{1}{K_{d0}} + \frac{1}{K_{d1}}) \frac{\sum_j SSW_{idj}}{v_d} \\
\overline{S_d^2} &= \frac{\sum_i S_{id}^2}{I} \\
B_d &= min \left[1, \frac{2}{v_d + 2} \frac{I-1}{I} + \frac{2}{v_d + 2} \frac{(I-1)(\overline{S_d^2})^2}{\sum_i (S_{id}^2 - \overline{S_d^2})^2} \right] \\
\hat{\sigma}_{id}^2 &= (1 - B_d) S_{id}^2 + B_d \overline{S_d^2}
\end{aligned} \tag{A.4}$$

Here $\overline{S_d^2}$ is the sample average of S_{id}^2s across all probes.

3. If IP and control samples are paired, dataset d has only one IP/control pair (i.e., $K_d = 1$), but at least one of the other datasets used in the joint analysis contains replicate pairs, then the probe variances σ_{id}^2 are estimated as follows.

$$\begin{aligned}
v &= \sum_{d'=1}^{D} (K_{d'} - 1) \\
S_i^2 &= \frac{1}{K_d} \frac{\sum_{d'} SSW_{id'}}{v} = \frac{\sum_{d'} SSW_{id'}}{v} \\
\overline{S^2} &= \frac{\sum_i S_i^2}{I} \\
B &= min \left[1, \frac{2}{v + 2} \frac{I-1}{I} + \frac{2}{v + 2} \frac{(I-1)(\overline{S^2})^2}{\sum_i (S_i^2 - \overline{S^2})^2} \right] \\
\hat{\sigma}_{id}^2 &= (1 - B) S_i^2 + B \overline{S^2}
\end{aligned} \tag{A.5}$$

Here $SSW_{id'}$ is computed using Formula A.1, and $\overline{S^2}$ is the sample average of S_i^2s across all probes.

4. If dataset d has only one IP and one control sample which are not paired (i.e., $K_{d0} = K_{d1} = 1$), but at least one of the other datasets used in the joint analysis contains replicates, then the probe variances are estimated as follows.

$$
\begin{aligned}
v &= \sum_{d'=1}^{D} \sum_{j} (K_{d'j} - 1) \\
S_i^2 &= (\frac{1}{K_{d0}} + \frac{1}{K_{d1}}) \frac{\sum_{d'} \sum_j SSW_{id'j}}{v} = \frac{2 \sum_{d'} \sum_j SSW_{id'j}}{v} \\
\overline{S^2} &= \frac{\sum_i S_i^2}{I} \\
B &= min \left[1, \frac{2}{v+2} \frac{I-1}{I} + \frac{2}{v+2} \frac{(I-1)(\overline{S^2})^2}{\sum_i (S_i^2 - \overline{S^2})^2} \right] \\
\hat{\sigma}_{id}^2 &= (1-B) S_i^2 + B \overline{S^2}
\end{aligned} \tag{A.6}
$$

Here $SSW_{id'j}$ is computed using Formula A.3, and $\overline{S^2}$ is the sample average of S_i^2s across all probes.

5. If none of the datasets used in the joint analysis contains replicate samples, and if IP and control samples are paired, then the following procedure is used to estimate the probe variances. For probe i we first compute the the log2 IP/control fold change Y_{id} for each dataset d, then the sample variance of Y_{id} across all datasets is computed. This variance is denoted as S_i^2. Let C be the 90-th percentile of all S_i^2s (i.e., the 90-th percentile across all probes). Define $\tilde{S}_i^2 = min(S_i^2, C)$. In other words, we truncate S_i^2s at their 90-th percentile and denote the truncated values as $\tilde{S}_i^2$s. Let D be the number of datasets. The probe variances σ_{id}^2 are then estimated as follows.

$$
\begin{aligned}
v &= D - 1 \\
\overline{S^2} &= \frac{\sum_i \tilde{S}_i^2}{I} \\
B &= min \left[1, \frac{2}{v+2} \frac{I-1}{I} + \frac{2}{v+2} \frac{(I-1)(\overline{S^2})^2}{\sum_i (\tilde{S}_i^2 - \overline{S^2})^2} \right] \\
\hat{\sigma}_{id}^2 &= (1-B) \tilde{S}_i^2 + B \overline{S^2}
\end{aligned} \tag{A.7}
$$

The reasoning behind this algorithm is explained below. In genomic regions without IP enrichment, the variance S_i^2 is an estimate (although not perfect) of the true probe variance σ_{id}^2. The background regions typically occupy more than 90% of the genome. In regions with TF binding, S_i^2 can often overestimate σ_{id}^2. This is because a TF can bind in some but not all datasets, whereas σ_{id}^2 is supposed to characterize variability across replicate samples. The overestimated probe variance can potentially reduce the sensitivity of peak detection. To avoid losing too much sensitivity, we choose to put a bound on S_i^2. Using the bounded $\tilde{S}_i^2$, the new shrinkage estimator provides relatively robust estimates of the probe variances.

Our experience shows that this procedure works better than an alternative approach in which σ_{id}^2 is set to zero when no replicate samples are available.

6. If none of the datasets used in the joint analysis contains replicate samples, and if IP and control samples are not paired, then for each probe i we first compute the sample variance of X_{idjk} in control samples (i.e., the variance of X_{id01} across all d). Let S_{i0}^2 denote this variance. The probe variances σ_{id}^2 are then estimated by

$$\begin{aligned} v &= D - 1 \\ S_i^2 &= (\frac{1}{K_{d0}} + \frac{1}{K_{d1}})S_{i0}^2 = 2S_{i0}^2 \\ \overline{S^2} &= \frac{\sum_i S_i^2}{I} \\ B &= min\left[1, \frac{2}{v+2}\frac{I-1}{I} + \frac{2}{v+2}\frac{(I-1)(\overline{S^2})^2}{\sum_i (S_i^2 - \overline{S^2})^2}\right] \\ \hat{\sigma}_{id}^2 &= (1-B)S_i^2 + B\overline{S^2} \end{aligned} \tag{A.8}$$

A.2 EM algorithm for estimating m and Ψ

Our estimation of $\mathbf{\Sigma}$ involves evaluation of closed-form formulas using all probes. Unlike $\mathbf{\Sigma}$, estimation of the other model parameters $\mathbf{m}$, $\mathbf{\Psi}$, π and $\mathbf{q}$ is based on iterative procedures. Directly applying these procedures to all probes is time-consuming. In practice, JAMIE uses a two-pass algorithm to keep the computation tractable. In the first pass, a simple and fast moving average method described in Ji and Wong (2005) is used to analyze each dataset separately to roughly locate peaks using a loose false discovery rate cutoff (default = 30%). If no peaks were found at the loose FDR cutoff in the first pass, it indicates that the data are too noisy, and the dataset will be excluded from the subsequent analysis (this did not happen in our real data tests). The data in the peaks will be used to estimate parameters $\mathbf{m}$, $\mathbf{\Psi}$, π and $\mathbf{q}$ (see below for details). In the second pass, the estimated parameters including $\mathbf{\Sigma}$, $\mathbf{m}$, $\mathbf{\Psi}$, π and $\mathbf{q}$ are used in conjunction with a sliding window to rescan the whole genome to detect peaks which will be reported as the final results.

This section describes the algorithm used for estimating $\mathbf{m}$ and $\mathbf{\Psi}$. After peaks are roughly located in the first pass scan, they are extended 1000bp on both ends. Probes within the peaks and extended regions are then collected. In addition, we randomly pick up the same number of probes from other regions in the genome to represent background. These two types of probes are mixed, and the following EM algorithm is applied to data from these selected probes.

Following the notations in the paper, Y_{id} is the observed IP-control difference in dataset d,

and μ_{id} is the true difference. Their probability distributions are parameterized by m_d and τ_d^2. The other parameters σ_{id}^2 and ϵ are assumed to be given for the purpose of estimating m_d and τ_d^2. We estimate m_d and τ_d^2 for each dataset separately, hence the dataset subscript d is dropped in this section to simplify notations.

The data generating model of $\mathbf{Y}$ can be translated into the following model:

$$\begin{aligned}
&Y_i|\mu_i \sim N(\mu_i, \sigma_i^2) \equiv \phi_{y_i} \\
&\mu_i|Z_i = 0 \sim N(0, \tau^2) \equiv \phi_{0i} \\
&\mu_i|Z_i = 1 \sim N(m, \tau^2) \equiv \phi_{1i} \\
&Pr(Z_i = 1|H_i = 0) = \epsilon \\
&Pr(Z_i = 1|H_i = 1) = 1 - \epsilon
\end{aligned} \tag{A.9}$$

In the JAMIE model used to scan the genome, the true Z_i states of probes are not independent (e.g., probes in the same peak should all have high probability of $Z_i = 1$). However, when we estimate m and τ^2, in order to simplify computation, we assume that Z_is of different probes are independent. In other words, probes on the array are viewed as a two-component mixture where $Z_i = 0$ and $Z_i = 1$ are the two components, and each probe independently decides which component it belongs to. This assumption is used only for the purpose of parameter estimation, and it is not used for peak detection. Let I be the total number of probes and $\lambda = Pr(Z_i = 1)$. Let $\mathbf{Z}$ be the collection of all Z_i, and $\mathbf{M}$ be the collection of all μ_i. The simplification above leads to the following joint density:

$$P(\mathbf{Y}, \mathbf{M}, \mathbf{Z}|\lambda, m, \tau^2) = \prod_{i=1}^{I} \{\phi_{y_i}[(1-\lambda)\phi_{0i}]^{1-Z_i}[\lambda\phi_{1i}]^{Z_i}\}$$

Treat $\mathbf{Y}$ as observed data, and treat $\mathbf{M}$ and $\mathbf{Z}$ as missing data. Let $\mathbf{\Theta} = (\lambda, m, \tau^2)$. The complete data log likelihood is

$$\begin{aligned}
l(\mathbf{\Theta}) &= \sum_{i=1}^{I} \{log\phi_{y_i} + (1 - Z_i)[log(1-\lambda) + log\phi_{0i}] + Z_i\,[log\lambda + log\phi_{1i}]\} \\
&= \sum_{i=1}^{I} \left\{ (1 - Z_i) \left[log(1-\lambda) - \frac{log\tau^2}{2} - \frac{\mu_i^2}{2\tau^2} \right] + Z_i \left[log\lambda - \frac{log\tau^2}{2} - \frac{(\mu_i - m)^2}{2\tau^2} \right] \right\} + c
\end{aligned} \tag{A.10}$$

Here c is a constant that does not contain the parameters to be estimated. Based on this log likelihood, the following EM algorithm can be developed to estimate m and τ^2. Let $\mathbf{\Theta}_t$ be the parameter estimates in step t.

E-step: Compute $E[Z_i|\mathbf{Y}, \mathbf{\Theta}_t]$, $E[\mu_i^2|\mathbf{Y}, \mathbf{\Theta}_t]$ and $E[Z_i\mu_i|\mathbf{Y}, \mathbf{\Theta}_t]$.

1. By integrating out μ_i from Equation A.9, one obtains

$$Y_i|Z_i = 0, \mathbf{\Theta} \sim N(0, \sigma_i^2 + \tau^2) \equiv f_{0i}$$
$$Y_i|Z_i = 1, \mathbf{\Theta} \sim N(m, \sigma_i^2 + \tau^2) \equiv f_{1i}$$

   and $E[Z_i|\mathbf{Y}, \mathbf{\Theta}] = P(Z_i = 1|\mathbf{Y}, \mathbf{\Theta}) = \frac{\lambda f_{1i}}{\lambda f_{1i} + (1-\lambda) f_{0i}}$. Define $z_{i,t} \equiv E[Z_i|\mathbf{Y}, \mathbf{\Theta}_t]$.

2. The distributions of μ_i conditional on $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{\Theta}$ are

$$\mu_i|Y_i, Z_i = 0, \mathbf{\Theta} \sim N\left(\frac{Y_i/\sigma_i^2}{1/\sigma_i^2 + 1/\tau^2}, \frac{1}{1/\sigma_i^2 + 1/\tau^2}\right),$$
$$\mu_i|Y_i, Z_i = 1, \mathbf{\Theta} \sim N\left(\frac{Y_i/\sigma_i^2 + m/\tau^2}{1/\sigma_i^2 + 1/\tau^2}, \frac{1}{1/\sigma_i^2 + 1/\tau^2}\right)$$

   Based on this,

$$E[\mu_i|\mathbf{Y}, \mathbf{\Theta}_t] = E[E[\mu_i|Y_i, Z_i, \mathbf{\Theta}_t]|\mathbf{Y}, \mathbf{\Theta}_t]$$
$$= z_{i,t}\frac{Y_i/\sigma_i^2 + m_t/\tau_t^2}{1/\sigma_i^2 + 1/\tau_t^2} + (1 - z_{i,t})\frac{Y_i/\sigma_i^2}{1/\sigma_i^2 + 1/\tau_t^2}$$

   Similarly, one can get:

3. $E[Z_i\mu_i|\mathbf{Y}, \mathbf{\Theta}_t] = z_{i,t}\frac{Y_i/\sigma_i^2 + m_t/\tau_t^2}{1/\sigma_i^2 + 1/\tau_t^2}$
4. $E[\mu_i^2|\mathbf{Y}, \mathbf{\Theta}_t] = z_{i,t}\left[\left(\frac{Y_i/\sigma_i^2 + m_t/\tau_t^2}{1/\sigma_i^2 + 1/\tau_t^2}\right)^2 + \frac{1}{1/\sigma_i^2 + 1/\tau_t^2}\right] + (1 - z_{i,t})\left[\left(\frac{Y_i/\sigma_i^2}{1/\sigma_i^2 + 1/\tau_t^2}\right)^2 + \frac{1}{1/\sigma_i^2 + 1/\tau_t^2}\right]$.

Plug in these expected values to Equation A.10, we get the Q function

$$Q \equiv Q(\mathbf{\Theta}|\mathbf{\Theta}_t) = E[l(\mathbf{\Theta})|\mathbf{Y}, \mathbf{\Theta}_t]. \tag{A.11}$$

M-step: Maximize the Q function with respect to $\mathbf{\Theta}$.

1. $\frac{\partial Q}{\partial \lambda} = 0 \Longrightarrow \lambda_{t+1} = \frac{\sum_i z_{i,t}}{I}$.
2. $\frac{\partial Q}{\partial m} = 0 \Longrightarrow m_{t+1} = \frac{\sum_i E[Z_i\mu_i|\mathbf{Y}, \mathbf{\Theta}_t]}{\sum_i z_{i,t}}$
3. $\frac{\partial Q}{\partial \tau^2} = 0 \Longrightarrow \tau_{t+1}^2 = \frac{\sum_i \left\{E[\mu_i^2|\mathbf{Y}, \mathbf{\Theta}_t] - 2m_{t+1}E[Z_i\mu_i|\mathbf{Y}, \mathbf{\Theta}_t] + m_{t+1}^2 z_{i,t}\right\}}{I}$

Upon convergence, the m and τ^2 from the last iteration will be used as the parameter values for JAMIE. The EM procedure converged well in real data and the estimates are reasonable.

A.3 EM algorithm for estimating π and q

For estimating π and $\mathbf{q}$, we assume that $\mathbf{\Sigma}$, $\mathbf{m}$ and $\mathbf{\Psi}$ are known. Again we first apply the simple moving average method to roughly locate peaks for each dataset separately, using a loose FDR cutoff (default = 30%). The union of peaks detected in all datasets are obtained. Each peak is truncated or extended to L bps. Resulting windows that are not overlapping are retained.

Let C be the indices of the probes that start the retained windows. Assume one can partition the rest of the genome into non-overlapping windows of L bps, and let $\bar{C}$ be the indices of the probes that start these windows. In reality, for windows in $\bar{C}$, B_i could either be 0 or 1. However, in order to simplify the computation, here we assume that (1) B_i of all windows are generated according to the probabilistic model described in METHODS of the main text; (2) for windows in $\bar{C}$, B_i and A_{id} are observed and are equal to 0; (3) for windows in C, B_i and A_{id} are unobserved (i.e., missing data) and can be 0 or 1. This simplification assumes that PBRs cannot occur in windows in $\bar{C}$, and it assumes that C and $\bar{C}$ are known before looking at the data. These simplifying assumptions provide useful approximations that allow us to develop computationally efficient algorithms to obtain rough estimates of π and $\mathbf{q}$. The assumption that PBRs only occur in C may cause π to be underestimated, which may lead to conservative false discovery rate estimates. The assumptions made by the simplification are only used for estimating π and $\mathbf{q}$. They are not required by the peak detection, in which the estimated parameters are used to scan the whole genome and all windows can have $B_i = 1$ with non-zero probability.

Now consider the joint probability of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{Y}$ for all genomic windows. Based on Equation 6 in the paper, this probability can be written as:

$$\prod_i P(\mathbf{Y}_i, \mathbf{A}_i, B_i|\mathbf{\Lambda}, \mathbf{U}) = \prod_{i\in\bar{C}} \left[(1-\pi)\prod_d p_{0id}\right]$$
$$\prod_{i\in C} \left\{ \left[(1-\pi)\prod_d \{(1-A_{id})p_{0id}\}\right]^{1-B_i} \left[\pi \prod_d [(1-q_d)p_{0id}]^{1-A_{id}}[q_d p_{1id}]^{A_{id}}\right]^{B_i} \right\} \quad (A.12)$$

$\mathbf{A}_i$'s and B_i's are partially missing, since their values are assumed to be known for $i \in \bar{C}$.

The complete data log-likelihood for parameters (π and $\mathbf{q}$) is

$$l(\pi, \mathbf{q}) = \sum_{i\in\bar{C}} log(1-\pi) + \sum_{i\in C}(1-B_i)log(1-\pi) +$$
$$\sum_{i\in C} B_i \left[log\pi + \sum_d \{(1-A_{id})log(1-q_d) + A_{id}logq_d\}\right] + c_1 \quad (A.13)$$

where c_1 is a constant that does not involve π and $\mathbf{q}$.

Based on this, an EM algorithm is developed to estimate π and $\mathbf{q}$. Let $\mathbf{\Omega} = (\pi, \mathbf{q})$, and $\mathbf{\Omega}_t$ be the parameter estimates at step t.

E-step: Compute the Q function $Q(\mathbf{\Omega}|\mathbf{\Omega}_t) = E(l(\mathbf{\Omega})|\mathbf{Y}, \mathbf{\Omega}_t)$ which involves calculating $E[B_i|\mathbf{Y}, \mathbf{\Omega}_t]$ and $E[B_i A_{id}|\mathbf{Y}, \mathbf{\Omega}_t]$.

1. For $i \in \bar{C}$, $B_i = 0$ and $B_i A_{id} = 0$.
2. For $i \in C$, $E[B_i|\mathbf{Y}, \mathbf{\Omega}_t] = \frac{Pr(\mathbf{Y}_i|B_i=1,\mathbf{\Omega}_t)\pi_t}{Pr(\mathbf{Y}_i|B_i=1,\mathbf{\Omega}_t)\pi_t + Pr(\mathbf{Y}_i|B_i=0,\mathbf{\Omega}_t)(1-\pi_t)} \equiv b_{i,t}$.
3. For $i \in C$, $E[B_i A_{id}|\mathbf{Y}, \mathbf{\Omega}_t] = \frac{q_{d,t} p_{1id} b_{i,t}}{q_{d,t} p_{1id} + (1-q_{d,t}) p_{0id}} \equiv a_{id,t}$

Here,

$$Pr(\mathbf{Y}_i|B_i = 1, \mathbf{\Omega}_t) = \prod_d \{q_{d,t} p_{1id} + (1 - q_{d,t}) p_{0id}\}$$
$$Pr(\mathbf{Y}_i|B_i = 0, \mathbf{\Omega}_t) = \prod_d p_{0id}$$

Plug $E[B_i|\mathbf{Y}, \mathbf{\Omega}_t]$ and $E[B_i A_{id}|\mathbf{Y}, \mathbf{\Omega}_t]$ into Equation A.13, we obtain the Q function

$$Q(\mathbf{\Omega}|\mathbf{\Omega}_t) = \sum_{i \in \bar{C}} log(1-\pi) + \sum_{i \in C} \{(1 - b_{i,t}) log(1-\pi) + b_{i,t} log\pi\} + \sum_{i \in C} \sum_d \{(b_{i,t} - a_{id,t}) log(1 - q_d) + a_{id,t} log q_d\} + c_2$$

M-step: Maximize the Q function with respect to π and $\mathbf{q}$.

1. $\frac{\partial Q}{\partial \pi} = 0 \Longrightarrow \pi_{t+1} = \frac{\sum_{i \in C} b_{i,t}}{\sum_i 1}$. The denominator is the total number of windows in C and $\bar{C}$, which is approximately equal to the total length of the genome covered by the tiling array divided by L, the length of PBR.
2. $\frac{\partial Q}{\partial q_d} = 0 \Longrightarrow q_{d,t+1} = \frac{\sum_{i \in C} a_{id,t}}{\sum_{i \in C} b_{i,t}}$.

B Assessing model assumptions

JAMIE is based on a number of model assumptions. In sections B.1 - B.5, we examine these assumptions in the context of real data analysis and discuss their implications and limitations.

In section B.6, we test JAMIE using simulations in which various model assumptions are not satisfied. Our results show that JAMIE is fairly robust to deviations from the model assumptions. It outperforms the other algorithms even in data where many model assumptions are not satisfied.

The choice of model assumptions represents a tradeoff between the ability to accurately describe the data and the complexity of the model. One can develop more complex models to faithfully reflect all data characteristics, but the model will be more difficult to implement and apply in reality.

B.1 Normality

Normality is a widely used assumption in both gene expression and genome tiling array data analysis. There are two normality assumptions in JAMIE. The first one is the normality of the true log ratios μ_{id} which are assumed to follow a mixture of normal distributions, and the second one is the normality of the observed log ratios Y_{id} conditional on μ_{id}.

Using normal QQ plots, we first explored the empirical distributions of the estimated μ_{id} in different datasets. For each dataset d, μ_{id} was estimated by $\hat{\mu}_{id} = Y_{id}$. Three representative normal QQ plots are shown in the upper panel of Figure S4. According to this analysis, $\hat{\mu}_{id}$ was approximately normal for most datasets (e.g., Gli1_limb and Sox2). Note that the heavier right tails in the plots are expected, as they correspond to probes from regions with enrichment signals (i.e., peaks). For a few datasets, $\hat{\mu}_{id}$ had heavier tails on both sides, suggesting that the normality assumption was not ideal. For example, the empirical distribution of $\hat{\mu}_{id}$ in the E2F4_G0 data was closer to a t-distribution with 8 degrees of freedom. However, even in the datasets where the normality assumption did not fit the data well, JAMIE still outperformed the other algorithms (e.g., see Figure 3). This suggests that the gain of JAMIE is reasonably robust to deviations from this assumption.

Next, we explored the normality of Y_{id} conditional on μ_{id}. For each dataset d, we first constructed pairs of IP and control samples. The real data we analyzed all contained the same number of IP and control samples. For a dataset with K IP and K control samples, we constructed K IP/control pairs. For each pair k, we computed $Z_{idk} = X_{id1k} - X_{id0k}$. Note that $Y_{id} = \sum_k Z_{idk}/K$. We then computed the standard deviation of Z_{idk}s for each probe (we denote it as s_{id}), and the standardized residual $e_{idk} = \frac{Z_{idk}-Y_{id}}{s_{id}}$. The normality of the standardized residuals e_{idk} was checked using normal QQ plots. Since Y_{id} is an average of Z_{idk}s, the normality of Z_{idk} implies the normality of Y_{id}. Representative QQ plots from the Gli and DREAM data are shown in the bottom panel of Figure S4. The Agilent data had only two replicates and the standardized e_{idk} was a constant. For this reason, QQ plots for the Agilent data were not available. The plots suggest that the normality of Z_{idk} (hence the normality of Y_{id}) conditional on μ_{id} is a reasonable

assumption that fits the real data well.

To summarize, our results show that the normality assumptions are reasonable in most cases. In all real data analyses (including those where the normality assumptions were not met), JAMIE performed better than or comparable to the other algorithms with respect to peak ranking (Figure 3). These results indicate that JAMIE is reasonably robust to deviations from the normality assumptions. We note that deviations from the normality assumptions may cause biased estimate of FDR, however as we discussed in the paper, accurate FDR estimation is not the primary goal of JAMIE, and we recommend users to use qPCR to obtain more reliable FDR estimates whenever possible.

B.2 Equal variance

JAMIE assumes that the noise and signal components of $f(\mu_{id}|H_{id})$ (namely $\phi(\mu_{id}; 0, \tau_d^2)$ and $\phi(\mu_{id}; m_d, \tau_d^2)$) have equal variance τ_d^2. Our initial design of JAMIE allowed unequal variances. In other words, we initially used $\phi(\mu_{id}; 0, \tau_d^2)$ and $\phi(\mu_{id}; m_d, \omega_d^2)$ where $\tau_d^2 \neq \omega_d^2$. With the unequal variance assumption, we estimated τ_d^2 and ω_d^2 in the real data using an EM algorithm similar to the one described in section A.2. The background standard deviation τ_d ranged from 0.3 to 0.4 in different datasets, and the signal standard deviation ω_d ranged from 0.4 to 0.5. Therefore, in real data, the variance of the signal and noise components are not equal.

However, we found that building JAMIE based on the unequal variance assumption caused problems. In particular, the algorithm reported peaks with negative log2 IP/control fold changes. This is because with unequal variances, the likelihood ratio between the peak and background states is no longer monotone with respect to the log ratio, and as a result, a region with a large negative log2 fold change can have higher probability to be classified as a peak as opposed to background. This is undesirable since one expects that real peaks should contain probes with positive log2 ratios.

To avoid this problem, we decided to force the two states to have equal variance. Although this is an idealized assumption, our simulations (see Section B.6) and real data tests (e.g., Figure 3) show that our current implementation based on this assumption works well. In both real data analyses and simulations where the equal variance assumption was not true, JAMIE robustly outperformed the other methods with respect to peak ranking. Similar to the normality assumptions, deviations from this assumption may result in biased FDR estimates. We are currently exploring methods to model unequal variances τ_d^2 without incurring the monotonicity issue above. Potential methods include using truncated normals or mixtures of normals (or mixtures of t distributions) constrained in some way to guarantee monotonicity. These will result in more sophisticated models. Developing a robust and computationally efficient algorithm to implement

these models to handle the large amount of real data is not straightforward. This topic deserves further investigation in future.

B.3 Independence

JAMIE assumes that μ_{id} and Y_{id} from different probes are independent conditional on probes' peak status H_{id} (see Formulas 1-2). To check whether this assumption is reasonable, we computed the lag-1 autocorrelation of Y_{id} in peak and background regions respectively. In the background regions (i.e., regions not declared as peaks by JAMIE), the lag-1 autocorrelation ranged from 0.05 to 0.2 in different datasets. In the peak regions, the lag-1 autocorrelation ranged from 0.1 to 0.3. These results suggest that conditional on H_{id}, there are weak correlations between neighboring probes. Assuming conditional independence ignores these correlations. However, the assumption simplifies the computation. For example, one can model correlations among probes by using a multivariate normal distribution to jointly describe probe intensities within a peak. However, if one uses this model, one needs to compute the inverse of a covariance matrix for each window, which is computationally prohibitive in the typical ChIP-chip data. Our real data analyses showed that even though there were weak correlations between probes, JAMIE performed better than the other algorithms with respect to peak ranking (e.g., Figure 3). Like the other assumptions, ignoring the between-probe correlations may result in biased estimates of FDR. Violations of the conditional independence assumption may also result in biased estimates of peak length. However, our simulations in section B.6 show that the peak length estimates provided by JAMIE in the correlated data usually are reasonable, and they are better than or comparable to the peak length estimates provided by the other algorithms. Therefore, we conclude that the conditional independence assumption is an appropriate assumption to use from a practical point of view.

It should be pointed out that our model does not assume independence of H_{id}s from different probes. In fact, if a probe is a start of an active binding peak and if the peak is W bp long, then all downstream probes within the W bps of the staring probe will be labeled as $H_{id} = 1$. Clearly this implies dependence among probes. When we estimate parameters $\mathbf{m}$ and $\mathbf{\Psi}$, we did assume that probes' hidden states are independent in order to derive a simple algorithm (see Section A.2). However, this simplification is only used for estimating $\mathbf{m}$ and $\mathbf{\Psi}$. After the parameters are estimated, we no longer assume independence of H_{id}s when we detect peaks.

B.4 Peak shape

In reality, ChIP-chip peaks tend to have a triangle or bell shape. This has not been reflected by our current model. In triangular- or bell-shaped peaks, probes from the peak centers have stronger

enrichment signals than those between the peak centers and flanking background regions. In our current implementation of JAMIE, however, it is assumed that probe signals within a peak are identically distributed. This implies that the peak shape is rectangular. Estimates of $\mathbf{m}$ based on this assumption represent biased estimates of real peak heights (i.e., the peak maxima), since both probes in the peak centers and those in the intermediate regions that rise from background to peak centers are considered to be signals. This bias could affect the sensitivity of JAMIE and the FDR estimation. The triangular- and bell-shaped peaks seen in reality could also affect the peak length estimation, since probes on the peak boundaries have weak enrichment signals, and it is difficult to distinguish them from background noise. Explicitly modeling the peak shape has the potential to further improve JAMIE. This was not explored in our current paper, but is worthwhile for future investigation.

B.5 Length of potential binding region

JAMIE asks users to specify the length (L) of PBR. This length is fixed in the peak detection. If the PBR length does not match the real peak length, the algorithm's sensitivity and specificity may be affected, and the peak length estimates may be biased. In practice, however, this is not a serious issue, since one can always perform exploratory analysis and visualize the data using existing software tools such as Integrated Genome Browser or CisGenome. One can then specify the PBR length to be the typical peak length (or a little longer than the typical length) observed in the data.

B.6 JAMIE's performance in simulations in which the assumptions are not satisfied

We performed a series of simulations to test JAMIE's performance when the model assumptions were violated. Similar to the simulation presented in Section 4.1 of the main document, we added simulated peaks to real input control data. As a result, probe intensity distributions and probe correlations in background regions in the real data were all retained.

In the first simulation, PBRs were 1000 bp long, peak lengths were uniformly drawn from U[300, 1000], and the peaks were rectangular. The true log fold changes (i.e., μ_{id}s) of different probes within the same peak were drawn from a multivariate normal distribution. The marginal distribution of each μ_{id} was $N(1, 0.25)$, and the pairwise correlation between any two probes' μ_{id}s within the same peak was 0.7. Since we used the real input control data as background, the correlations among probes in the background regions were completely determined by the real data. The variances τ_d^2 in the background regions were estimated to be around 0.09 in the real

data. In this simulation, both the equal variance (τ_d^2) assumption and the assumption that μ_{id}s are independent conditional on H_{id}s were violated.

Figure S5 shows the performance of various algorithms. Only one dataset is shown. Results for the other datasets were similar. The analysis shows that JAMIE pooling still outperformed the other three methods with respect to peak ranking, and JAMIE pooling still had higher sensitivity compared to JAMIE single. At the same nominal FDR cutoff, JAMIE pooling also reported more peaks than MAT, however this was partly due to the conservative FDR estimates provided by MAT in this simulation. Compared to TileMap, JAMIE's FDR estimates were more conservative, but JAMIE pooling still reported more peaks (hence more sensitive) when the nominal FDR was smaller than 0.08.

In the second simulation, peaks were simulated to have a triangle shape, and their heights (i.e., the maximal μ_{id} of each peak) were sampled from a uniform distribution U[0.5, 2]. The lengths of the peaks were uniformly distributed between 300 and 1000 bps. For each peak, given the height, width and peak center, μ_{id}s of probes within the peak were computed deterministically according to the triangle shape (rather than randomly and independently sampled from $N(m_d, \tau_d^2)$), and the μ_{id}s were added to the original input probe intensities (which were normalized and log2 transformed). Other aspects of the simulation such as the PBR lengths were kept the same as the previous simulation. In this new simulation, probe signals within each peak were not independent conditional on H_{id}, the true peak shape was not the shape assumed by the JAMIE model, and the distribution of μ_{id}s within the peaks was no longer a normal distribution. Figure S6 shows the performance of various algorithms in this new simulation. JAMIE pooling again performed better than the other algorithms in terms of peak detection accuracy.

In the third simulation, peaks were triangular. Peak heights were drawn from a Gamma distribution with mean of μ_{id} being 1, and variance of μ_{id} being 0.2. PBR lengths were 1500 bps, and peak lengths were uniformly distributed between 300 and 1500 bps. When JAMIE was used to detect peaks, the PBR length was set to $L = 1000$ bps, which did not match the real peak lengths. The results are shown in Figure S7. Overall, JAMIE pooling still performed the best with respect to peak ranking.

In addition to peak ranking and FDR, all the algorithms tested here provide start and end coordinates of predicted peaks. Based on this information, one can compute the peak lengths. We compared the peak length estimates of JAMIE pooling, MAT and TileMap to the true peak lengths. Table S2- S5 listed the relative errors of peak length estimates. The relative error was defined as (Estimated peak length - True peak length) / True peak length.

Table S2 shows the results for the simulation presented in Section 4.1 of the main manuscript. Table S3 shows the results for the first simulation in the present section. These tables show that

JAMIE was able to provide relatively accurate peak length estimates for rectangular peaks. The estimates were slightly down biased, but the JAMIE's estimates had smaller relative errors (in terms of magnitude) than the errors of TileMap and MAT.

Table S4 shows the results for the second simulation in this section. In this simulation, peaks were triangular. Compared to Table S2 and S3, the peak length estimates of JAMIE showed bigger negative biases. This was likely due to the fact that probes at the boundaries of triangular peaks tend to have weak enrichment signals which made it difficult to distinguish them from noise. In this simulation, the TileMap peak length estimates had the smallest bias (in terms of magnitude), the bias associated with JAMIE's estimates were intermediate, and the length estimates of MAT had big positive biases.

Table S5 shows the results for the third simulation in this section. Here the peaks were triangular and the PBR length used to detect peaks was smaller than the true peak lengths. Similar to Table S4, JAMIE's length estimates were negatively biased. This was partly due to the triangle peak shape and partly due to the incorrect specification of the PBR length. However, compared to the other algorithms, JAMIE's estimates were still quite reasonable. In fact, it had smaller biases in most length intervals compared to MAT.

Finally, it should be pointed out that in practice, one can obtain high resolution (usually a few base pairs) predictions of TFBSs from ChIP-chip data by performing *de novo* motif discovery or by mapping known transcription factor binding motifs to the peak DNA sequences. For this reason, accurately estimating the peak lengths is relatively less important compared to accurately identifying the peak locations. We therefore conclude that the bias of JAMIE's peak length estimate was within a reasonable range from a practical point of view.

Together, the simulation results in this section showed that JAMIE performed reasonably well when the model assumptions were violated, and it robustly outperformed the other algorithms with respect to peak ranking. This observation was consistent with the real data analyses, in which JAMIE consistently showed favorable performance. In terms of peak length estimates, JAMIE also performed reasonably well compared to the other algorithms.

C Summary of datasets used in the real data tests

We performed three real data tests. Table S6 lists the data used in these tests. The first test involved three ChIP-chip datasets for detecting TFBSs of OCT4, SOX2 and NANOG in human embryonic stem (ES) cells (Boyer et al., 2005). The data were generated using Agilent promoter tiling arrays (ArrayExpress: E-WMIT-5) and is called "Agilent data" in the paper. Each dataset

contained two replicates. The average $log_2(Cy5/Cy3)$ ratio were used as Y_{id} in our analysis. The three TFs were previously known to cooccupy many human promoters, although each TF also had their own specific binding sites. The second test ("Gli data") contained four Gli datasets generated on Affymetrix mouse promoter 1.0R arrays (GEO: GSE11062, GSE17682). The data were produced by two labs to map TFBSs of Gli1 and Gli3 in different developmental and pathological contexts. Each dataset had 3 IP and 3 control samples. The third test ("DREAM data") involved four datasets (GEO: GSE7516) on Affymetrix human promoter 1.0R arrays. The data were used to identify DNA binding of a p130 complex termed DREAM (DP, RB-like, E2F, and MuvB) (Litovchick et al., 2007). It was reported that the DREAM complex binds to more than 800 human promoters in G0 phase of the cell cycle, but dissociates in S phase. ChIP-chip was performed to detect binding sites of four proteins, including p130, E2F4 and the mammalian homologs of synMuvB proteins LIN9 and LIN54. Our test analyzed the four binding datasets in G0-arrested cells. Each dataset had 3 IP and 3 control samples. All data were quantile normalized before processing.

D Motif enrichment in detected peaks in real data

Figures S9-S14 compare the motif contents in the detected peaks in the three real data tests. In all figures, X-axis is the number of top ranked peaks. Y-axis is the percentage of peaks with at least one motif site. Figure S9 shows the enrichment of Oct4 motif in peaks from Agilent data (Boyer *et al.*, 2005). Figure S10 shows the enrichment of Gli motif in peaks from Gli data (Vokes *et al.*, 2008). Figures S11-S14 shows the enrichment of four different motifs in peaks detected from DREAM data: E2F4, nMyc, NRF2 and CREB (Litovchick *et al.*, 2007). These results show that the peaks detected by JAMIE pooling have higher or similar motif content compared with peaks detected from other methods.

E Supplementary Tables

Table S1: Control files used in simulation

Dataset	GEO accession number
1 and 2	GSE12283
3 and 4	GSE11062

Table S2: Average relative errors of the peak length estimates in the simulation in Section 4.1

	Relative error (%)		
True peak length (bp)	JAMIE	MAT	TileMap
(300,600]	-2.6	179.6	28.5
(600,800]	-9.0	113.7	10.3

Table S3: Average relative errors of the peak length estimates in the first simulation in Section B.6

	Relative error (%)		
True peak length (bp)	JAMIE	MAT	TileMap
(300,600]	-5.7	186.7	26.6
(600,800]	-8.3	115.4	16.8
(800,1000]	-8.6	97.5	12.6

Table S4: Average relative errors of the peak length estimates in the second simulation in Section B.6

	Relative error (%)		
True peak length (bp)	JAMIE	MAT	TileMap
(300,600]	-17.3	172.2	17.0
(600,800]	-30.4	108.6	-0.1
(800,1000]	-36.0	72.3	-9.1

Table S5: Average relative errors of the peak length estimates in the third simulation in Section B.6

	Relative error (%)		
True peak length (bp)	JAMIE	MAT	TileMap
(300,600]	-24.2	168.3	21.0
(600,900]	-42.0	93.5	-10.5
(900,1200]	-47.2	61.6	-21.7
(1200,1500]	-53.9	34.2	-29.2

Table S6: Real datasets used in the paper

Dataset	Accession number	Platform	Reference
Agilent data	ArrayExpress E-WMIT-5	Agilent Human promoter 44k	Boyer *et al.* (2005)
Gli data	GEO GSE11062 GSE17682	Affymetrix mouse promoter 1.0R	Vokes *et al.* (2008) Lee *et al.* (2010)
DREAM data	GEO GSE7516	Affymetrix human promoter 1.0R	Litovchick *et al.* (2007)

F Supplementary Figures

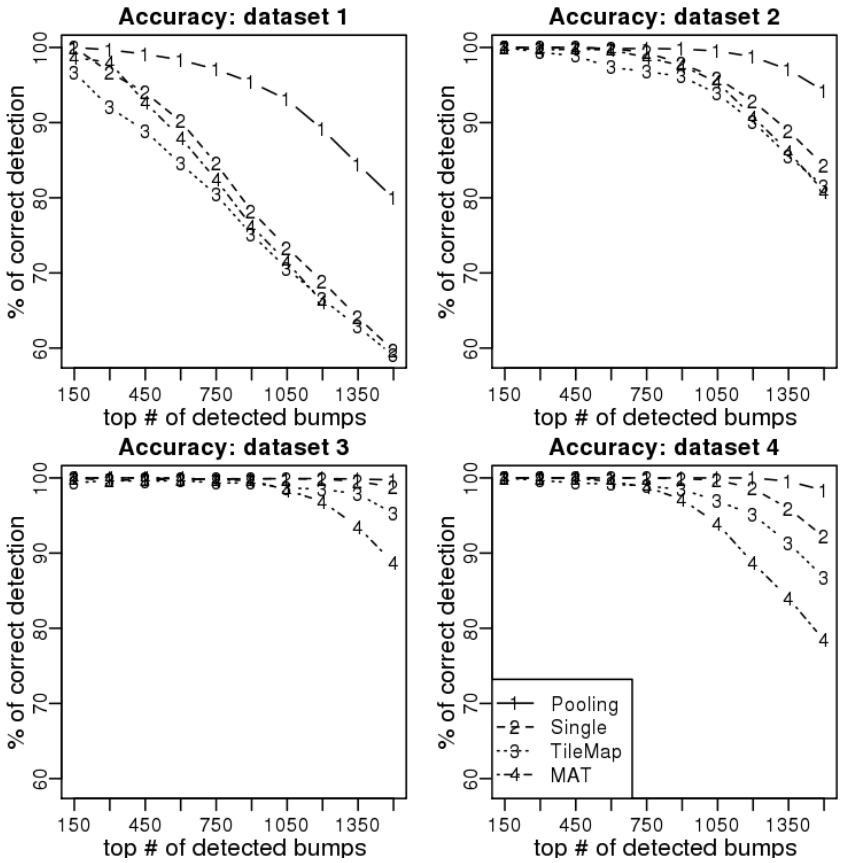

Figure S1: Comparisons of peak detection accuracy of different methods (JAMIE pooling, JAMIE single, TileMap and MAT) in four simulated datasets. X axis is number of top ranked peaks. Y axis is the percentage of peaks being true positives.

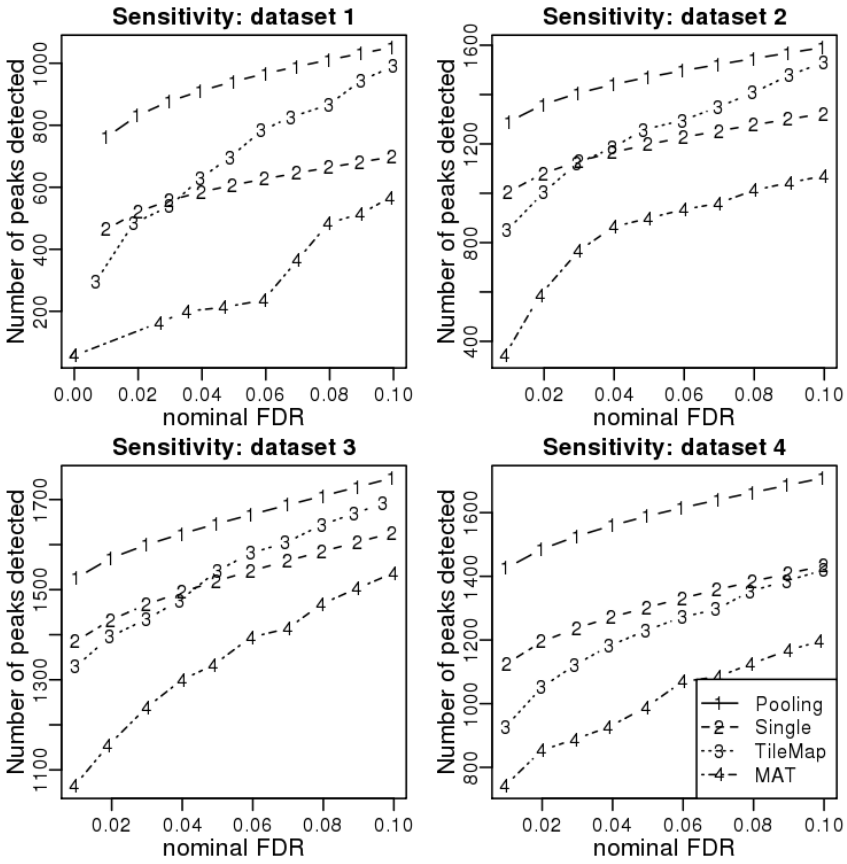

Figure S2: Comparisons of peak detection sensitivity at various nominal FDR cutoffs. X axis is the nominal FDR. Y axis is the number of peaks reported.

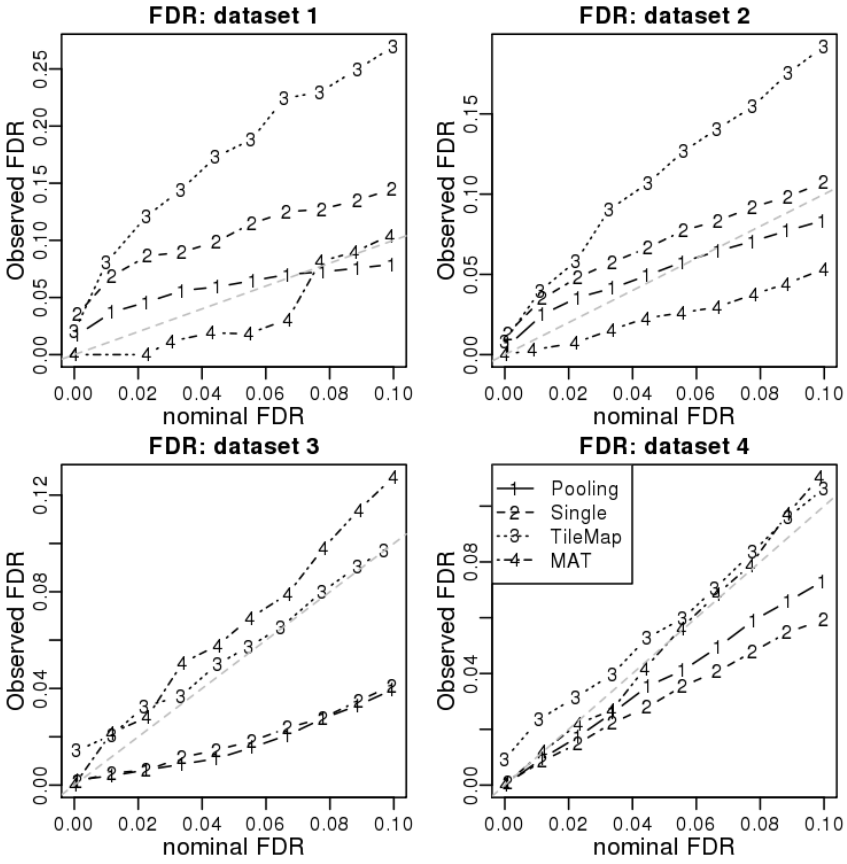

Figure S3: Comparisons of observed versus nominal FDRs of different methods in four simulated datasets.

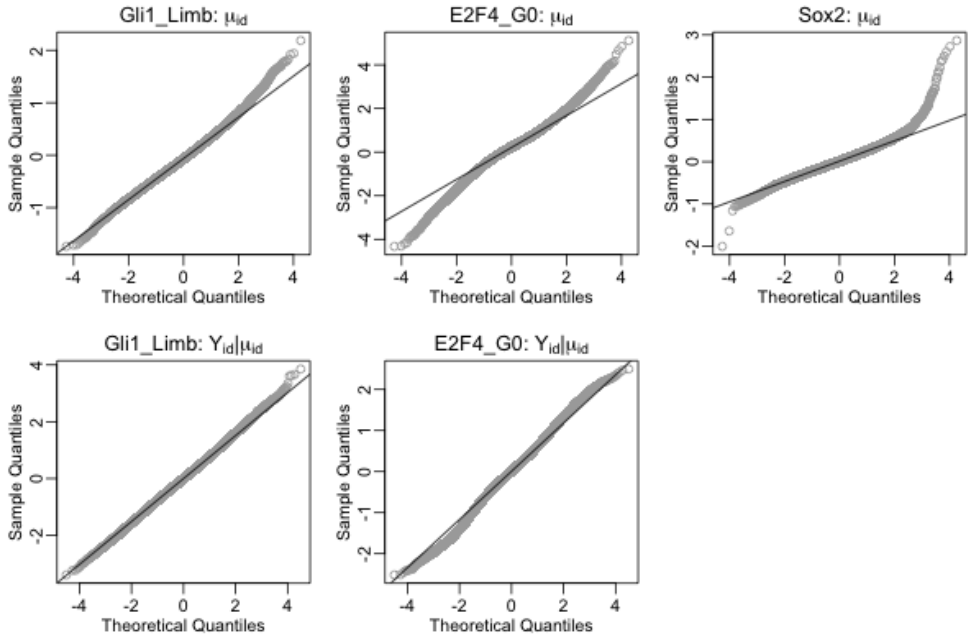

Figure S4: Normal QQ plots for checking normalities of μ_{id} and $Y_{id}|\mu_{id}$.

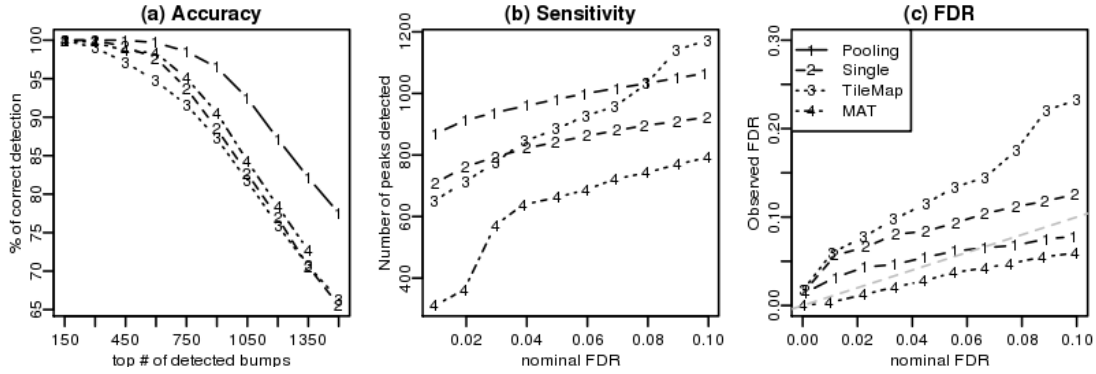

Figure S5: Simulation results when peaks are rectangular, probes within a peak are correlated, and the equal variance assumption does not hold true.

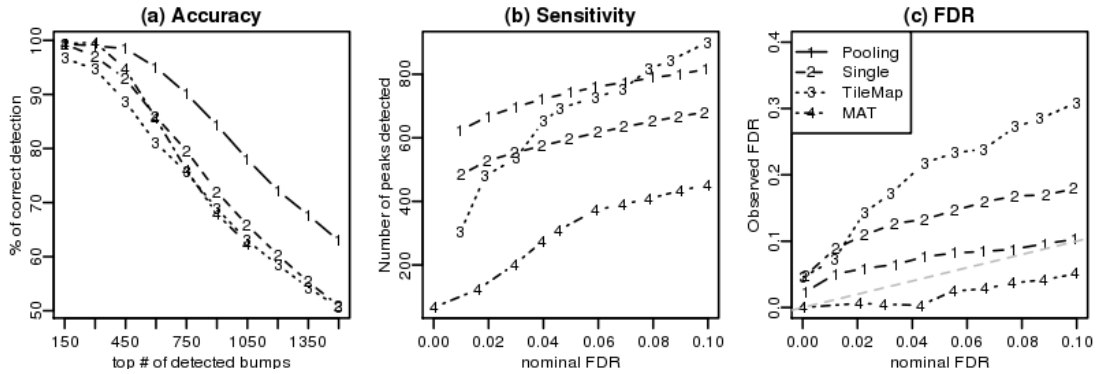

Figure S6: Simulation results when peaks are triangular and peak heights follow a uniform distribution.

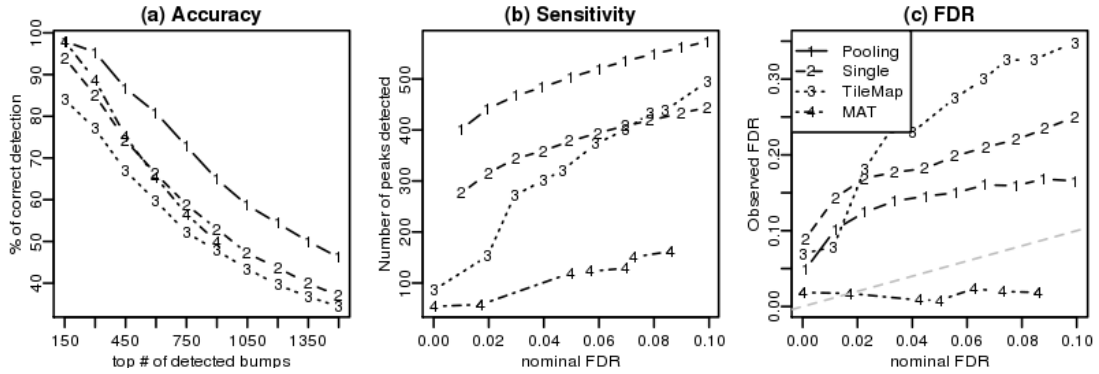

Figure S7: Simulation results when peaks are triangular and peak heights follow a Gamma distribution.

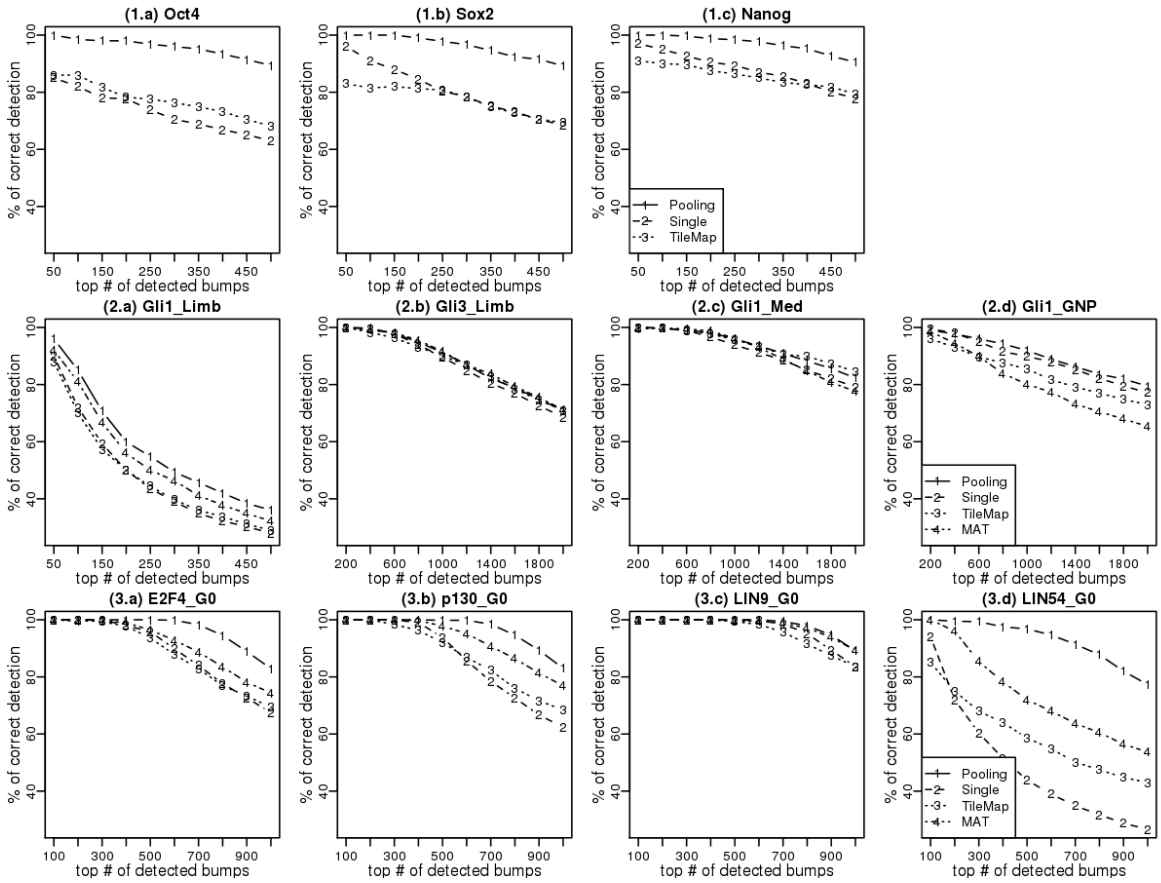

Figure S8: Comparisons of peak detection consistency of JAMIE pooling, JAMIE single, TileMap and MAT in three real ChIP-chip data. This figure compares the self-consistency of different algorithms in the three real data tests. For each test, each of the four algorithms was applied to the reduced data and compared to the gold standard constructed by itself using full data (i.e., all replicates). For each dataset in the test, the number of peaks in the gold standard was kept the same for all algorithms and was chosen as the minimal number of peaks reported by the four algorithms at the 30% FDR cutoff. X-axis is the number of top ranked peaks. For each algorithm, the Y-axis shows the average percentage of correct detections, where correctness is evaluated by the gold standard. The first row shows the results for Agilent data, the second row is for Gli data, and the third row is for DREAM data. JAMIE pooling shows the best overall results with respect to peak detection consistency.

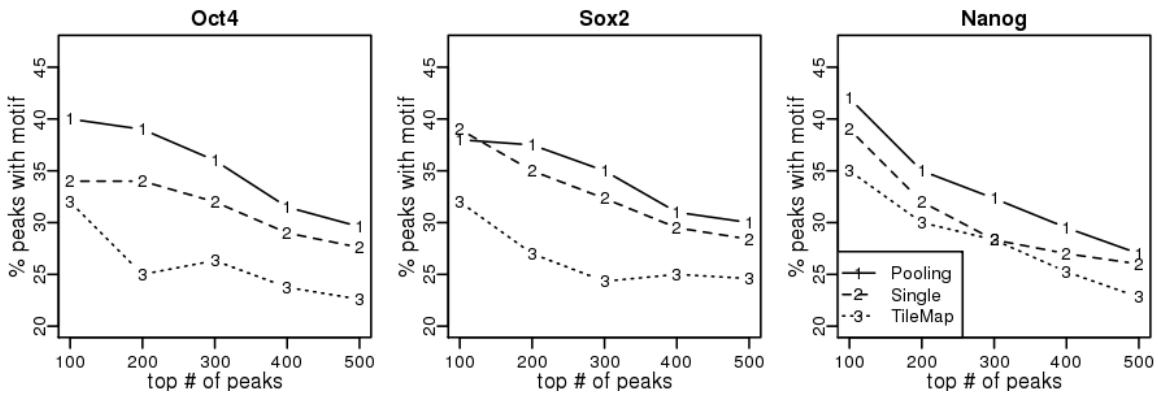

Figure S9: Percentage of peaks with at least one Oct4 motif in the Agilent data.

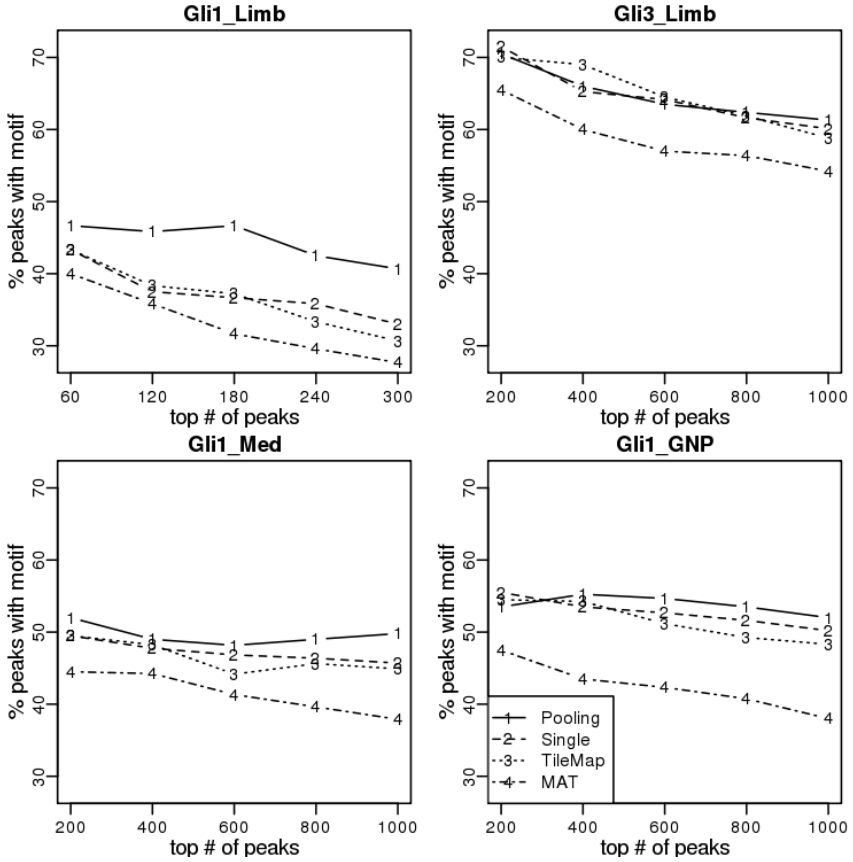

Figure S10: Percentage of peaks with at least one Gli motif in the Gli data. The motif enrichment in the Gli3 dataset is higher than the other three Gli1 datasets. This has several possible explanations. First, it was observed that the Gli3_Limb data had higher signal-to-noise ratio which might be due to technological reasons such as protocols or quality of antibodies. Owing to the high signal-to-noise ratio, it is reasonable to expect that, at the same rank level, peaks detected from this dataset are more likely to be true binding sites. Second, it is also possible that in vivo binding of Gli3 to the Gli motif has stronger affinity than Gli1 binding. Computationally we cannot tell which one is the true explanation behind the observed differences in motif enrichment.

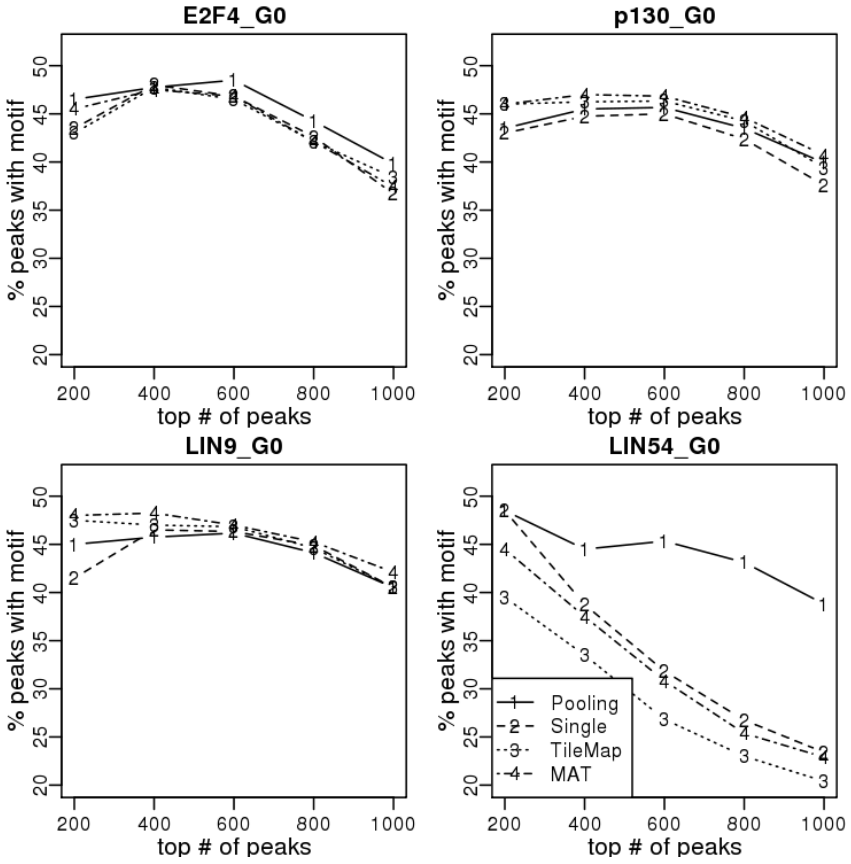

Figure S11: Percentage of peaks with at least one E2F4 motif in the DREAM data.

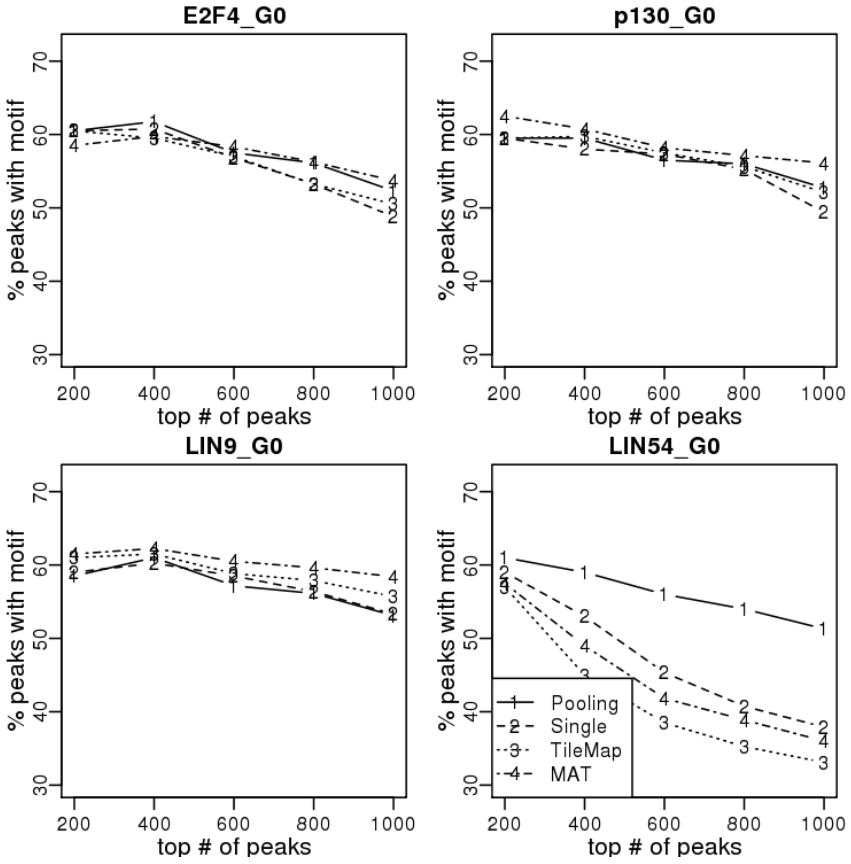

Figure S12: Percentage of peaks with at least one nMyc motif in the DREAM data.

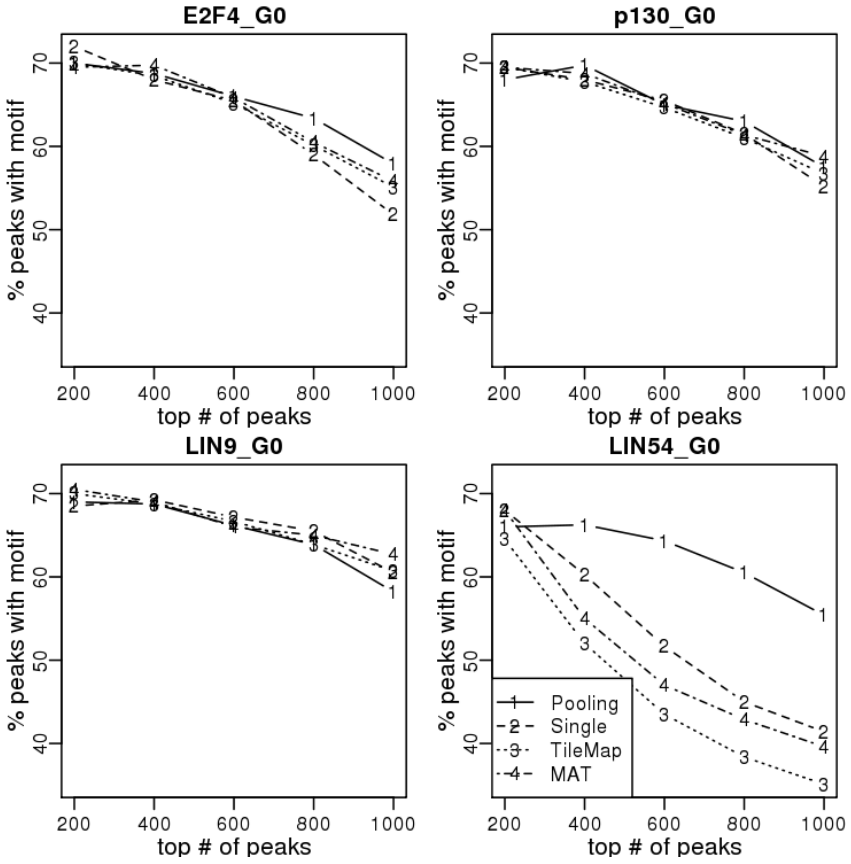

Figure S13: Percentage of peaks with at least one NRF2 motif in the DREAM data.

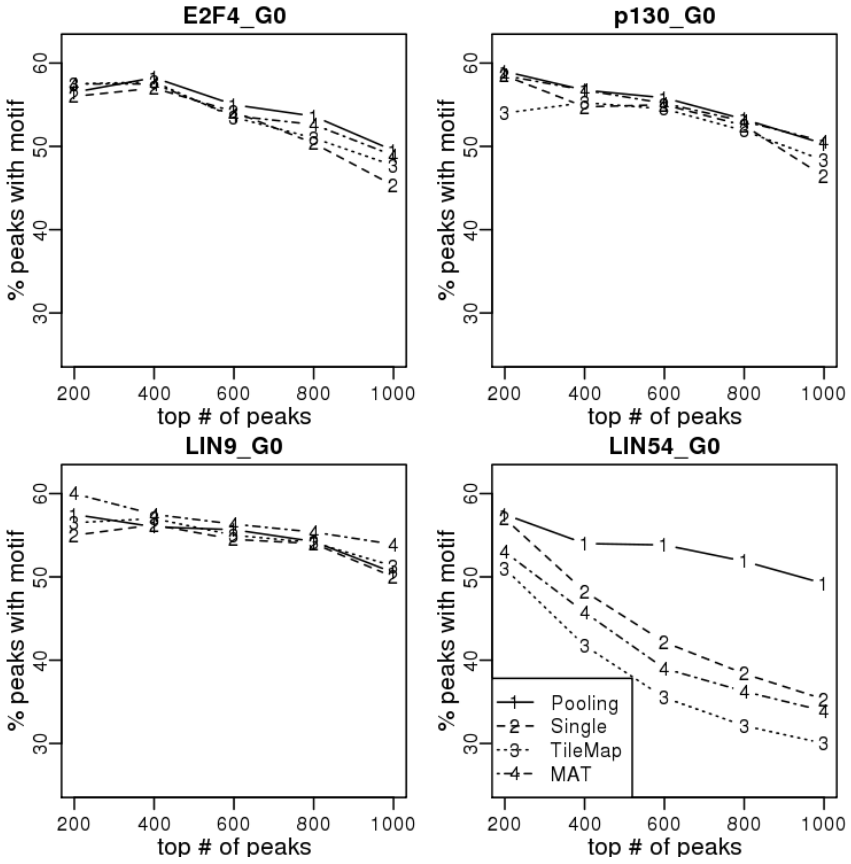

Figure S14: Percentage of peaks with at least one CREB motif in the DREAM data.