PICS: Probabilistic Inference for ChIP-seq

Xuekui Zhang,1 Gordon Robertson,2 Martin Krzywinski,2 Kaida Ning,3 Arnaud Droit,4
Steven Jones,2 and Raphael Gottardo1,4,5,*

1Department of Statistics, University of British Columbia, Vancouver, BC, Canada
2BC Cancer Agency Genome Sciences Centre, Vancouver, BC, Canada
3Bioinformatics Training Program, University of British Columbia, Vancouver, BC, Canada
4Clinical Research Institute of Montreal (IRCM), 110 Avenue des Pins Ouest, Montreal, QC, Canada
5Université de Montreal, Montreal, QC, Canada
*email: raphael.gottardo@ircm.qc.ca

Summary. ChIP-seq combines chromatin immunoprecipitation with massively parallel short-read sequencing. While it can profile genome-wide in vivo transcription factor-DNA association with higher sensitivity, specificity, and spatial resolution than ChIP-chip, it poses new challenges for statistical analysis that derive from the complexity of the biological systems characterized and from variability and biases in its sequence data. We propose a method called PICS (Probabilistic Inference for ChIP-seq) for identifying regions bound by transcription factors from aligned reads. PICS identifies binding event locations by modeling local concentrations of directional reads, and uses DNA fragment length prior information to discriminate closely adjacent binding events via a Bayesian hierarchical t-mixture model. It uses precalculated, whole-genome read mappability profiles and a truncated t-distribution to adjust binding event models for reads that are missing due to local genome repetitiveness. It estimates uncertainties in model parameters that can be used to define confidence regions on binding event locations and to filter estimates. Finally, PICS calculates a per-event enrichment score relative to a control sample, and can use a control sample to estimate a false discovery rate. Using published GABP and FOXA1 data from human cell lines, we show that PICS’ predicted binding sites were more consistent with computationally predicted binding motifs than the alternative methods MACS, QuEST, CisGenome, and USeq. We then use a simulation study to confirm that PICS compares favorably to these methods and is robust to model misspecification.

Key words: Bayesian hierarchical model; ChIP-seq; EM algorithm; Mappability; Missing values; Mixture model; Transcription factor; Truncated data; t-distribution.

1. Introduction

ChIP-seq, which combines chromatin immunoprecipitation with massively parallel short-read sequencing, offers high specificity, sensitivity, and spatial resolution in profiling in vivo protein-DNA association; histones, histone variants, and modified histones; nucleosome positioning; polymerases and transcriptional machinery complexes; and DNA methylation (Holt and Jones, 2008; Park, 2009). While sequencing overcomes certain limitations of profiling with microarrays (ChIP-chip), it raises statistical and computational challenges, some of which are related to those for ChIP-chip, and others that are novel. A typical ChIP-seq data set consists of millions or tens of millions of sequence reads that are generated from ends of immunoprecipitated DNA fragments. The quality of called bases varies along and between reads, and, as the sequencing technology evolves, read lengths and quality are increasing, as are the number of reads generated in a machine run. Current read lengths are generally between 36 and 75 base pairs (bp). While pairs of end reads can be generated from each DNA fragment, current ChIP-seq data typically consist of single-end reads, in which each DNA fragment contributes a directional read from only one randomly selected fragment end ((7–8) in Figure 1).

After read sequences have been aligned to a reference genome (e.g., Li and Durbin, 2009), the aligned read data are transformed into a form that reflects local densities of immunoprecipitated DNA fragments, and, in the work described here, into estimates of locations where transcription factors were associated with DNA in the experimental cellular system. The analysis is complicated by biases and artifacts in local read densities that can be introduced in sequencing and aligning, and by chromatin structure and genome copy number variations (Johnson et al., 2008; Kharchenko, Tolorukov, and Park, 2008; Rozowsky et al., 2009). As well, repetitive sequences can prevent aligning (mapping) reads to unique genomic locations (Robertson et al., 2008; Rozowsky et al., 2009), and reads that cannot be uniquely aligned are typically removed from the analysis. In typical mammalian experiments, 30–40% of reads may be discarded, but higher rates can be encountered in particular experiments. Because of ChIP-seq’s cost-effectiveness, such global losses are usually not an important practical consideration; however, few analysis methods correct for the local biases in aligned read densities that are caused by repetitive regions.

Certain types of biases in read density profiles can be estimated by sequencing a control sample in addition to the...
Figure 1. Details of a ChIPseq experiment. A DNA-binding protein is cross-linked to its in vivo genomic DNA targets, and the chromatin (a complex of DNA and protein) is isolated (1). The DNA with the bound proteins is extracted from the cells, and is sheared by sonication into fragments of average length 500–1,000 bp (2). DNA fragments that are cross-linked to the protein of interest are enriched by immunoprecipitation with an antibody that specifically binds that protein (3–4). After the immunoprecipitation step, the DNA is separated from the protein (5), the resulting suspension of IP-enriched DNA is size selected on a gel, and only smaller fragments (e.g., 100–300 bp) are retained (6). Then, the size-selected, IP-enriched DNA is sequenced to generate millions of short reads, each of which represents either a fragment start or end (7–8). (In an alternative “paired end” experiment that is rarely used for ChIP-seq, a read is generated from each end of each DNA fragment.) After read sequences are aligned to a reference genome, read positions can be used to infer binding site positions. (8) shows two binding sites, with the right-hand site more enriched in DNA fragments. Fragments that do not align with a binding site reflect biases like nonspecific immunoprecipitation, misalignment, etc. This figure appears in color in the electronic version of this article.
Probabilistic Inference for ChIP-seq

and we refer the reader to Buck and Lieb (2004) for an overview.

In summary, once reads have been aligned to a reference genome, there are at least four central analysis issues: interpreting the information in local densities of directional reads; identifying local read densities that represent false positives; addressing biases in read densities that arise from local variations in the efficiency with which reads can be aligned to unique genomic locations; and segmenting the enrichment profile in order to identify a statistically and biologically meaningful subset of enriched regions.

Short-read sequencing technology has evolved rapidly since its commercial introduction early in 2007, and, as was the case while ChIP-chip developed as an experimental approach (e.g., Johnson et al., 2006; Gottardo et al., 2008), statistical analysis methods for ChIP-seq are actively being developed. For example, Valouev et al. (2008) introduced QuEST, a method based on kernel density estimates of forward and reverse read counts, which allows estimating DNA fragment lengths. The forward and reverse profiles are then combined to estimate binding site locations and quantify local enrichment. Given control sample data, QuEST can estimate a false discovery rate (FDR). Like QuEST, MACS (Zhang et al., 2008) uses both forward and reverse read profiles to empirically model the “shift size” of ChIP-seq reads, and uses this to improve the spatial error of the predicted binding sites. Instead of using kernel density estimates, MACS uses a parametric model based on a dynamic Poisson distribution to identify and quantify locations where the protein of interest binds. Ji et al. (2008) introduced a “CisGenome” analysis pipeline for the analysis of ChIP-chip and ChIP-seq data. Their method is also based on a negative binomial background model, and includes functionality not offered by MACS and QuEST (e.g., filtering atypical regions, and different types of FDR estimates). Another method that has been shown to perform well is USeq, which uses window-level binomial p-values followed by an FDR correction to call individual binding events (Nix et al., 2008).

While these methods have established statistical approaches for ChIP-seq analysis, model-based and Bayesian approaches are in earlier stages of development. In the work described here, we introduce PICS, a method for probabilistic inference of ChIP-seq data that is based on a Bayesian hierarchical truncated $t$-mixture model. PICS integrates four important components. First, it jointly models local concentrations of directional reads. Second, it uses mixture models to distinguish closely adjacent binding events. Third, it incorporates prior information for the length distribution of immunoprecipitated DNA fragments to help resolve closely adjacent binding events and to identify events that have atypical fragment lengths. Fourth, it uses precalculated whole-genome read mappability profiles to adjust local read densities for reads that are missing due to genome repetitiveness. For each binding event, PICS returns an enrichment score that is relative to a control sample when such a sample is available, and it can use a control sample to estimate a false discovery rate. Finally, because it is based on a probabilistic model, PICS can compute measures of uncertainty for binding site estimates; these can be used to estimate binding site locations and to filter out low-confidence regions.

The article is organized as follows. In Section 2, we introduce the data structure and some notation. In Section 3, we present our Bayesian hierarchical truncated $t$-mixture model. Section 4 discusses parameter estimation, inference, and detection of binding sites. In Section 5, we apply PICS to two published human ChIP-seq data sets and compare its results to those from QuEST, MACS, CisGenome, and USeq. Section 6 presents the results of a simulation study that evaluates the robustness of our model to mis specification and compares the performance of PICS and the four other methods. In Section 7, we briefly discuss our results and possible extensions.

2. Data, Preprocessing, and Notation
We used two ChIP-seq data sets that have been analyzed by other groups. Zhang et al. (2008), using MACS, identified binding sites of FOXA1 (hepatocyte nuclear factor 3α) in human MCF7 breast cancer cells. Valouev et al. (2008), using QuEST, identified binding sites of the growth-associated binding protein (GABP) in human Jurkat T cells. Each data set consists of single-end reads for a treatment (ChIP) and an input DNA control sample. The FOXA1 data consist of 3,909,507 treatment reads and 5,233,322 control reads, while the GABP data consist of 7,830,602 treatment reads and 17,028,066 control reads.

Because ChIP-seq aligned-read data are usually sparse, consisting largely of regions in which few or no reads are observed, we first preprocess the read data by segmenting the genome into regions, each of which has a minimum number of reads that aligned to forward and reverse strands. We detect such regions using a $w$-bp sliding window with an $s$-bp step size, counting the number of forward and reverse strand reads in the left and right half-windows, respectively, and we retain windows that contain at least one forward read and one reverse read. For each chromosome, after merging overlapping windows and removing merged regions that have fewer than two forward or reverse reads, we obtain a disjoint set of candidate regions, each of which we analyze separately. Because DNA fragments are often between 100 and 300 bp long after gel size selection, for the work described here we chose $w = 100$ bp, and set $s = 10$ bp for computational convenience. We tested other values for $w$ and $s$ and obtained essentially the same candidate regions.

From this point, we will use the following terminology. A candidate region is a region obtained by our sliding window approach, i.e., one with enough forward/reverse reads to be processed by PICS. A binding event refers to a location where the protein of interest is associated with DNA. At a binding event the {in vivo} protein was either interacting with the DNA directly at a binding site, i.e., a DNA motif that the protein recognizes and binds to (D’haeseleer, 2006), or indirectly via membership in a DNA-associated protein complex. We will define a binding site’s midpoint as its position or location. Note that a candidate region can result from more than one binding event, and so can contain more than one binding site.

3. Model and Priors
In this section, we use $Ga(\alpha, \beta)$ to denote a Gamma distribution with shape parameter $\alpha$ and scale parameter $\beta$. 
Similarly, \( N(\mu, \sigma^2) \) denotes a Normal distribution with mean \( \mu \) and variance \( \sigma^2 \); while \( t_4(\mu, \sigma^2) \) denotes a \( t \)-distribution with four degrees of freedom, mean \( \mu \), and variance parameter \( \sigma^2 \). The exact density of the distributions used is given in the Web Appendix.

3.1 Modeling a Single Binding Event

Having segmented the read data into candidate regions (Section 2), we first assume that each region contains a single transcription factor binding site. An extension to the case of multiple binding sites is treated below. Let us denote by \( f_i \) and \( r_j \) the \( i \)th forward and \( j \)th reverse read positions in a given region, with \( i = 1, \ldots, n_f \) and \( j = 1, \ldots, n_r \). Note that the number of forward reads, \( n_f \), and reverse reads, \( n_r \), will typically vary between candidate regions. We jointly model the forward and reverse read positions as:

\[
f_i \sim t_4(\mu - \delta/2, \sigma_f^2) \quad \text{and} \quad r_j \sim t_4(\mu + \delta/2, \sigma_r^2),
\]

where \( \mu \) represents the binding site position; \( \delta \) is the distance between the maxima of the forward and reverse read position densities, which corresponds to the average DNA fragment length for the binding event in question; and \( \sigma_f \) and \( \sigma_r \) measure the corresponding variability in DNA fragment end positions. Note that this approach differs from that typically used in analyzing sequencing data, in that we do not model the sequence counts, but rather the distributions of the fragment end positions, for which we have more prior information. Figure 2a shows a candidate region with one binding event, along with the corresponding PICS parameter estimates.

3.2 Modeling Multiple Binding Events

We use mixture models to address the possibility that the sets of forward and reverse reads within a candidate region were generated by multiple closely spaced binding events. We model the forward and reverse read positions using \( t \)-mixture distributions:

\[
f_i \sim \sum_{k=1}^{K} \frac{w_k}{4} t_4(\mu_{fk}, \sigma_{fk}^2) \sim g_f(f_i | \mu, \sigma_f),
\]

\[
r_j \sim \sum_{k=1}^{K} \frac{w_k}{4} t_4(\mu_{rk}, \sigma_{rk}^2) \sim g_r(r_j | \mu, \sigma_r),
\]

where \( \mu_{fk} = \mu_k - \delta_k/2 \) and \( \mu_{rk} = \mu_k + \delta_k/2 \) and \( \mu_k, \delta_k, \sigma_{fk}, \sigma_{rk} \) are defined as in (1), but have an index \( k \) that corresponds to the binding event \( k \), while \( w_k \) is the mixture weight of component \( k \), which represents the relative proportion of reads coming from the \( k \)th binding event. For simplicity we denote by \( g_f \) and \( g_r \) the resulting p.d.f. of the forward and reverse mixture distributions.

Figure 2b shows a candidate region that has two binding events, along with the corresponding PICS parameter estimates.

As described in (1–2), PICS uses \( t \)-distributions with four degrees of freedom to model local distributions of the forward and reverse read positions. While the \( t \) distribution is similar in shape to the Gaussian distribution, its heavier tails make it more robust to outliers (Lange, Little, and Taylor, 1989; Lo, Brinkman, and Gottardo, 2008). Since a DNA fragment should contribute a forward read or a reverse read with equal probability, we use the same mixture weight \( w_k \) for both forward and reverse distributions. Finally, to accommodate possible biases that result in asymmetric forward and reverse peaks, we use different forward and reverse variance parameters \( \sigma_{fk}^2 \) and \( \sigma_{rk}^2 \).

3.3 Prior Distributions

Typically, the library construction process makes prior information available for the length of the DNA fragments. We can use a Bayesian approach to take advantage of this information by allowing the \( \delta_k \) ’s for all binding sites to derive from a common prior fragment length distribution. Similarly, we can also put a common prior distribution on \( \sigma_{fk}^2 \) and \( \sigma_{rk}^2 \), which allows us to incorporate prior information about the variability of the

Figure 2. Predicted binding events in two candidate regions in the GABP data. PICS detected one binding event in the region in (a) and two binding events in the region in (b). The observed data of forward and reverse strand aligned reads are shown by > and < arrowheads, respectively. Mappability profiles are shown as black/white lines, in which the white intervals show nonmappable regions. In (a), the distribution of reverse read positions was biased by a region with low mappability. This figure appears in color in the electronic version of this article.
DNA fragment length across a set of binding events, and to regularize variance estimates when few reads are available. For computational convenience, we use a Normal–Gamma conjugate prior, given by

$$
\begin{align*}
\left(\sigma^2_{f,k} + \sigma^2_{r,k}\right) & \sim \mathcal{G}(\alpha, \beta) \quad \text{and} \\
\left(\delta_k \mid \sigma_{f,k}^2, \sigma_{r,k}^2\right) & \sim \mathcal{N}(\xi, \rho^{-1}/(\sigma_{f,k}^2 + \sigma_{r,k}^2)),
\end{align*}
$$

(3)

where $\xi$ represents our best prior guess about the mean fragment length across all binding sites, and $\rho$ controls the spread around this guess. In the analysis of experimental data reported here, we chose $\alpha = 20$, $\beta = 40,000$, $\xi = 175$, and $\rho = 1$. These values were based on knowing that DNA fragments should be on the order of 50–250 bp after gel size selection for both data sets considered (Valouev et al., 2008; Zhang et al., 2008), and resulted in a fairly diffuse prior for the DNA fragment length, with a mean of 175 bp and a standard deviation of approximately 50 bp.

### 3.4 Summary of Model and Priors

In this section, we summarize our model and priors using a graphical representation (Figure 3) and the compact set of equations given by (4–9).

In the PICS model, the forward and reverse read positions are linked by the shared parameters, which are denoted by the equations given by (4–9).

3.5 Accounting for Missing Reads

Building on (1–2), we now consider the case where some reads are missing due to one or more nonmappable regions intersecting a candidate region. For clarity, we again focus on a single candidate region, whose genomic extent is denoted by $I$. For each chromosome, a mappability profile for a specific read length (e.g., 36 bp) consists of a vector that lists an estimated read mappability “score” for each base pair in the chromosome (Robertson et al., 2008). A score of one at a genomic position means that we should be able to uniquely align a read that overlaps that position, while a score of zero indicates that no read of that length should be uniquely alignable at that position. As noted above, typically only reads that map to unique genomic locations are retained for analysis. For convenience, and because transitions between mappable and nonmappable regions are often much shorter than the regions, we compactly summarize each chromosome’s mappability profile as a disjoint union of nonmappable intervals that specify only zero-valued profile regions (Figure 2).

Let us assume that a candidate region intersects one or more of these intervals. We can write $I = \bigcup_{l=1}^{L} I_l$, where $I_l = [a_l, b_l]$ denotes the $l$th nonmappable interval, with $l = 1, \ldots, L$; and $I_D$ denotes the union of intervals that have high mappability, and so should have no missing reads. In $I_l$, the $f_l$ ($i = 1, \ldots, n_{fl}$) and $r_l$ ($j = 1, \ldots, n_{rl}$) denote independent forward read positions and $n_{rl}$ independent reverse read positions. Note that only the quantities with $l = 0$ are

$$
\begin{align*}
\mu_{f,k} &= \mu_k - \delta_k / 2, & \mu_{r,k} &= \mu_k + \delta_k / 2 \\
(d_i \mid U_{dik} = u_{dik}, z_{dik} = 1, \mu_k, \delta_k) & \sim \mathcal{N}\left(\mu_{dik}, \sigma_{dik}^2\right) \\
(U_{dik} \mid z_{dik} = 1) & \sim \mathcal{G}(v/2, \nu/2) \\
(z_{f,i}, z_{r,j} & \sim \mathcal{M}(w) \\
(\sigma^2_{f,k} + \sigma^2_{r,k}) & \sim \mathcal{G}(\alpha, \beta) \\
(\delta_k \mid \sigma^2_{f,k}, \sigma^2_{r,k}) & \sim \mathcal{N}(\xi, \rho^{-1}/(\sigma^2_{f,k} + \sigma^2_{r,k})),
\end{align*}
$$

(4–9)

Figure 3. Directed acyclic graph representation of the PICS model. Reads aligned to forward and reverse strands are denoted by $f$ and $r$. Dashed rectangles represent fixed hyper-parameters. Gray-filled circles indicate missing data that are incorporated into the model to facilitate inference via the EM algorithm while plain circles represent unknown variables that need to be estimated. Plain squares represent the data (aligned forward and reverse read positions). Finally, thick arrows indicate deterministic relationships, while thin arrows indicate stochastic relationships. The index $k$ refers to the mixture component $k$. 

3.4 Summary of Model and Priors

In this section, we summarize our model and priors using a graphical representation (Figure 3) and the compact set of equations given by (4–9).

In the PICS model, the forward and reverse read positions are linked by the shared parameters, which are denoted by the nodes in the figure’s center column. The left and right sides of the figure, respectively, show the information related to the forward and reverse strand reads. Given the average fragment length $\delta_k$ and the binding site positions $\mu_k$, the centers of the forward and reverse read position density distributions can be calculated by (4). Conditional on the binding event memberships (the $z_{f,i}$’s), the read positions follow a Normal–Gamma compound distribution given by (5–6). In other words, conditionally on the z’s, the read positions follow a t-distribution with $\nu$ degrees of freedom. The unknown membership indicator $z_{f,i}$ (respectively $z_{r,j}$) indicates which binding event a forward read $i$ (respectively reverse read $j$) comes from; both $z_{f,i}$ and $z_r$ share the same multinomial distribution given by (7).

As a consequence, the marginal distribution of the read positions is a mixture of t-distributions whose weights are given by the parameters of the multinomial distribution in (7). The introduction of the missing data (the $z$’s and the $\mathbf{u}$’s) allows us to perform parameter estimation via an EM algorithm as described in Section 4. Finally, the average fragment length parameter $\delta_k$ and peak variances $\sigma_{f,k}^2, \sigma_{r,k}^2$ are assigned a Normal–Gamma conjugate prior, given by (8–9).

$$
\begin{align*}
\mu_{f,k} &= \mu_k - \delta_k / 2, & \mu_{r,k} &= \mu_k + \delta_k / 2 \\
(d_i \mid U_{dik} = u_{dik}, z_{dik} = 1, \mu_k, \delta_k) & \sim \mathcal{N}\left(\mu_{dik}, \sigma_{dik}^2\right) \\
(U_{dik} \mid z_{dik} = 1) & \sim \mathcal{G}(v/2, \nu/2) \\
(z_{f,i}, z_{r,j} & \sim \mathcal{M}(w) \\
(\sigma^2_{f,k} + \sigma^2_{r,k}) & \sim \mathcal{G}(\alpha, \beta) \\
(\delta_k \mid \sigma^2_{f,k}, \sigma^2_{r,k}) & \sim \mathcal{N}(\xi, \rho^{-1}/(\sigma^2_{f,k} + \sigma^2_{r,k})).
\end{align*}
$$

(4–9)
observed, while all others are unobserved random variables. Also, note that \( n_f, n_r, \) and \( L \) will vary across candidate regions.

Based on (2), \( f_i \) and \( r_j, \) \( i = 1, \ldots, L, \) follow a truncated \( t \)-mixture model, which is given by \( g_j \) and \( g_i, \) truncated on \( I_l. \) The only information carried in the mappability profile is the locations and lengths of intervals \( I_l; \) these affect the estimation of the model parameters shared between the observed and unobserved reads, i.e., \( w, \mu, \delta, \sigma_f, \) and \( \sigma_r, \) as described in Section 4.

4. Estimation and Inference

4.1 Parameter Estimation Using the EM Algorithm

Given the conjugacy of the prior chosen, an expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) can be derived to find the maximum a posteriori estimates (MAP) of the unknown parameter vector \( \Theta = (\theta_1, \ldots, \theta_K) \) where \( \theta_k = (w_k, \mu_k, \delta_k, \sigma^2_f, \sigma^2_r). \) Our algorithm is similar to those used in \( t \)-mixture models and Bayesian regularization for mixture models (Peel and McLachlan, 2000; Fraley and Raftery, 2007). In the presence of missing reads, we use an algorithm similar to that developed by McLachlan and Jones (1998) for grouped and truncated data. The algorithm is described in detail in the Web Appendix.

4.2 Inference and Detection of Binding Sites

Choosing the number of binding events in each region. In practice, \( K, \) the number of binding events, is unknown and needs to be estimated. For each candidate region, we fit our PICs model with \( K \) taking values from 1 to 15, and select the value of \( K \) that has the largest BIC (Schwarz, 1978), which in our case is given by

\[
BIC = 2Q(\Theta = \hat{\Theta} \mid \hat{\Theta}) - (5K - 1)\ln(n_f + n_r),
\]

where \( \hat{\Theta} \) is the final estimate for the parameters \( \Theta, \) and \( Q \) is the log-likelihood as defined in the Web Appendix.

Uncertainty of parameter estimates. It is useful to extend the point estimates for the parameters of interest, \( \mu \) and \( \delta, \) by deriving measures of uncertainty for them. Within our framework of mixture models with truncated data, we use the approach described in McLachlan and Krishnan (1997) to derive an approximation of the observed information matrix for the parameters. From the observed information matrix, we can then obtain approximate standard errors for both \( \hat{\mu} \) and \( \hat{\delta}. \)

We can use these standard errors to define the starts and ends of binding event neighborhoods, filter out noisy predictions, and estimate confidence intervals for binding site positions.

Binding event neighborhoods. Because PICs models local concentrations of bidirectional read positions, we can define "high confidence" neighborhoods whose extents are given by the maxima of forward and reverse read position densities. Using our PICs parameters, and taking into consideration the standard errors of the estimates, for a given binding event this neighborhood is defined as the interval \( \hat{\mu} \pm \frac{\hat{\delta}}{2}, \) extended by three standard errors on each side (i.e., \( \text{SE}(\hat{\mu} - \frac{\hat{\delta}}{2}) \) for the left limit and \( \text{SE}(\hat{\mu} + \frac{\hat{\delta}}{2}) \) for the right limit). These high confidence neighborhoods can define "enriched" regions in a file that can be visualized in a genome browser (Kuhn et al., 2009).

Peak merging and filtering. We use BIC to estimate the number of binding events within each candidate region. While BIC is well suited for selecting the number of mixture components required to estimate an underlying probability density, it sometimes overestimates the number of components (Baudry et al., 2010). In our case, this can occur when a candidate region contains hundreds of reads. To address this, we merge peaks that have overlapping binding event neighborhoods. The parameters of the merged peaks are obtained by moment-matching conditions (see the Web Appendix). Since the combined parameter estimates \( \hat{\mu} \) and \( \hat{\delta} \) are linear combination of the original ones, the original information matrix can be used to recompute the standard errors. For the GABP and FOXA1 data described below, this approach merged less than 1% of the binding events.

In addition to merging overlapping events, we also filter out binding events that have noisy (i.e., poorly estimated) or atypical parameter estimates, as these could affect downstream analysis. Specifically, we remove binding events that fail to satisfy any of the following three criteria: (i) \( \text{SE}(\hat{\mu}) < 50, \) (ii) \( 50 < \delta < 200, \) and (iii) \( \hat{\delta} < 150, d = \{f, r\}. \)

Essentially, (i) filters events that have poor binding site position estimates, while (ii) and (iii) filter events that have atypical DNA fragment distributions, given the fragment size selection applied in library construction, and so are likely to be artifactual regions (e.g., events that have high fractional overlaps with simple tandem repeats; Johnson et al., 2008).

Scoring and ranking binding events. In order to identify and rank a statistically meaningful subset of binding events, we define an enrichment score for each binding event. For a given event, we define \( F_{\text{ChIP}} \) and \( R_{\text{ChIP}}, \) the number of observed forward/reverse ChIP ("treatment") read positions that fall within the 90% contours of the forward/reverse read position densities, i.e., within \( \hat{\mu} \pm c \cdot \hat{\sigma} \) where \( d = \{f, r\} \) and \( c \approx 2.13 \) is the 95% quantile of the \( t_d \) distribution. For each binding event, we define the enrichment score as \( O_{\text{ChIP}} = F_{\text{ChIP}} + R_{\text{ChIP}}, \) which is an estimate of the observed number of DNA enriched fragments falling within the binding events after removing outliers. When a control sample is available, we also define \( O_{\text{cont}} = F_{\text{cont}} + R_{\text{cont}}, \) by computing the number of observed forward/reverse reads in the control sample that fall within the 90% contour of the forward/reverse read position densities estimated from the ChIP sample. Using this information, we define an enrichment score for the treatment relative to the control as \( S = (N_{\text{control}}/N_{\text{ChIP}}) \cdot O_{\text{ChIP}}/(O_{\text{cont}} + 1), \) where the addition of the constant one prevents a division by zero, and \( N_{\text{control}} \) (respectively \( N_{\text{ChIP}} \)) denote the total read count in control sample (respectively IP sample). The scaling of the enrichment score by \( N_{\text{control}}/N_{\text{ChIP}} \) accounts for the control and ChIP samples having different numbers of reads (sequence depth).

False discovery rate. Given control data, we can estimate the false discovery rate as a function of the enrichment score. We do this by repeating the analysis after swapping the control sample for the ChIP sample and recomputing our enrichment scores (i.e., fit PICs to the control data, and define control-over-IP enrichment scores), which we call "null"
enrichment scores and denote by $S_0$. Then the FDR, as a function of the threshold value $q$, can be computed as:

$$FDR(q) = \frac{\#S_0 : S_0 > q}{\#S : S > q},$$

that is, as the ratio of the number of null enrichment scores greater than $q$, divided by the number of observed enrichment scores greater than $q$.

5. Application to Experimental Data

We applied PICS to the two experimental data sets described in Section 2, obtaining 75,451 candidate regions and 86,262 binding events for the GABP data, and 51,843 candidate regions and 53,740 binding events for FOXA1 data. Web Figure 1 shows histograms of estimated average DNA fragment lengths for both data sets. For the FOXA1 data the estimated average fragment size was approximately 150 bp, consistent with Zhang et al. (2008); it was somewhat smaller for the GABP data. Web Figure 1 also shows that most of the regions had DNA fragment lengths between 50 and 200 bp, which supports our filtering atypical regions by this parameter.

Noting that some of the algorithms responded very differently in terms of estimated FDR (see Web Figure 2), we compared the methods by identifying conserved DNA sequence motifs in the 5,000 top-ranked predictions from each method, using 200-bp-wide regions that were centered on each method’s binding site estimates. For motif analysis, we used GADEM (Li, 2009), which can process large sets of ChIP-seq data. Web Figure 1 also shows histograms of estimated average DNA fragment lengths for both data sets. For the FOXA1 data, the estimated average fragment size was approximately 150 bp, consistent with Zhang et al. (2008); it was somewhat smaller for the GABP data. Web Figure 1 also shows that most of the regions had DNA fragment lengths between 50 and 200 bp, which supports our filtering atypical regions by this parameter.

Table 2 summarizes the results for 5,000 predictions for the GABP and FOXA1 data. Overall, in terms of occurrence rate, PICS performed better than the other methods. MACS and USeq were relatively close to PICS, while QuEST and cisGenome performed comparably, but were less effective than PICS, MACS, and USeq. For spatial error, PICS and MACS were most effective, followed closely by USeq, and then QuEST and cisGenome.

As stated in Section 3, we use mixture models to address the possibility that the sets of forward and reverse reads within a candidate region were generated by multiple binding events. In order to assess how PICS’ mixture model can detect multiple binding sites within a candidate region, we used our predicted transcription factor binding motifs for the top-ranked 5,000 PICS predictions for the GABP and FOXA1 data. In each case, we determined the percentage of binding events from single- and multiple-component candidate regions (i.e., regions with multiple predicted events) that could be associated with at least one motif site. Table 1 shows these results as a function of the number of PICS predictions (i.e., mixture components) in a candidate region. Four times more GABP regions than FOXA1 regions had two components (940 vs. 208), and seven times more had at least three components (236 vs. 33), even though candidate regions had comparable sizes. Because shorter DNA fragments should support detecting more adjacent binding sites, these differences may be related to the smaller average fragment size estimated for the GABP data. For both data sets, the percentage of binding events that was associated with a predicted binding motif was relatively insensitive to the number of predictions in a region. Because cells can use multiple closely spaced transcription factor binding sites to establish progressive expression responses to cellular signals, we also assessed how effectively PICS can detect closely adjacent binding sites. To evaluate PICS and all other methods compared here, we generated another table, but this time considered predicted binding events that had at least one other event within a fixed distance $d$. Table 2 summarizes the results for $d = 250, 500$, and 1000 bp. For these data, PICS, cisGenome, and USeq were the most effective at identifying proximal binding events, and large fractions of these events were associated with a predicted motif site. While cisGenome and USeq also predicted large numbers of proximal binding events, a larger fraction of those reported by PICS were associated with predicted binding motifs. In comparison, MACS returned the lowest number of proximal binding events, which suggests that MACS is less effective in discriminating proximal binding events, even though it performs relatively well in terms of the overall number of predictions that are associated with a motif (Figure 4). These results suggest that our mixture model was effective in distinguishing biologically meaningful proximal binding events.

As described in Section 4, PICS can compute approximate standard errors for its model parameter estimates. In
Figure 4. Motif occurrence rate and spatial error (see text) for GABP and FOXA1 data, as a function of region enrichment rank, for the 5000 top-ranked regions for each method. This figure appears in color in the electronic version of this article.

Table 1
Number of 5000 top-ranked PICS predictions from GABP and FOXA1 data in candidate regions that had 1, 2, or at least 3 predicted binding events (i.e., mixture components). The first row gives the number of binding events in each category, while the second row gives the percentage of predicted events that could be associated with an expected motif. For example, 80% of the 940 binding events in two-component GABP regions could be associated with a predicted GABP site.

| Number of components in region | GABP | FOXA1 |
|-------------------------------|------|-------|
|                               | 1    | 2     | 3+   | 1   | 2   | 3+   |
| Number of events              | 3824 | 940   | 236  | 4759| 208 | 33   |
| % of motifs                    | 81   | 80    | 77   | 86  | 88  | 94   |

particular, we can derive an approximate confidence interval for a given predicted binding event location as $\hat{\mu} \pm c \cdot SE(\hat{\mu})$, where $c$ is a constant to be chosen as a function of the motif coverage desired. For example, assuming that $\hat{\mu}$ is approximately Normal, $c = 1.96$ should give us an approximate 95% confidence interval for our binding site position.

Using the set of motifs identified by GADEM, we evaluated the actual coverage of our confidence intervals for different values of $c$. Figure 5 shows the occurrence frequency of GABP motifs (left) and FOXA1 motifs (right) within $c \cdot SE(\hat{\mu})$ of peaks’ centers. Using three standard errors, the coverage was approximately 65% and 85% for the GABP and FOXA1 data. While these results indicate that the current version of PICS provides a capable modeling framework, they also suggest that certain biases remain unaddressed. For example, PICS models the binding site as the midpoint between the F/R peak summits, but fragmentation biases during sonication due to local chromatin structure could result in a binding site being asymmetrically positioned with respect to its associated DNA fragment densities. Locations where motifs fall outside event confidence intervals identify a subset of regions on which to focus ongoing work.

Finally, we evaluated the effect of the mappability profiles on the parameter estimates. We re-did the analysis while ignoring mappability, and compared the spatial error, i.e., the distance to the closest computationally verified binding
Table 2

Number of proximal binding events found in the 5000 top-ranked regions identified by each method in GABP and FOXA1 data, as a function of the motif “proximity” distance \(d\). The numbers in parentheses give the percentage of predicted binding events that could be associated with at least one predicted motif site. For example, in the GABP data, PICS identified 128 binding event locations that were within 250 bp of another location, and 86% of these predicted events could be associated with a motif.

| \(d\) (bp) | PICS  | QuEST | MACS  | cisGen | USeq  | PICS  | QuEST | MACS  | cisGen | USeq  |
|-----------|-------|-------|-------|--------|-------|-------|-------|-------|--------|-------|
| 250       | 128 (86) | 0     | 0     | 0      | 6 (90) | 2 (100)| 0     | 0     | 4 (75) | 4 (83) |
| 500       | 437 (84) | 2 (50) | 18 (83) | 16 (57) | 106 (77) | 62 (85)| 18 (72) | 18 (100) | 34 (82) | 70 (74) |
| 1000      | 517 (81) | 405 (63) | 54 (81) | 76 (63) | 269 (72) | 108 (87)| 74 (76) | 54 (88) | 105 (85) | 149 (77) |

Figure 5. Fraction of predicted binding events with a GABP (a) or FOXA1 (b) motif site within \(c \cdot \text{SE}(\hat{\mu})\) of the predicted event location, \(\hat{\mu}\), as a function of region enrichment rank. This figure appears in color in the electronic version of this article.
has support over \((\mu_k, \infty)\), while the forward read position distribution is defined as the mirror image of the same Gamma about \(\mu_k\).

### 6.2 Simulation Results

We analyzed each data set with PICS, MACS, cisGenome, QuEST, and USeq. In each case, we summarized the results with a receiver operating characteristic (ROC) curve, which shows the relationship of sensitivity to specificity when varying the cutoff for each method, and a plot of the nominal FDR against the true FDR. Figure 6 shows that the results for the simulation scenarios were similar to those for experimental data. In this figure, we only show the ROC curves in the region where 1-specificity is between 0 and 0.2, since that is the region of most interest. The full ROC curves (from 0 to 1) are shown in Web Figure 5. PICS was slightly better than MACS and USeq under scenarios 1 and 3, whereas under scenario 2 USeq was slightly better than PICS, though the differences were small. FDR results show larger differences; PICS was closest to the expected \(y = x\) line, while MACS, QuEST, and USeq tended to underestimate the FDR and cisGenome tended to overestimate it. Note that in the results presented here we only used a single simulation because some of the software (e.g., USeq, cisGenome) require manual user input at different analysis stages and thus could not be run in a fully automated way. This said, the results for PICS and MACS were consistent over a few repeats (data not shown). Finally, even though PICS’ estimation of the FDR is good, it is still slightly overestimated at low values, which could be due to the slight model misspecification caused by injecting control reads into the IP sample. Overall, this suggests that FDR estimates should be treated carefully for all methods.

As with the FOXA1 and GABP data, we also compared the methods in terms of their spatial error (see Web Figure 3). Again PICS was the most accurate, followed closely by USeq, MACS, and QuEST. In addition, we also assessed the performance of our measures of uncertainty by calculating the actual coverage rate of our approximate 95% confidence intervals.
interval calculated as $\pm 1.96 \cdot \text{se}(\hat{\mu})$. Web Figure 4 shows the coverage rate as a function of peak rank. The coverage rate was between 90% and 95% even under misspecified models, which suggests that our approach is reasonably accurate and robust to misspecification.

Finally, we reassessed the ability of each method to detect closely adjacent binding sites. Since our simulation scheme included closely adjacent binding sites that were generated in spike-in regions, we compared the number of sites predicted by each method to the true number of sites in each region. Web Tables 1–3 show the results for each method and simulation scenario. As expected, PICS performed best at discriminating closely adjacent binding sites, and did this even under misspecification. This suggests that BIC is an effective metric for estimating the number of binding events.

7. Discussion

We have developed PICS, a probabilistic framework for detecting transcription factor binding events from ChIP-seq experiments. The approach integrates a number of important factors in interpreting aligned read data, including using prior information for input DNA fragment lengths and correcting for reads that are missing due to genome repetitiveness. Working with two published ChIP-seq data sets from human cell lines, we compared PICS to four alternative analysis methods. While additional methods are available (e.g., Fejes et al., 2008; Kharchenko et al., 2008; Rozowsky et al., 2009), the four methods that we used have been shown to perform well, and so offer reasonable performance baselines from a range of algorithms. For both experimental data sets, the binding events predicted by PICS were the most consistent with computationally identified motif sites. Our simulation study confirmed PICS’ effectiveness and also showed that its model is robust to misspecifications.

To address read “outliers” due to biological and technical biases, we used a $t$-distribution as a robust alternative to the Gaussian distribution. We fixed the degrees of freedom parameter at four because this value has been shown to work well in many applications (Lange et al., 1989). To confirm that this was an appropriate choice, we estimated the degrees of freedom for each candidate region and found that it was less than 10 in about 20% of all regions. This suggests that many of the regions contain read outliers and would be poorly modeled by a Gaussian distribution. In addition, many of the candidate regions have very few reads, and it is possible that the number of reads was not large enough to properly estimate the degrees of freedom. We believe that with more sequence reads, the overall percentage of candidate regions with low degrees of freedom would have been even larger. For that reason, and because estimating the degrees of freedom is computationally intensive, we used a fixed four degrees of freedom.

We showed that PICS’ mixture model addresses multiple adjacent enrichment events, and can fit a different DNA fragment length value for each binding event in a mixture. Our estimation of the number of binding events in a region is based on the Bayesian information criteria. Although regularity conditions required for consistency of the BIC estimate do not hold for mixture models, there is considerable theoretical and practical support for its use in this context (Roeder and Wasserman, 1997; Fraley and Raftery, 1998). Our simulation study confirmed this, and suggests that our estimation of the number of binding events performs relatively well. In our BIC estimation, we allowed the mixture model to detect up to 15 components per candidate region, but this limit can be readily adjusted.

Because it is based on mixture models and accounts for missing reads, PICS is computationally intensive. The results reported here were generated with an implementation of PICS that was written in the R programming language (Ihaka and Gentleman, 1996). Processing a $\sim 10M$ read data set required an average computing time of 30 minutes on a machine with dual quad-core 3 GHz CPUs. We believe this to be acceptable, but if speed were an issue, PICS could be rewritten in C. PICS will be made freely available via Bioconductor (Gentleman et al., 2004).

Paired-end (PE) data should resolve a subset of read alignments that would be nonunique in SE data, and offer direct information on DNA fragment lengths that is not available for single-end (SE) reads. However, PE experiments require more input DNA and more complex library preparation, and double the sequencing machine time; to our knowledge, all published short read ChIP-seq data are SE rather than PE. We anticipate that PICS’ probabilistic approach will remain useful for PE data, where the defined fragment lengths should simplify the modeling framework.

As a first step in implementing a probabilistic approach for ChIP-seq data, we have shown how to incorporate prior information about the DNA fragment lengths using a Bayesian approach. We can extend the PICS framework to include more types of prior information. For example, we could place a prior distribution on $\mu$, the binding site position, and could include in this information about nucleosome occupancy and computationally derived motifs. Such extensions should allow us to improve the detection of biologically relevant binding sites. More generally, we anticipate that extended probabilistic methods for ChIP-seq will contribute to mechanistic biological insights by offering principled ways for addressing backgrounds, noise and biases, and for integrating diverse types of biological information. For example, histone modifications are an important part of cellular biology, and their aligned reads present a wider range of nucleosome-based density distributions than the localized transcription factors that PICS currently addresses. As part of ongoing work in this area, we are extending PICS to infer nucleosome positioning from histone modification data.

8. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 3, 4, 5 and 6 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

Acknowledgements

We gratefully acknowledge Inanc Birol for discussions related to read mappability. We thank Martin Hirst, Anthony Fejes, Misha Bilenky, Nina Thiessen, three anonymous referees, the editor, and associate editor for suggestions that clearly improved an earlier draft of the article. This
research is supported by an NSERC Discovery Grant (RG and XZ).

REFERENCES

Baudry, J. P., Raftery, A. E., Celeux, G., Lo, K., and Gottardo, R. (2010). Combining mixture components for clustering. To appear in Journal of Computational and Graphical Statistics.

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.

Cicatiello, L., Addeo, R., Sasso, A., Altucci, L., Petrizzi, V. B., Borgo, R., Cancemi, M., Caporali, S., Caristi, S., Scafoglio, C., Teti, D., Bresciani, F., Perillo, B., and Weisz, A. (2004). Estrogens and progesterone promote persistent CCND1 gene activation during G1 by inducing transcriptional derepression via c-Jun/c-Fos/estrogen receptor (progesterone receptor) complex assembly to a distal regulatory element and recruitment of cyclin D1 to its own gene promoter. Molecular and Cellular Biology 24, 7260–7274.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B 39, 1–38.

D’haeseleer, P. (2000). What are DNA sequence motifs? Nature Biotechnology 24, 423–425.

Eeckhoute, J., Carroll, J. S., Geistlinger, T. R., Torres-Arzayus, M. I., and Brown, M. (2006). A cell-type-specific transcriptional network required for estrogen regulation of cyclin D1 and cell cycle progression in breast cancer. Genes and Development 20, 2513–2526.

Fejes, A. P., Robertson, A. G., Bilenky, M. B., Varhol, R., Bainbridge, M. N., and Jones, S. J. (2008). FindPeaks 3.1: A java application for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24, 1729–1730.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. Computer Journal 41, 578–588.

Fraley, C. and Raftery, A. E. (2007). Bayesian regularization for Normal mixture estimation and model-based clustering. Journal of Classification 24, 155–181.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dubois, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 5, R80.1–R80.16.

Gottardo, R., Li, W., Johnson, W. E., and Liu, X. S. (2008). A flexible and powerful Bayesian hierarchical model for ChIP-chip experiments. Biometrics 64, 468–478.

Holt, R. A. and Jones, S. J. M. (2008). The new paradigm of flow cell sequencing. Genome Research 18, 839–848.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 5, 299–314.

Ji, H., Jiang, H., Ma, W., Johnson, D. S., Myers, R. M., and Wong, W. H. (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nature Biotechnology 26, 1293–1300.

Johnson, D. S., Li, W., Gordon, D. B., Blattner, F. C., Curry, A., Ghosh, J., Brizuela, L., Carroll, J. S., Brown, M., Flick, P., Koch, C. M., Dunham, I., Bieda, M., Xu, X., Farhmann, P. J., Kapranov, P., Nix, D. A., Gingeras, T. R., Zhang, X., Holster, H., Jiang, N., Green, R. D., Song, J. S., McCuine, S. A., Anton, E., Nguyen, L., Trinklein, N. D., Ye, Z., Ching, K., Hawkins, D., Ren, B., Scacheri, P. C., Rozovsky, J., Karpikov, A., Euskirchen, G., Weissman, S., Gerstein, M., Snyder, M., Yang, A., Mogiaderi, Z., Hirsch, H., Shulha, H. P., Fu, Y., Weng, Z., Struhl, K., Myers, R. M., Lieb, J. D., and Liu, X. S. (2008). Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. Genome Research 18, 393–403.

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

Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature Biotechnology 26, 1351–1359.

Kuhn, R. M., Karolchik, D., Zweig, A. S., Wang, T., Smith, K. E., Rosenblloom, K. R., Rhead, B., Raney, B. J., Pohl, A., Pheasant, M., Meyer, L., Hsi, F., Hinrichs, A. S., Harte, R. A., Giardine, B., Fujita, P., Diekhans, M., Dresser, T., Clawson, H., Barber, G. P., Haussler, D., and Kent, W. J. (2009). The UCSC genome browser database: Update 2009. Nucleic Acids Research 37, D755–D761.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. Journal of the American Statistical Association 84, 881–896.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Li, L. (2009). GADEM: A genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. Journal of Computational Biology 16, 317–329.

Lo, K., Brinkman, R. R., and Gottardo, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. Cytometry A 73A, 321–332.

Lupien, M., Eeckhoute, J., Meyer, C. A., Wang, Q., Zhang, Y., Li, W., Carroll, J. S., Liu, X. S., and Brown, M. (2008). FoxAI translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell 132, 958–970.

Mahony, S., Auron, P. E., and Benos, P. V. (2007). DNA familial binding profiles made easy: Comparison of various motif alignment and clustering strategies. PLoS Computational Biology 3, e61.

McLachlan, G. J. and Jones, P. N. (1998). Fitting mixture models to grouped and truncated data via the em algorithm. Biometrics 44, 571–578.

McLachlan, G. J. and Krishman, T. (1997). The EM Algorithm and Extensions, 2nd edition. Hoboken, New Jersey: Wiley.

Milde-Langosch, K. (2008). The Fos family of transcription factors and their role in tumourigenesis. European Journal of Cancer 44, 2449–2461.

Nix, D. A., Courdy, S. J., and Boucker, K. M. (2008). Empirical methods for controlling false positives and estimating confidence in ChIP-seq peaks. BMC Bioinformatics 9, 1–9.

Park, P. J. (2009). ChiP-seq: Advantages and challenges of a maturing technology. Nature Reviews Genetics 10, 669–680.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. Statistics and Computing 10, 339–348.

Robertson, A. G., Bilenky, M., Tam, A., Zhao, Y., Zeng, T., Thiessen, N., Cezard, T., Fejes, A. P., Wedrell, E. D., Cullum, R., Euskirchen, G., Kryzinswki, M., Birol, I., Snyder, M., Hoodless, P. A., Hirst, M., Marra, M. A., and Jones, S. J. (2008). Genome-wide relationship between histone H3 lysine 4 monom- and trimethylation and transcription factor binding. Genome Research 18, 1906–1917.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. Journal of the American Statistical Association 92, 894–902.
Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology* **27**, 66–75.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-seq data. *Nature Methods* **5**, 829–834.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based Analysis of ChIP-seq (MACS). *Genome Biology* **9**, R137.1–R137.9.

Received March 2009. Revised March 2010. Accepted March 2010.