Supplementary Materials: A Statistical Framework for the Analysis of ChIP-Seq Data

Pei Fen Kuan*

Departments of Statistics and of Biostatistics and Medical Informatics

Dongjun Chung

Departments of Statistics and of Biostatistics and Medical Informatics

Guangjin Pan†

Genome Center of Wisconsin and Morgridge Institute for Research

James A. Thomson

Department of Anatomy, Genome Center of Wisconsin,

Wisconsin National Primate Research Center and

Morgridge Institute for Research

Ron Stewart

Genome Center of Wisconsin and Morgridge Institute for Research

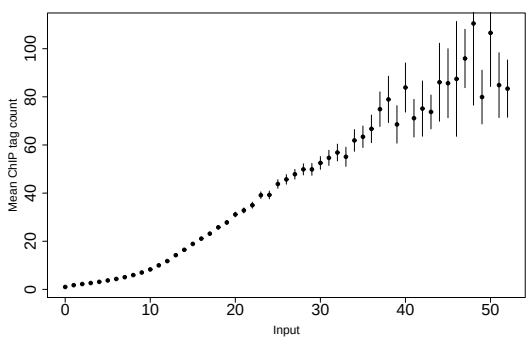
and Sündüz Keleş‡

Departments of Statistics and of Biostatistics and Medical Informatics

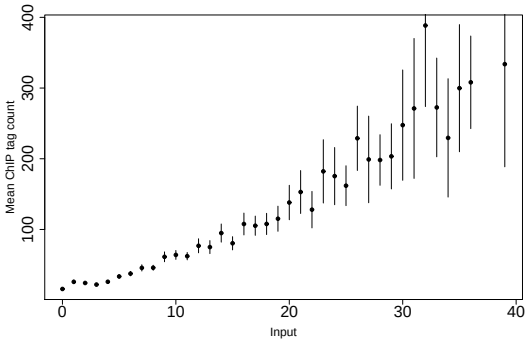
University of Wisconsin, Madison, WI 53706

May 9, 2011

1 Relationship between ChIP and Input tag counts



(a) ChIP vs. input (non-peaks)



(b) ChIP vs. input (peaks).

Figure 1: Mean ChIP vs. input DNA tag counts. Data displayed are from non-peak regions (Panel (a)) and peak regions (Panel (b)) of the STAT1 ChIP-Seq data and its matching input DNA control.

2 Adaptive gridding for strata-specific estimation of the non-enriched distribution

We implement an adaptive gridding scheme in Step 1(b) of our estimation procedure to account for strata with too few bins. These strata may result in poor estimates of a_i and μ_i. The basic idea is to "rescue" these strata to improve the estimates of β_0, β_M, $\boldsymbol{\beta}_{GC}$, and $\boldsymbol{\beta}_X$ in Step 1(d). In general, adaptive gridding rescues more points at boundaries of the M and GC ranges (small M values or small/large GC values), hence, provides more stable fits for strata at the boundaries. For expository purposes, we describe the adaptive gridding procedure for the one-sample model. Extension to the two-sample model is straightforward. The candidate grid sizes are

*Current Position: Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599

†Current Position: Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, 510530 China

‡Corresponding Author: keles@stat.wisc.edu

chosen to be 0.01, 0.02, 0.04, 0.10, 0.20, and 0.50 so that the grid size in the next iteration is approximately twice as large as that of the current iteration.

Let C denote the set of candidate strata at a given iteration. When the grid size is initialized at 0.01, C denotes all of the unique (M, GC) pairs, where M and GC are rounded to the nearest hundredth. Let E denote the final set of strata that will be utilized in Steps 1(d) and 1(e). Initially, set E is empty and all the strata are in set C. Adaptive gridding is an iterative application of the following procedure for grid sizes of 0.01, 0.02, 0.04, 0.10, 0.20, and 0.50:

1. For the current grid size, estimate a_i and μ_i for all the strata with size, i.e., number of bins in the strata, greater than N_{min} (50 for our case studies). These strata are then removed from set C and become part of set E.

2. For the remaining strata in C, increase the grid size and redefine strata by replacing the M and GC values of the bins within the strata by the median M and GC values of the strata. For example, consider the bins with $M_{i_1} \in (0.40, 0.42)$ and $GC_{i_2} \in (0.40, 0.42)$ as a result of increasing the grid size from 0.01 to 0.02. We reset their M and GC values as the median M and GC values of the strata defined by the $M_{i_1} \in (0.40, 0.42)$ and $GC_{i_2} \in (0.40, 0.42)$ pair. These redefined strata are then used to estimate a_i and μ_i.

3. Repeat the procedure until the grid size reaches 0.50. In the final iteration, move all the remaining strata from set C to set E.

This procedure enables use of all the data and prevents strata from having too few bins. Although it does not quarantee each final strata to have at least N_{min} bins, we observe this to be the case for the case studies presented in the paper.

3 Simulation studies for evaluating the estimation procedure of MOSAiCS

3.1 Evaluation of the estimation of the model parameters

In Section 3 of the main text, we proposed a computationally efficient estimation procedure for fitting the MOSAiCS model. Given the volume of high throughput sequencing data, computational speed is an important factor in developing methods. Here, we evaluate the performance of the proposed procedure with simulation studies. Let Y_j be the tag count for bin j and

$$P(Y_j = y) \;\; = \;\; \pi_0 P(N_j = y) + (1 - \pi_0) P(N_j + S_j + k = y),$$

where $N_j \sim NegBin(a, a/\mu_j)$, $\mu_j = \exp(\beta_0 + \beta_M \log_2(M_j + 1) + \beta_{GC} GC_j)$, $S_j \sim NegBin(b, c)$ and k is a fixed constant. Here, we consider the simpler functional form for the GC contribution; however, a smaller scale simulation study confirms that similar conclusions hold for the actual functional form of GC in the MOSAiCS model. In our modeling framework, we assume that 0, 1, and 2 counts are from the background distribution which implies that $k = 3$. In our first simulation, we generate $\beta_0 \sim U(-4, -2)$, $\beta_M \sim U(1, 3)$, $\beta_{GC} \sim U(1, 3)$, $\pi_0 \sim U(0.6, 0.99)$, $a \sim U(1, 5)$, $S \sim U(5, 10)$, $b \sim U(0.2, 2)$, $c = S/b$. We use the mappability and GC content from chromosome 18 of the dataset as our covariates, and compare the β estimates from our procedure to the β estimates obtained from the `glm.nb` function of the `MASS` library in `R`. This function fits a negative binomial family generalized linear model using a maximum likelihood approach. We apply `glm.nb` only to the subset of bins generated from the background distribution. However, our proposed algorithm is applied to the whole data to obtain the parameters of

both the background and enriched distribution simultaneously.

We let $\hat{\mu}_j^{MOSAiCS} = \exp(\hat{\beta}_0^{MOSAiCS} + \hat{\beta}_M^{MOSAiCS} \log_2(M_j + 1) + \hat{\beta}_{GC}^{MOSAiCS} GC_j)$ and $\hat{\mu}_j^{glm} = \exp(\hat{\beta}_0^{glm} + \hat{\beta}_M^{glm} \log_2(M_j + 1) + \hat{\beta}_{GC}^{glm} GC_j)$ denote the fitted means from MOSAiCS and `glm.nb`, respectively. Supplementary Figure 2 compares the mean square error for MOSAiCS ($\sum_j (\mu_j - \hat{\mu}_j^{MOSAiCS})^2/nsim$) and `glm.nb` ($\sum_j (\mu_j - \hat{\mu}_j^{glm})^2/nsim$), where $nsim$ is the number of simulations. We further define an analog of multiple r-squared value as in least squares regression: $R^2_{MOSAiCS} = 1 - \sum_j (\mu_j - \hat{\mu}_j^{MOSAiCS})^2 / \sum_j (\mu_j - \bar{\mu}_j)^2$ and $R^2_{glm} = 1 - \sum_j (\mu_j - \hat{\mu}_j^{glm})^2 / \sum_j (\mu_j - \bar{\mu}_j)^2$. The results over 100 simulations are summarized in Supplementary Table 1. Both the mean squared error and multiple r-squared comparisons of the two approaches exhibit little difference between them. Although `glm.nb` has slightly lower mean squared error than the estimation procedure of MOSAiCS, both mean squared errors are on average smaller than 0.05. As emphasized above, the `glm.nb` is fitted using the subset of true unbound bins, whereas MOSAiCS estimates the proportion of unbound bins π_0 and thus is expected to be less efficient. Despite this, MOSAiCS estimation procedure generates estimators comparable to the maximum likelihood estimators in `glm.nb`.

Table 1: Comparison of estimators of background model parameters by MOSAiCS and `glm.nb`

	MOSAiCS	`glm.nb`
Sim 1:$Median(R^2) \pm mad(R^2)$	0.9872± 0.0079	0.9999± 0.000038

NOTE: Results are reported over 100 simulation replicates.

Next, we evaluate the performance of MOSAiCS in estimating π_0, a, b and c. Supplementary Figure 3 compares the estimates of these parameters to their true values across 100 simulated datasets and provides strong support that the estimation procedure of MOSAiCS works well.

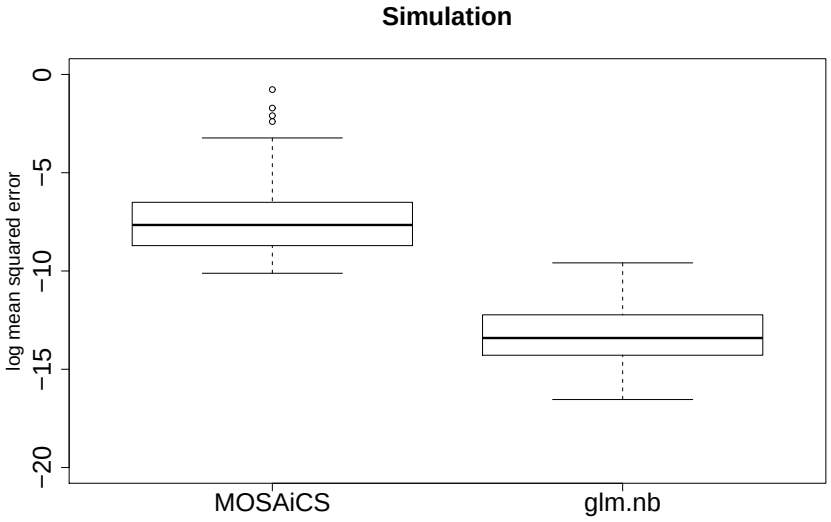


Figure 2: Mean squared error (log scale) comparison of MOSAiCS and `glm.nb`.

3.2 False discovery rate (FDR) under MOSAiCS

We evaluate the performance of MOSAiCS estimates in terms of the FDR control of the model. The first simulation setting is the same as the previous section where all the modelling assumptions of MOSAiCS are satisfied. For this evaluation, we stratify the simulations in terms of the proportion of unbound bins π_0 which is equivalent to the proportion of null hypotheses. Supplementary Figure 4 depicts empirical FDR from the MOSAiCS model against the nominal FDR and illustrates that for a wide range π_0 values, FDR is well controlled in the MOSAiCS model.

Next, we evaluate the consequences of violating the assumption that $k = 3$. Specifically, we generate data from $P(Y_j = y|Z_j = 1) = P(N_j + S_j + k = y)$ for $k = 0$, (i.e., 0, 1 and 2 counts bins can be generated from the enriched distribution) but enforce $k = 3$ during the estimation procedure. In Supplementary Figure 5, we compare the empirical FDR fom the model against the nominal FDR for different strata of π_0. As expected, there is an underestimation of FDR in all cases. However,

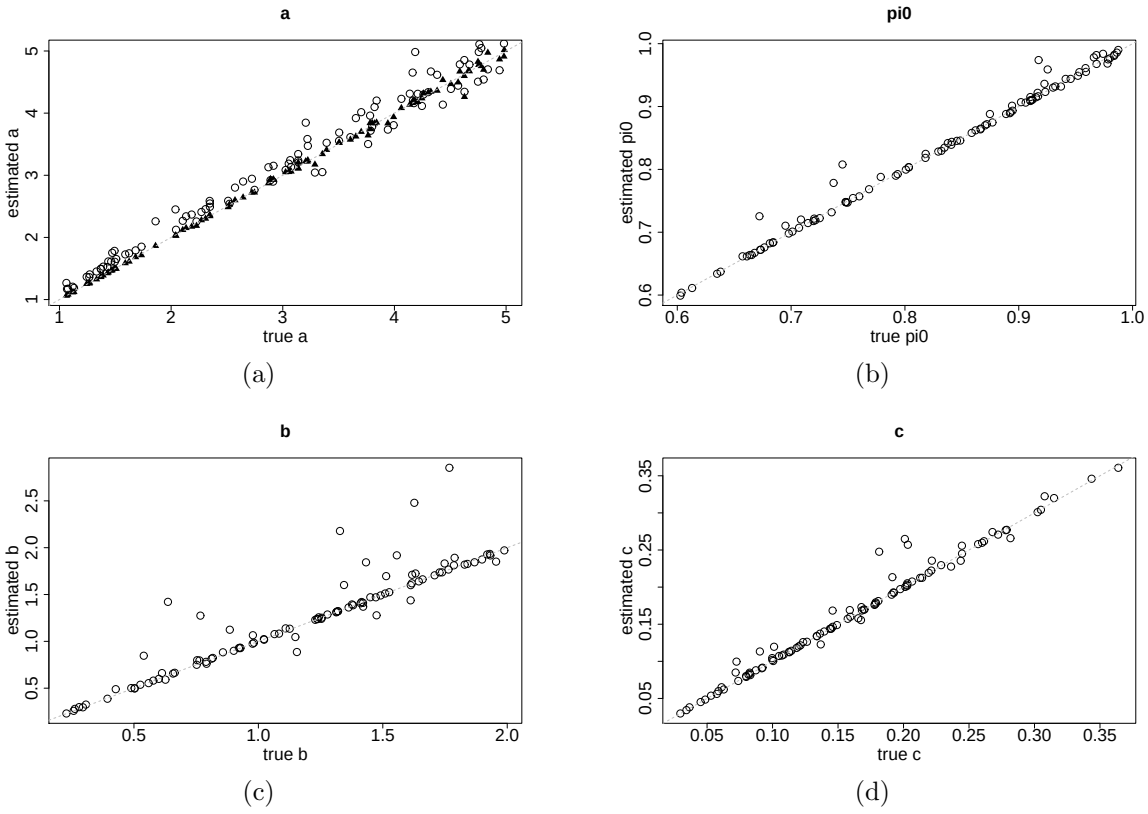
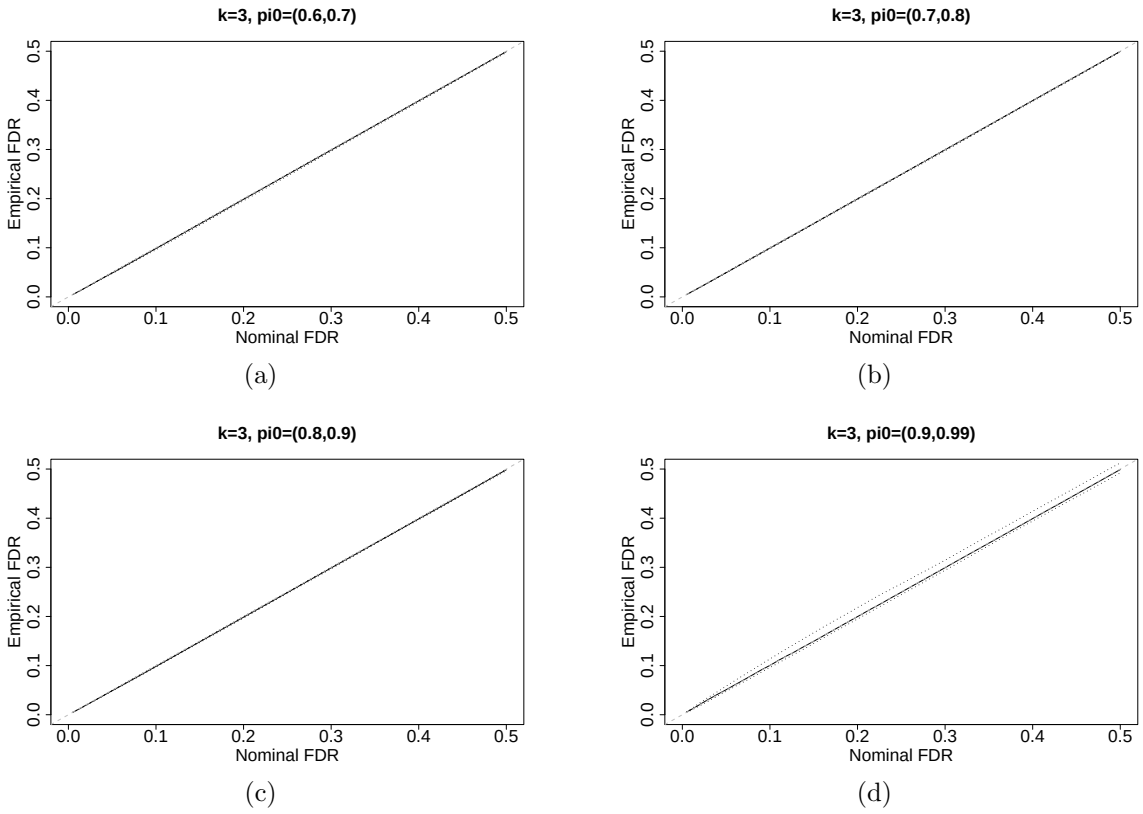


Figure 3: Performance of estimators of a, π_0, b, and c. Open circles in panels (a)-(d) denote estimators of a, π_0, b, and c versus their true values in 100 simulations. In panel (a), solid triangle compares the estimated a from `glm.nb` to its true value.



(a) (b) (c) (d)

Figure 4: Empirical versus nominal FDR for $k = 3$. Each panel compares the empirical FDR against nominal FDR for different π_0. Black solid line is the median empirical FDR, whereas dotted lines are the first and third quartiles of empirical FDRs over 100 simulations. The 45° line is depicted with a dashed gray line in each panel.

at $\pi_0 \in (0.9, 0.99)$, which is usually the case for transcription factor binding ChIP-Seq data, the empirical FDR is only slightly underestimated. We check that this range also covers binding of elonglation factors such as RNA Polymerase II based on data Pol II binding ChIP-Seq data in unstimulated HeLa S3 cells (Rozowsky et al., 2009). PeakSeq estimated π_0 for Pol II is between 0.986 and 0.990 across different chromosomes at FDR level of 0.05 both in one- or two-sample analysis using naked DNA or input DNA control.

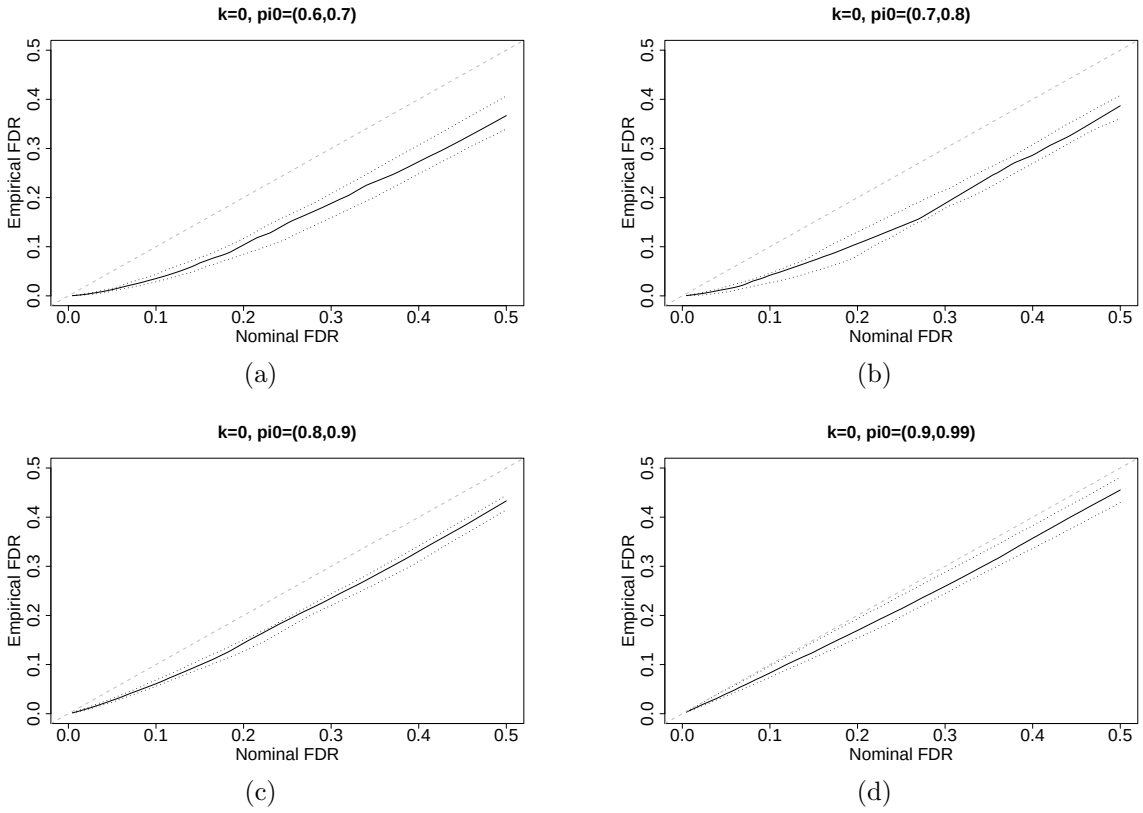


Figure 5: Empirical vs. nominal FDR when k is misspecified as 0 in the data generating mechanism. Each panel compares the empirical FDR against nominal FDR under model misspecification of k for different π_0. Black solid line is the median empirical FDR, whereas dotted lines are the first and third quartiles of the empirical FDR over 100 simulations.

4 Mappability in PeakSeq

We investigate how PeakSeq utilizes mappability by a simple computational experiment as follows. We create a pseudo mappability score for each chromosome by fixing the mappability scores for each 1 Mb to be a constant value (third quartile of the actual mappability scores for a chromosome). We run PeakSeq-1S twice (Run1 and Run2) with two different starting seeds for random permutation using the actual mappability score, and once using the pseudo mappability score for Pol II ChIP-Seq data (Rozowsky et al., 2009). In top panels of Supplementary Figure 6, we compare the window specific thresholds to detect bound regions by using each of these scores whereas in the bottom panels, we compare the number of peaks obtained by using actual and pseudo mappability scores. As evident from these plots, the results in PeakSeq-1S using either the actual or the pseudo mappability scores are very similar. That is, although PeakSeq-1S aims to incorporate mappability bias, the simulation based approach in PeakSeq-1S down-weighs the effect of mappability bias in a local region of 1 Mb, i.e., the variability in mappability across segments of 1 Mb is almost constant.

Supplementary Figure 7 compares the effect of mappability against the genomic window size in Pol II ChIP-Seq data. These plots show that the effect of mappability is only apparent using a shorter segment (e.g.,1 kb).

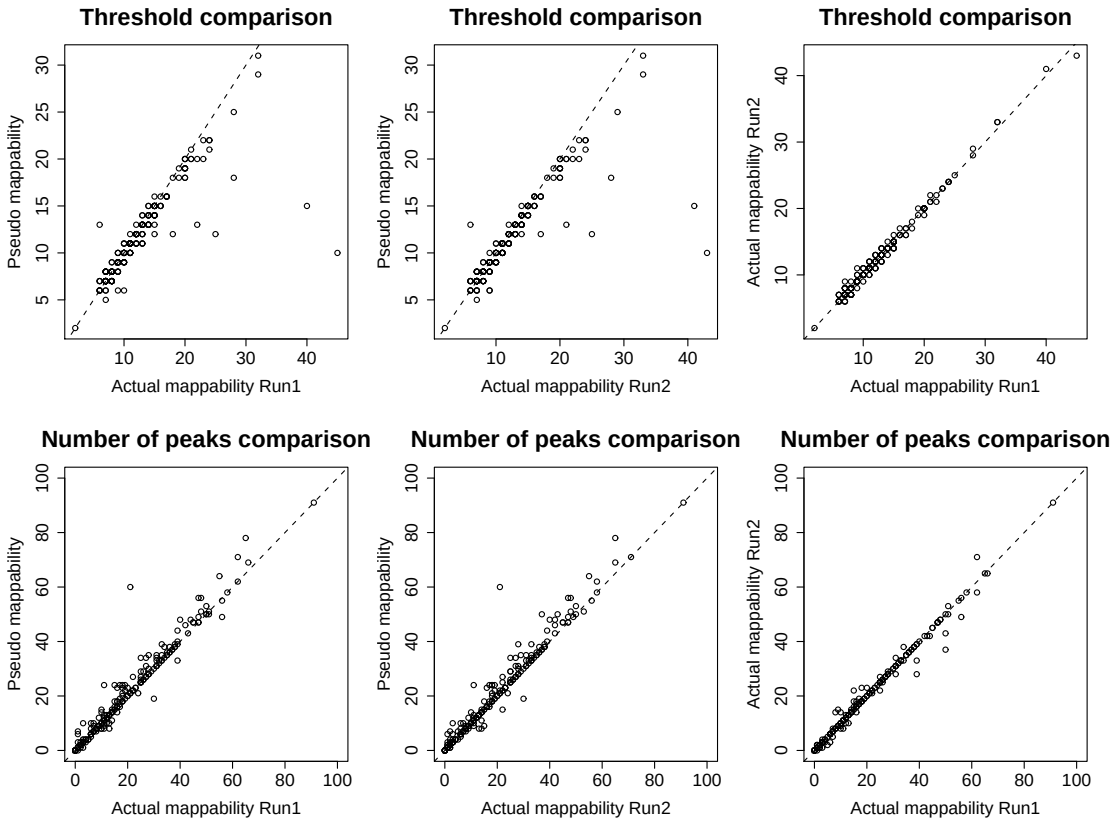
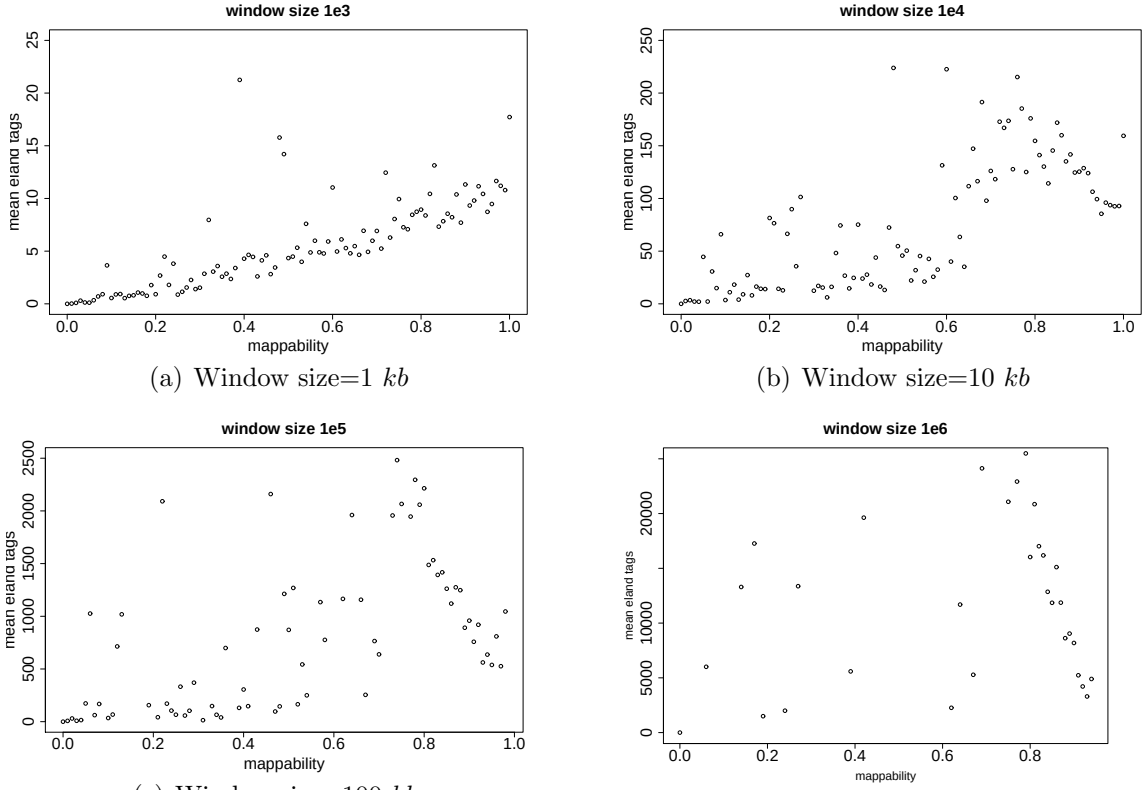


Figure 6: Comparison of results in PeakSeq-1 using actual and pseudo mappability (constant mappability across all the segments). Top panels compare thresholds for detecting bound regions between analyses using actual and pseudo mappability scores. Bottom panels compare the number of peaks obtained. Each data point corresponds to a 1 Mb segment.



(a) Window size=1 kb

(b) Window size=10 kb

(c) Window size=100 kb

(d) Window size=1 Mb

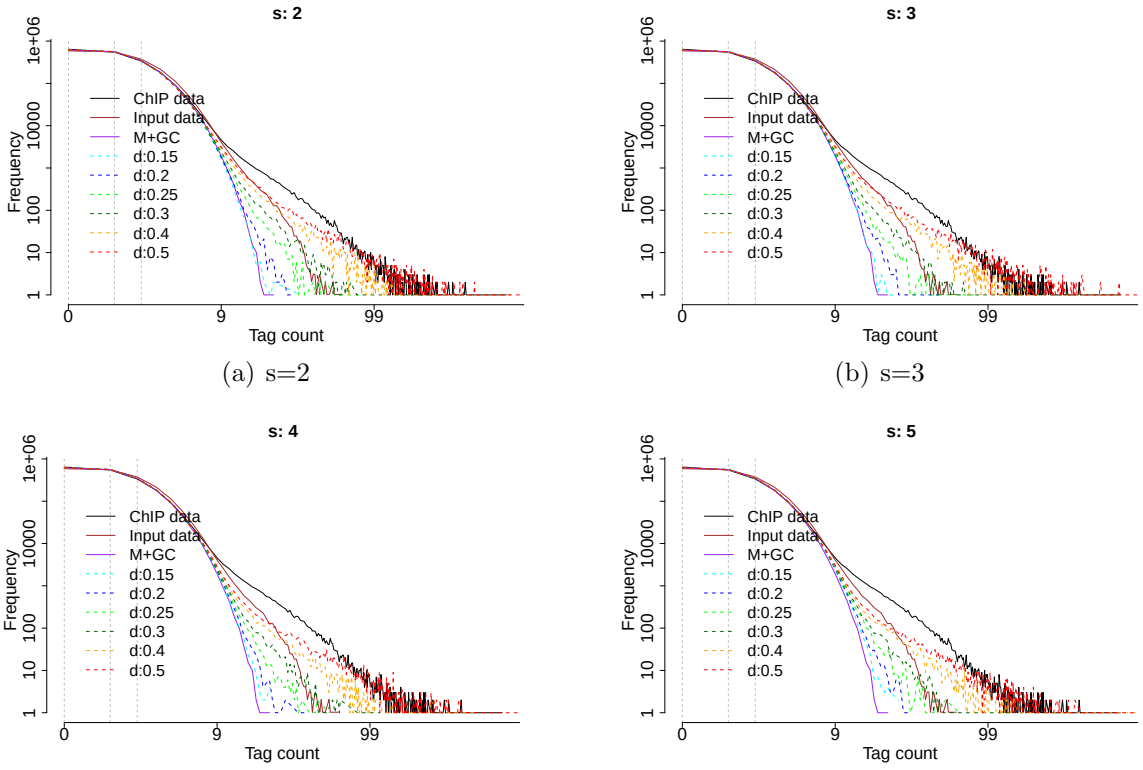
Figure 7: Mappability against mean tag counts. Each panel plots the mappability against average number of Eland tags for segments of length 1 kb to 1 Mb, respectively.

5 Choosing the s and d parameters of MOSAiCS two-sample model

In the two-sample analysis of ChIP-Seq data, we model the background mean as

$$\mu_j = \exp\left\{\beta_0 + \left[\beta_M \log_2(M_j+1) + \boldsymbol{\beta}'_{GC}\boldsymbol{S}p(GC_j) + \beta_{X1}X_j^d\right] \mathrm{I}(X_j \leq s) + \beta_{X2}X_j^d \mathrm{I}(X_j > s)\right\},$$

where s and d are tuning parameters. We consider $s \in \{2,3,4,5\}$ and $d \in \{0, 15, 0.25, 0.3, 0.4, 0.5\}$. Note that we do not consider larger s values since such values generate strata with too few number of bins in the estimation procedure. Supplementary Figure 8 compares the goodness of fit for STAT1 ChIP-Seq data for chromosome 1 (the other chromosomes result in relatively similar patterns). Solid brown line in each panel is the empirical distribution of the matching input DNA sample. The goal is to tune s and d such that the estimated background is close to the empirical distribution of matching input DNA sample. We choose $s = 2$ and $d = 0.25$ for STAT1 ($s = 4$ and $d = 0.25$ for GATA1, plots not shown). In our software implementation, s and d are set to default values from STAT1; however users can vary these parameters and compare both the GOF plots and the BIC scores of the resulting models. Overall, s is expected to be inversely related to the sequencing depth of the input control sample. When the sequencing depth of the input control sample is high, s is expected to be small because mappability and GC content effects will be absorbed into input counts.



(a) s=2 (b) s=3

(c) s=4 (d) s=5

Figure 8: Tuning of s and d in the two-sample MOSAiCS model. Each panel plots the goodness of fit for $d = 0, 15, 0.25, 0.3, 0.4, 0.5$ at a particular s value. Solid purple line is the estimated background using only mappability and GC content in one-sample problem. Solid brown line is the empirical distribution of matching input DNA.

6 Further results on STAT1 ChIP-Seq data

6.1 Comparison of methods on qPCR validated STAT1 target regions from Euskirchen et al. (2007)

We compare the overlap of the peak sets identified by each method with a gold-standard set of 280 (120 positive, 160 negative) ChIP-chip target sites validated independently by qPCR (Euskirchen et al., 2007) (http://encode.gersteinlab.org/data/Euskirchen_etal/) (PCR validated coordinates were lifted over from HG17 to HG18). The results on top 10000 peaks are presented in Supplementary Table 2. We report the results using both the original peak boundaries and also refining each peak boundry to be 2500 *bp* to account for the differences in peak sizes because many qPCR validated targets are adjacent to each other giving an advantage to methods with wider peaks. Among one-sample analyses, MOSAiCS-1S and PeakSeq-1S are comparable in terms of their true positive and negative rates, whereas CisGenome-1S and MACS-1S have lower true positive rates. CisGenome-2S performs slightly worse than the other two-sample methods and MOSAiCS and PeakSeq capture the highest number of true negatives in the one- and two-sample comparisons.

6.2 Comparison of one-sample methods when a two-sample analysis with input DNA is used as the gold standard

We repeat the analysis presented in Table 3 of the main text by replacing naked DNA with input DNA in the binomial test to obtain the gold-standard peak set. The resulting bin level sensitivity and specificity, and peak level sensitivity of each method are given in Supplementary Table 3. The relative performances of the different methods are comparable to the results using naked DNA as the gold standard.

Table 2: Comparisons of top 10000 peaks with the qPCR validated regions

Method	TP	TN	Ave	TP*	TN*	Ave*
MOSAiCS-1S	69 (26)	156	0.775	73	156	0.792
PeakSeq-1S	68 (25)	156	0.771	70	155	0.776
CisGenome-1S	39 (22)	158	0.656	64	155	0.751
MACS-1S	49 (22)	154	0.685	52	152	0.692
MOSAiCS-2S	64 (23)	158	0.760	71	156	0.783
PeakSeq-2S	70 (26)	157	0.782	73	155	0.789
CisGenome-2S	35 (20)	158	0.640	64	153	0.745
MACS-2S	79 (26)	154	0.810	73	153	0.782

NOTE: TP: True Positive, TN: True Negative. Numbers in parentheses under the column TP refers to numbers of unique peaks that overlap with the qPCR positive regions. "Ave" is computed by [(TP/120) + (TN/160)]/2. * Denotes calculations based on refining each peak to be of size 2500 *bp*. Each peak is resized by using both the start and the end position of the peak as the anchor and the best result is reported. Ranking of the methods remains robust to changing peak widths to smaller sizes.

Two-sample analysis with input DNA as the gold standard yields a smaller set gold standard peak set and, therefore, the sensitivities of all the methods increase, whereas the specificities of all the methods except CisGenome decrease.

6.3 FIMO analysis for comparing PeakSeq and MOSAiCS-2S (Input + M + GC)

We scanned the ranked peaks of PeakSeq-2S and MOSAiCS-2S (Input + M + GC) with the two available STAT1 position weight matrices from the JASPAR database (Portales-Casamar et al., 2010). Scoring on each peak set was conducted with the FIMO tool of the MEME suite (Bailey and Elkan, 1994; Bailey et al., 2009). FIMO evaluates the significance of each subsequence in a given dataset by comparing the likelihoods of the subsequence under the motif position weight matrix (PWM) model and a background model. For each peak set, we allowed the background model to be

Table 3: Bin and peak level sensitivity and specificity for one-sample analysis of STAT1 ChIP-Seq data

STAT1 ChIP	MOSAiCS-1S	CS-1S	MACS-1S(1)	MACS-1S(2)	PS-1S
Sensitivity (peak)	0.994	0.608	0.866	0.875	0.999
Sensitivity (bin)	0.994	0.363	0.882	0.863	0.998
Specificity (bin)	0.985	0.999	0.981	0.957	0.984

NOTE: Sensitivity and specificity of different methods for one-sample analysis of STAT1 ChIP-Seq data are reported by assuming bound regions from a two-sample comparison with input DNA to be the gold-standard set. MACS-1S(1) and MACS-1S(2) correspond to two different thresholds of p-value = 10^{-5} and p-value = 10^{-2}, respectively. CS-1S and PS-1S refer to CisGenome-1S and PeakSeq-1S, respectively.

estimated from the sequences of all the peaks. The differences in peak lengths were taken into account by controlling peak level FDR at 0.1. The following multiple hypotheses testing framework was utilized for the FDR control. Let the motif width be w. For a peak of length L ($L \geq w$), there are $2(L-w+1)$ subsequences of length w. The factor 2 accounts for the reverse complement. A p-value for each of the $2(L-w+1)$ subsequences is computed with the FIMO tool. The overall p-value for the peak is then adjusted using the Benjamini-Hochberg FDR control method (Benjamini and Hochberg, 1995) by taking into account a total of $2(L-w+1)$ tests.

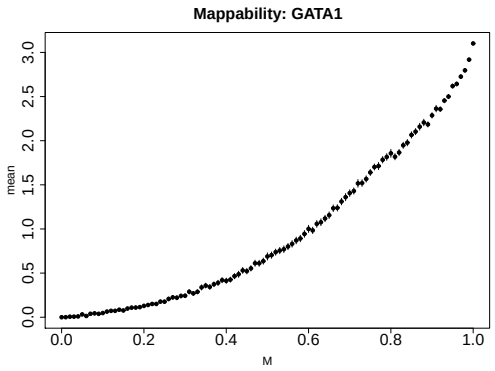
7 Further results on GATA1 ChIP-Seq data

7.1 Mappability and GC content biases

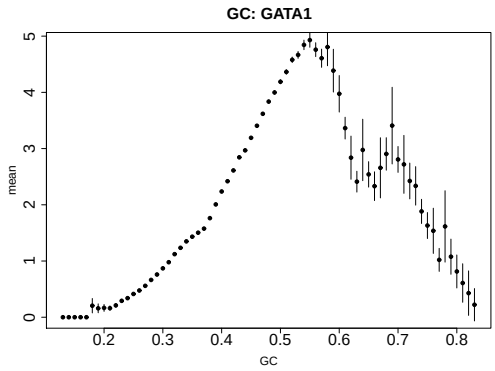
See Supplementary Figure 9.

7.2 Model selection

See Supplementary Table 4.



(a) Mean ChIP tag count vs. mappability.



(b) Mean ChIP tag count vs. GC content.

Figure 9: Mappability and GC content biases in the GATA1 ChIP-Seq sample. Left and right panels plot mean ChIP tag counts against the mappability score M_j and GC content GC_j.

Table 4: Model selection based on BIC scores for the GATA1 ChIP sample

MOSAiCS	1S (1 NB)	1S (2 NB)	2S (Input Only)	2S (Input+M+GC)
BIC	3653449	3655712	3814563	3588403

NOTE: Each cell reports BIC score for one-sample (1S) and two-sample (2S) MOSAiCS.

7.3 Motif analysis

A recent study on GATA1 (Zhang et al., 2009) showed that the consensus sequence [A/T]GATA[A/G] is necessary for GATA1 binding but its occurrence alone does not guarantee binding of GATA1. Specifically, while more than 90% of GATA1-bound regions contain this motif, less than 1% of regions that contain the motif are actually bound by GATA1. Zhang et al. (2009) further showed that multiple occurrences of the consensus sequence [A/T]GATA[A/G] strongly discriminate GATA1-bound regions from the unbound regions with the consensus sequence, i.e., the average number of occurrences of the [A/T]GATA[A/G] motif is about 2.3 in bound regions, compared to 1.1 in the unbound regions. Supplementary Figure 10 compares the different methods by scanning ranked peaks for one or more [A/T]GATA[A/G] motif occurrences. Consistent with the findings of Zhang et al. (2009), scanning for ≥ 1 motif occurrences is unable to discriminate the top ranking peaks from peaks in lower ranks. On the other hand, scanning for ≥ 2 motif occurrences is associated with the binding specificity of GATA1 (Figure 7(c) of the main text).

8 A generalized E-M algorithm when the signal component is a mixture of two negative binomial random variables

These derivations closely follow the derivations of the simpler model when the signal component is characterized by a single negative binomial distribution (Step 3(a)). We have $Y_j|Z_j = 1 \sim p_1(N_j + S_{1j}) + (1-p_1)(N_j + S_{2j}) + k$ and $S_{1j} \sim NegBin(b_1, c_1)$, $S_{2j} \sim NegBin(b_2, c_2)$. We introduce the latent variable G_j such that $Y_j|Z_j = 1 \sim N_j + S_{1j} + k$ if $G_j = 1$ and $Y_j|Z_j = 1 \sim N_j + S_{2j} + k$ if $G_j = 2$.

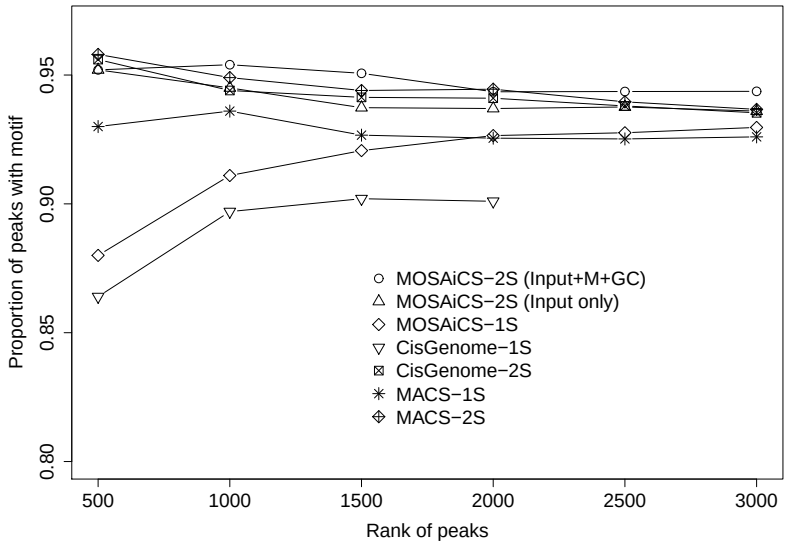


Figure 10: Scanning of one or more GATA1 consensus binding sequence occurrences in the top 3000 peaks.

The expected complete data likelihood is given by

$$\sum_{j=1}^{T} \{\mathrm{I}(Z_j = 0)[\log \pi_0 + \log P(Y_j|Z_j = 0)] \\ +\mathrm{I}(Z_j = 1, G_j = 1)[\log(1-\pi_0) + \log p_1 + \log P(Y_j|Z_j = 1, G_j = 1)] \\ + \mathrm{I}(Z_j = 1, G_j = 2)[\log(1-\pi_0) + \log(1-p_1) + \log P(Y_j|Z_j = 1, G_j = 2)]\}\,.$$

Then, the expected complete data likelihood is

$$\begin{aligned} Q \;=\; & \sum_{j=1}^{T} \{P(Z_j = 0|Y_j)[\log \pi_0 + \log P(Y_j|Z_j = 0)] \\ & +P(Z_j = 1, G_j = 1|Y_j)[\log(1-\pi_0) + \log p_1 + \log P(Y_j|Z_j = 1, G_j = 1)] \\ & + P(Z_j = 1, G_j = 2|Y_j)[\log(1-\pi_0) + \log(1-p_1) + \log P(Y_j|Z_j = 1, G_j = 2)]. \end{aligned}$$

The E- and M-steps for iteration t follow as:

E-step:

$$\begin{aligned}
g_j^{(t)} &= P(G_j = 1 | Z_j = 1, Y_j = y) \\
&= \frac{p_1^{(t-1)} P(Y_j = y | Z_j = 1, G_j = 1)}{p_1^{(t-1)} P(Y_j = y | Z_j = 1, G_j = 1) + (1 - p_1^{(t-1)}) P(Y_j = y | Z_j = 1, G_j = 2)} \\
&= \frac{p_1^{(t-1)} P(N_j = y | Z_j = 1, G_j = 1)}{p_1^{(t-1)} P(N_j + S_{1j} + k = y) + (1 - p_1^{(t-1)}) P(N_j + S_{2j} + k = y)}, \\
z_j^{(t)} &= P(Z_j = 1 | Y_j = y) \\
&= \frac{(1 - \pi_0)[p_1^{(t-1)} P(N_j + S_{1j} + k = y) + (1 - p_1^{(t-1)}) P(N_j + S_{2j} + k = y)]}{\pi_0 P(N_j = y) + (1 - \pi_0)[p_1^{(t-1)} P(N_j + S_{1j} + k = y) + (1 - p_1^{(t-1)}) P(N_j + S_{2j} + k = y)]}.
\end{aligned}$$

M-step:

$$\begin{aligned}
\frac{\partial Q}{\partial p_1} &= \frac{\sum_{j=1}^T P(Z_j = 1, G_j = 1 | Y_j)}{p_1} - \frac{\sum_{j=1}^T P(Z_j = 1, G_j = 2 | Y_j)}{1 - p_1} = 0 \\
\Rightarrow p_1^{(t)} &= \frac{\sum_{j=1}^T P(Z_j = 1, G_j = 1 | Y_j)}{\sum_{j=1}^T P(Z_j = 1, G_j = 1 | Y_j) + P(Z_j = 1, G_j = 2 | Y_j)} \\
&= \frac{\sum_{j=1}^T g_j^{(t)} z_j^{(t)}}{\sum_{j=1}^T z_j^{(t)}}.
\end{aligned}$$

Similar to step 3(a) in the manuscript, we update b_1, c_1, b_2 and c_2 by method of moments as follows:

$$\begin{aligned}
c_1^{(t)} &= \frac{E(Y_j | Z_j = 1, G_j = 1) - E(N_j) - k}{Var(Y_j | Z_j = 1, G_j = 1) - Var(N_j) - E(Y_j | Z_j = 1, G_j = 1) + E(N_j) + k}, \\
b_1^{(t)} &= \frac{[E(Y_j | Z_j = 1, G_j = 1) - E(N_j) - k]^2}{Var(Y_j | Z_j = 1, G_j = 1) - Var(N_j) - E(Y_j | Z_j = 1, G_j = 1) + E(N_j) + k}, \\
c_2^{(t)} &= \frac{E(Y_j | Z_j = 1, G_j = 2) - E(N_j) - k}{Var(Y_j | Z_j = 1, G_j = 2) - Var(N_j) - E(Y_j | Z_j = 1, G_j = 2) + E(N_j) + k},
\end{aligned}$$

$$b_2^{(t)} = \frac{[E(Y_j|Z_j=1,G_j=2)-E(N_j)-k]^2}{Var(Y_j|Z_j=1,G_j=2)-Var(N_j)-E(Y_j|Z_j=1,G_j=2)+E(N_j)+k},$$

where

$$\begin{aligned}
E(Y_j|Z_j=1,G_j=1) &= \frac{\sum_{j=1}^T z_j^{(t)} g_j^{(t)} Y_j}{\sum_{j=1}^T z_j^{(t)} g_j^{(t)}},\\
Var(Y_j|Z_j=1,G_j=1) &= \frac{\sum_{j=1}^T z_j^{(t)} g_j^{(t)} [Y_j - E(Y_j|Z_j=1,G_j=1)]^2}{\sum_{j=1}^T z_j^{(t)} g_j^{(t)}},\\
E(Y_j|Z_j=1,G_j=2) &= \frac{\sum_{j=1}^T z_j^{(t)} (1-g_j^{(t)}) Y_j}{\sum_{j=1}^T z_j^{(t)} (1-g_j^{(t)})},\\
Var(Y_j|Z_j=1,G_j=2) &= \frac{\sum_{j=1}^T z_j^{(t)} (1-g_j^{(t)}) [Y_j - E(Y_j|Z_j=1,G_j=2)]^2}{\sum_{j=1}^T z_j^{(t)} (1-g_j^{(t)})}.
\end{aligned}$$

9 Derivation of the full E-M algorithm

We provide the derivation of the E-M algorithm for our proposed model without making any simplifying assumptions to speed up the computations. However, as pointed out in Section 3 of the main text, the enriched distribution in MOSAiCS is a convolution of negative binomials involving the non-enriched distribution. This makes implementation of the full E-M algorithm highly unappealing since the M-step would have to rely on numerical optimization. Therefore, our software implements the procedure proposed in the main text.

1. *Model 1:* $Y_j|Z_j=1 \sim N_j + S_j$

   (a) Consider Model 1: $Y_j|Z_j=0 \sim N_j$ and $Y_j|Z_j=1 \sim N_j+S_j$ for $j=1,\cdots,T$, where $N_j \sim NegBin(a, a/\mu_j)$, $\mu_j = \exp(\beta_0 + f(M_j, GC_j, X_j \mid \beta_M, \boldsymbol{\beta}_{GC}, \boldsymbol{\beta}_X))$, and $S_j \sim NegBin(b,c)$.

The complete data likelihood is given by

$$L = \prod_{j=1}^{T} \left[\pi_0 P\left(Y_j|Z_j=0\right)\right]^{\mathrm{I}(Z_j=0)} \left[\left(1-\pi_0\right) P\left(Y_j|Z_j=1\right)\right]^{\mathrm{I}(Z_j=1)},$$

$$\begin{aligned}\log L &= \sum\nolimits_{j=1}^{T} \left[\mathrm{I}\left(Z_j=0\right)\left\{\log\pi_0 + \log P\left(Y_j|Z_j=0\right)\right\}\right. \\ &\quad \left.+\mathrm{I}\left(Z_j=1\right)\left\{\log\left(1-\pi_0\right) + \log P\left(Y_j|Z_j=1\right)\right\}\right].\end{aligned}$$

Then, the expected complete data likelihood is given by

$$\begin{aligned}Q &= \sum\nolimits_{j=1}^{T} \left[P\left(Z_j=0|Y_j\right)\left\{\log\pi_0 + \log P\left(Y_j|Z_j=0\right)\right\}\right. \\ &\quad \left.+P\left(Z_j=1|Y_j\right)\left\{\log\left(1-\pi_0\right) + \log P\left(Y_j|Z_j=1\right)\right\}\right].\end{aligned}$$

(b) *E-step:*

$$\begin{aligned}z_j^{(t)} &= P(Z_j=1|Y_j=y) \\ &= \frac{P(Z_j=1)P(Y_j|Z_j=1)}{P(Z_j=0)P(Y_j=y|Z_j=0)+P(Z_j=1)P(Y_j=y|Z_j=1)} \\ &= \frac{(1-\pi_0^{(t-1)})P(N_j+S_j=y)}{\pi_0^{(t-1)}P(N_j=y)+(1-\pi_0^{(t-1)})P(N_j+S_j=y)}.\end{aligned}$$

(c) *M-step:*

$$\frac{\partial Q}{\partial \pi_0} = \frac{\sum_{j=1}^{T} P\left(Z_j=0|Y_j\right)}{\pi_0} - \frac{\sum_{j=1}^{T} P\left(Z_j=1|Y_j\right)}{1-\pi_0} = 0$$

$$\begin{aligned}
\Rightarrow \pi_0^{(t)} &= \frac{\sum_{j=1}^{T} P\left(Z_j = 0 | Y_j\right)}{\sum_{j=1}^{T} P\left(Z_j = 0 | Y_j\right) + \sum_{j=1}^{T} P\left(Z_j = 1 | Y_j\right)} \\
&= \frac{\sum_{j=1}^{T} \left(1 - z_j^{(t)}\right)}{T}.
\end{aligned}$$

Since we do not have close form solutions for a, β_0, β_M, $\boldsymbol{\beta}_{GC}$, $\boldsymbol{\beta}_X$, b, c, we can estimate them by numerical maximization of the Q function by using, for example, `R` optimization function `optim()`.

2. *Model 2:* $Y_j|Z_j = 1 \sim p_1(N_j + S_{1j}) + (1 - p_1)(N_j + S_{2j})$

   (a) Now consider Model 2: $Y_j|Z_j = 0 \sim N_j$ and $Y_j|Z_j = 1 \sim p_1(N_j + S_{1j}) + (1 - p_1)(N_i + S_{2j})$ for $j = 1, \cdots, T$, where $N_j \sim NegBin(a, a/\mu_j)$, $\mu_i = \exp(\beta_0 + f(M_j, GC_j, X_j \mid \beta_M, \boldsymbol{\beta}_{GC}, \boldsymbol{\beta}_X))$, $S_{1j} \sim NegBin(b_1, c_1)$, and $S_{2j} \sim NegBin(b_2, c_2)$. We introduce the latent variable G_j such that $Y_j|Z_j = 1 \sim N_j + S_{1j}$ if $G_j = 1$ and $Y_j|Z_j = 1 \sim N_j + S_{2j}$ if $G_j = 2$.

   The complete data likelihood is given by

$$\begin{aligned}
L &= \prod_{j=1}^{T} \left[\pi_0 P\left(Y_j|Z_j = 0\right)\right]^{\mathrm{I}(Z_j=0)} \\
&\quad \left[\left(1 - \pi_0\right) p_1 P\left(Y_j|Z_j = 1, G_j = 1\right)\right]^{\mathrm{I}(Z_j=1, G_j=1)} \\
&\quad \left[\left(1 - \pi_0\right)\left(1 - p_1\right) P\left(Y_j|Z_j = 1, G_j = 2\right)\right]^{\mathrm{I}(Z_j=1, G_j=2)},
\end{aligned}$$

$$\begin{aligned}
\log L &= \sum_{j=1}^{T} \left[\mathrm{I}\left(Z_j = 0\right)\left\{\log \pi_0 + \log P\left(Y_j|Z_j = 0\right)\right\}\right. \\
&\quad + \mathrm{I}\left(Z_j = 1, G_j = 1\right)\left\{\log\left(1 - \pi_0\right) + \log p_1 + \log P\left(Y_j|Z_j = 1, G_j = 1\right)\right\} \\
&\quad \left. + \mathrm{I}\left(Z_j = 1, G_j = 2\right)\left\{\log\left(1 - \pi_0\right) + \log\left(1 - p_1\right) + \log P\left(Y_j|Z_j = 1, G_j = 2\right)\right\}\right].
\end{aligned}$$

Then, the expected complete data likelihood equals

$$\begin{aligned} Q &= \sum_{j=1}^{T} [P\left(Z_j = 0|Y_j\right)\left\{\log \pi_0 + \log P\left(Y_j|Z_j = 0\right)\right\} \\ &\quad + P\left(Z_j = 1, G_j = 1|Y_j\right)\left\{\log\left(1-\pi_0\right) + \log p_1 + \log P\left(Y_j|Z_j = 1, G_j = 1\right)\right\} \\ &\quad + P\left(Z_j = 1, G_j = 2|Y_j\right)\left\{\log\left(1-\pi_0\right) + \log\left(1-p_1\right) + \log P\left(Y_j|Z_j = 1, G_j = 2\right)\right\}]. \end{aligned}$$

(b) *E-step:*

$$\begin{aligned} g_j^{(t)} &= P(G_j = 1|Z_j = 1, Y_j = y) \\ &= \frac{p_1^{(t-1)} P(Y_j = y|Z_j = 1, G_j = 1)}{p_1^{(t-1)} P(Y_j = y|Z_j = 1, G_j = 1) + (1 - p_1^{(t-1)}) P(Y_j = y|Z_j = 1, G_j = 2)} \\ &= \frac{p_1^{(t-1)} P(N_j + S_{1j} = y)}{p_1^{(t-1)} P(N_j + S_{1j} = y) + (1 - p_1^{(t-1)}) P(N_j + S_{2j} = y)}, \\ z_j^{(t)} &= P(Z_j = 1|Y_j) \\ &= \frac{(1 - \pi_0^{(t-1)})[p_1^{(t-1)} P(N_j + S_{1j} = y) + (1 - p_1^{(t-1)}) P(N_j + S_{2j} = y)]}{\pi_0^{(t-1)} P(N_j = y) + (1 - \pi_0^{(t-1)})[p_1^{(t-1)} P(N_j + S_{1j} = y) + (1 - p_1^{(t-1)}) P(N_j + S_{2j} = y)]}. \end{aligned}$$

(c) *M-step:*

$$\frac{\partial Q}{\partial \pi_0} = \frac{\sum_{j=1}^{T} P\left(Z_j = 0|Y_j\right)}{\pi_0} - \frac{\sum_{j=1}^{T} P\left(Z_j = 1|Y_j\right)}{1 - \pi_0} = 0$$

$$\begin{aligned} \Rightarrow \pi_0^{(t)} &= \frac{\sum_{j=1}^{T} P\left(Z_j = 0|Y_j\right)}{\sum_{j=1}^{T} P\left(Z_j = 0|Y_j\right) + \sum_{j=1}^{T} P\left(Z_j = 1|Y_j\right)} \\ &= \frac{\sum_{j=1}^{T} \left(1 - z_j^{(t)}\right)}{T}. \end{aligned}$$

$$\frac{\partial Q}{\partial p_1} = \frac{\sum_{j=1}^{T} P\left(Z_j = 1, G_j = 1 | Y_j\right)}{p_1} - \frac{\sum_{j=1}^{T} P\left(Z_j = 1, G_j = 2 | Y_j\right)}{1 - p_1} = 0$$

$$\begin{aligned} \Rightarrow p_1^{(t)} & \\ &= \frac{\sum_{j=1}^{T} P\left(Z_j = 1, G_j = 1 | Y_j\right)}{\sum_{j=1}^{T} P\left(Z_j = 1, G_j = 1 | Y_j\right) + \sum_{j=1}^{T} P\left(Z_j = 1, G_j = 2 | Y_j\right)} \\ &= \frac{\sum_{j=1}^{T} g_j^{(t)} z_j^{(t)}}{\sum_{j=1}^{T} z_j^{(t)}}. \end{aligned}$$

Since we do not have close form solutions for a, β_0, β_M, $\boldsymbol{\beta}_{GC}$, $\boldsymbol{\beta}_X$, b_1, c_1, b_2, c_2, we can estimate them by numerical maximization of the Q function using, for example, `R` optimization function `optim()`.

References

Bailey, T. and Elkan, C. (1994), "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, California: AAAI Press, pp. 28–36, `http://meme.sdsc.edu/meme4_3_0/fimo-intro.html`.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009), "MEME Suite: tools for motif discovery and searching," *Nucleic Acids Research*, 37, W202–W208.

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

Euskirchen, G., Rozowsky, J., Wei, C., Lee, W., Zhang, Z., Hartman, S., Emanuelsson, O., Stolc, V., Weissman, S., Gerstein, M., Ruan, Y., and Snyder, M. (2007), "Mapping of transcription factor binding regions in mammalian cells by ChIP: Comparison of array- and sequencing-based technologies," *Genome Research*, 17, 898–909.

Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W., and Sandelin, A. (2010), "JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles," *Nucleic Acids Research*, 38, D105–10, `http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl`.

Rozowsky, J., Euskirchen, G., Auerbach, R., Zhang, D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. (2009), "PeakSeq enables systematic scoring of ChIP-Seq experiments relative to controls," *Nature Biotechnology*, 27, 66–75.

Zhang, Y., Wu, W., Cheng, Y., King, D., Harris, R., Taylor, J., Chiaromonte, F., and Hardison, R. (2009), "Primary sequence and epigenetic determinants of in vivo occupancy of genomic DNA by GATA1," *Nucleic Acids Research*, 37, 7024–7038.