The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1348–1372
DOI: 10.1214/16-AOAS938

A HIERARCHICAL FRAMEWORK FOR STATE-SPACE MATRIX INFERENCE AND CLUSTERING

BY CHANDLER ZUO, KAILEI CHEN, KYLE J. HEWITT, EMERY H. BRESNICK AND SÜNDÜZ KELEŞ[1]

University of Wisconsin–Madison

Integrative analysis of multiple experimental datasets measured over a large number of observational units is the focus of large numbers of contemporary genomic and epigenomic studies. The key objectives of such studies include not only inferring a hidden state of activity for each unit over individual experiments, but also detecting highly associated clusters of units based on their inferred states. Although there are a number of methods tailored for specific datasets, there is currently no state-of-the-art modeling framework for this general class of problems. In this paper, we develop the MBASIC (*M*atrix *B*ased *A*nalysis for *S*tate-space *I*nference and *C*lustering) framework. MBASIC consists of two parts: state-space mapping and state-space clustering. In state-space mapping, it maps observations onto a finite state-space, representing the activation states of units across conditions. In state-space clustering, MBASIC incorporates a finite mixture model to cluster the units based on their inferred state-space profiles across all conditions. Both the state-space mapping and clustering can be simultaneously estimated through an Expectation-Maximization algorithm. MBASIC flexibly adapts to a large number of parametric distributions for the observed data, as well as the heterogeneity in replicate experiments. It allows for imposing structural assumptions on each cluster, and enables model selection using information criterion. In our data-driven simulation studies, MBASIC showed significant accuracy in recovering both the underlying state-space variables and clustering structures. We applied MBASIC to two genome research problems using large numbers of datasets from the ENCODE project. The first application grouped genes based on transcription factor occupancy profiles of their promoter regions in two different cell types. The second application focused on identifying groups of loci that are similar to a GATA2 binding site that is functional at its endogenous locus by utilizing transcription factor occupancy data and illustrated applicability of MBASIC in a wide variety of problems. In both studies, MBASIC showed higher levels of raw data fidelity than analyzing these data with a two-step approach using ENCODE results on transcription factor occupancy data.

1. Introduction. The flow of genetic information through DNA transcription and RNA translation is a highly regulated process. The underlying mechanisms of regulation by both genomic and epigenomic factors are central targets in large

Received May 2015; revised January 2016.

[1] Supported in part by National Institutes of Health Grants HG006716 and HG007019.

Key words and phrases. State-space, clustering, E-M algorithm, transcription factors, ChIP-seq.

numbers of genomic and epigenomic studies. This paper is motivated by a number of such studies that aim to elucidate genomic regulatory mechanisms across multiple biological conditions. Experiments that investigate such processes produce a plethora of data types. For example, relevant to DNA transcription is transcription factor occupancy data that quantify which regions of DNA are occupied by DNA binding proteins that can enhance or reduce gene expression. Histone modification data map covalent post-translational modifications to histone proteins, core proteins that make up the nucleosome structure of DNA. Such modifications impact DNA transcription by altering the chromatin structure.

Computational and statistical analysis of these data often involve identifying genomic loci that show a significant signal, that is, enrichment, compared to background noise in the experimental measurements, with the operating principle that multiple loci might exhibit a similar signal profile over different biological conditions.

Improvements in the next-generation sequencing technology further accelerated rapid generation of these types of data. In return, the vast availability of such data has revolutionized the scope of genome regulation studies. Previous analyses had been restricted to detecting regions of the genome that were associated with one particular factor in one single organism. Many recent studies focus on detecting more complex functional patterns that integrate data from multiple organisms under multiple conditions, namely, the associations between DNA elements and how they change across biological and/or experimental conditions have been the focus of many integrative modeling approaches. Examples of these studies include the following:

Differential binding analysis among multiple ChIP-seq data. One of the key mechanisms of gene expression regulation is through differential activities of transcription factors and epigenetic modifications. Currently, chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq) is the state-of-the-art method for genome-wide profiling of protein-DNA interactions. Two such key interactions are DNA occupancy by transcription factors and histone modifications. Most transcription factors, that is, DNA binding proteins, recognize DNA in a sequence-specific manner and promote or repress gene expression. Similarly, histone modifications can induce diverse biological consequences such as transcriptional activation/deactivation.

The study of gene regulation often involves comparing transcription factor occupancy and histone modifications across multiple biological conditions. Such conditions can be different treatment levels, time points of measurements, or different dosage levels [Anders and Huber (2010), Ji et al. (2013), Liang and Keles (2012), Wei, Tenzen and Ji (2015)].

Transcription factor regulatory network analysis. The combinatorial nature of transcription factor regulation underlies the large diversity observed in eukaryotic gene control. This largely motivates construction of regulatory networks that model gene expression as a combinatorial function of regulatory interactions between DNA and different transcription factors. The large-scale data from the ENCODE project [ENCODE Project Consortium (2012)] now enable joint analyses of over one hundred human transcription factors across multiple cell types. Such analyses are posed to reveal a great amount of information about co-association patterns between different TFs, hierarchical network organizations, and systems-level integration of complex cellular signals [Cheng et al. (2011), Gerstein et al. (2012), Neph et al. (2012), Zeng et al. (2013)]. While the large number of TFs makes it computationally formidable to exhaust all possible combinatorial associations for such analyses, it is important to detect the most significant combinatorial patterns that preserve global regulatory dynamics.

Comparative functional genomic studies across different species. Functional genomics analysis compares gene expressions or TF occupancy profiles between multiple species. The main task is to identify divergent and conserved functional modules that are central to evolutionary relationships [e.g., Kunarso et al. (2010), Schmidt et al. (2010)]. Existing methods, that build on hidden Markov models [Roy et al. (2013)] or biclustering [Waltman et al. (2010)], implicitly assume that the functional modules should at least have similar signal profiles (i.e., expression, occupancy) among some subsets of the species under consideration. For these analyses, it is also important to identify functional modules that are fully divergent across species. These regions play an equally important role in understanding connectivity among species over the evolutionary history.

Although the types of data for these different studies vary, the underlying statistical principles are largely shared. Therefore, we propose a unified framework for the analysis of such data by formalizing the shared aspects. We formulate the underlying statistical problem as follows. Suppose a dataset $\{Y_{ik}\}$ is collected over a set of observational units (e.g., loci in genomic experiments) $i = 1, 2, \dots, I$ under conditions $k = 1, 2, \dots, K$. Inferring the association patterns within a single experiment involves mapping the corresponding set of observations $\{Y_{ik} : i = 1, 2, \dots, I\}$ to a finite discrete state-space, $\mathscr{S} = \{1, 2, \dots, S\}$. This space contains different levels of association (e.g., enrichment/nonenrichment indicating the status of occupancy in ChIP-seq experiments, expressed/not expressed in RNA-seq gene expression experiments). This falls under the classical finite-mixture modeling framework where a latent state variable $\theta_{ik} \in \mathscr{S}$ is inferred for each observational unit Y_{ik}. A higher level of modeling on the matrix $\Theta = (\theta_{ik})_{1 \le i \le I, 1 \le k \le K}$ is required for integrating the association patterns under different conditions. We call this matrix the *state-space matrix* since it describes the latent states of individual observations.

We propose the following framework to model the state-space matrix Θ. We assume that rows of Θ can be partitioned into $J + 1$ subsets: $\{1, \dots, I\} = \mathscr{C}_0 \cup \mathscr{C}_1 \cup \cdots \cup \mathscr{C}_J$. Rows of Θ within partition $\mathscr{C}_j$, $j \ge 1$, are generated by the same distribution parametrized by $w_{j\cdot} = (w_{jk})_{1 \le k \le K}$:

$$\theta_{ik} \sim g(\cdot | w_{jk}), \qquad i \in \mathscr{C}_j,$$

while the rows of $\mathscr{C}_0$, which denotes the group of "singleton" units, that is, units that do not cluster in any of the J groups, are generated by row-specific distributions. The goal of this model is thus to estimate a partitioning that best characterizes the row associations of the state-space matrix Θ.

We refer to the proposed framework as the *M*atrix *B*ased *A*nalysis for *S*tate-space *I*nference and *C*lustering (*MBASIC*). MBASIC is related to classical factor analysis which considers the problem of projecting one dimension (either row or column) of large noisy matrices into low-dimensional spaces. MBASIC has two distinguished features compared to the existing literature in these areas. First, MBASIC deals with matrices with discrete entries, while most existing methods are designed for matrices on continuous scales. Second, MBASIC estimates the low-dimensional projection by grouping the rows of the original matrix in contrast to the Principle Component Analysis (PCA) approaches which form linear combinations of the rows [e.g., Ji et al. (2013), Lee, Huang and Hu (2010)]. This is motivated by the following arguments:

1. In MBASIC, each factor estimate $w_{j\cdot}$ characterizes the commonality of a group of rows and is easily interpretable in practice. Such interpretability can further be enhanced by imposing structural restrictions on the $w_{j\cdot}$ vector for practical purposes. Examples of such constraints are described in Section 3.3;
2. PCA for high-dimensional matrices are often accompanied by regularization techniques which are computationally prohibitive for many epigenetic datasets. In contrast, clustering the matrix rows can be implemented very efficiently and in a straightforward manner.

The hierarchical structure of MBASIC is similar to two other recently proposed statistical models: iASeq [Wei et al. (2012)] and Cormotif [Wei, Tenzen and Ji (2015)]. Both these models incorporate a state-space clustering structure similar to MBASIC. MBASIC extends these models in several critically essential directions. First, MBASIC is developed for general purposes and can be easily implemented for a wide range of parametric distributions, while Cormotif and iASeq operate with specific distributions targeting the problems of differential expression and allele-specific binding. Second, neither of these models include a group of singletons with idiosyncratic state-space profiles. When we are agnostic about the "true" clustering structure in applications, separating the singletons can reduce their influence on the estimation of clustering parameters. Third, both iASeq and Cormotif separate estimation for the distributional parameters from the clustering structure, while MBASIC jointly fits all model parameters. A limiting assumption of MBASIC compared to these models is that MBASIC does not allow the distributional parameters within the same state to be heterogeneous. However, a preprocessing step that accounts for the heterogeneity can overcome such a limitation. We evaluate and discuss all of these features with extensive simulation studies in this paper.

This paper is organized as follows. We start with a formal description of MBASIC in Section 2, followed by model estimation and selection methods in Section 3.

We also investigate general features of MBASIC compared to iASeq and Cormotif with extensive simulations in this section. Section 4 presents results from several real data examples. Mathematical details of the algorithm are included in the Supplementary Material [Zuo et al. (2016)].

2. The hierarchical mixture model framework.

Consider a dataset with observations from I different *observational units* under K different *conditions*. For each condition $k \in \{1, 2, \dots, K\}$, there are n_k replicate experiments, indexed by $l = 1, 2, \dots, n_k$. We use Y_{ikl} to denote the observation for the lth replicate of unit i under condition k. For each condition k at unit i, there exists a hidden state variable $\theta_{ik} \in \mathscr{S} = \{1, 2, \dots, S\}$. The MBASIC model consists of the following components:

1. *State-space mapping*:

$$Y_{ikl} | \theta_{ik} = s \overset{\text{ind.}}{\sim} f_s(\cdot | \mu_{kls}, \sigma_{kls}, \gamma_{ikls}). \tag{2.1}$$

2. *State-space clustering*: θ_{ik}'s are independently sampled from $\mathscr{S}$ with the sampling probability:

$$P(\theta_{ik} = s) = \zeta p_{is} + (1 - \zeta) \sum_{j=1}^{J} \pi_j w_{jks}. \tag{2.2}$$

In (2.1), μ_{kls} and σ_{kls} are the parameters related to the mean and dispersion for the sth state for replicate l under condition k, and γ_{ikls} is the covariate encoding known information for unit i. In (2.2), p_{is}, ζ, π_j, and w_{jks} are additional nonnegative parameters subject to restrictions:

$$0 \le \zeta \le 1; \qquad \sum_{j=1}^{J} \pi_j = 1; \qquad \sum_{s=1}^{S} w_{jks} = 1, \qquad \forall j, k;$$
$$\sum_{s=1}^{S} p_{is} = 1, \qquad \forall i.$$

We further discuss these parameters in Section 2.2.

2.1. *State-space mapping*.

Equation (2.1) partitions observational units $i = 1, \dots, I$ into S subsets according to their hidden states. Within the same replicate, data from the same hidden state follow the same distribution $f_s(\cdot | \mu_{kls}, \sigma_{kls}, \gamma_{ikls})$. MBASIC assumes that the hidden states θ_{ik}'s are independent of the replicate index l, which means all replicates under the same condition have the same set of hidden states. However, distributional parameters for a given state can be different among replicates. Such a setting allows for the flexibility of modeling the heterogeneity in replicate experiments.

The density function f can be from an arbitrary parametric distribution. We consider three fundamental families of distributions commonly used for genomic data analysis:

- *Log-normal distribution.* $\mathrm{LN}(\mu_{kls}\gamma_{ikls}, \sigma_{kls})$ with a density function:

$$f_s(y|\mu_{kls}, \sigma_{kls}, \gamma_{ikls}) = \frac{1}{\sqrt{2\pi}\sigma_{kls}} \exp\left\{-\frac{(\log(y+1) - \mu_{kls}\gamma_{ikls})^2}{2\sigma_{kls}^2}\right\}. \tag{2.3}$$

- *Negative binomial distribution.* $\mathrm{NB}(\mu_{kls}\gamma_{ikls}, \sigma_{kls})$ with a density function:

$$f_s(y|\mu_{kls}, \sigma_{kls}, \gamma_{ikls}) = \frac{\Gamma(y+\sigma_{kls})}{\Gamma(\sigma_{kls})\Gamma(y)} \frac{(\mu_{kls}\gamma_{ikls})^y \sigma_{kls}^{\sigma_{kls}}}{(\mu_{kls}\gamma_{ikls} + \sigma_{kls})^{y+\sigma_{kls}}}. \tag{2.4}$$

- *Binomial distribution.* $\mathrm{Binom}(\gamma_{ikls}, \mu_{kls})$ with a density function:

$$f_s(y|\mu_{kls}, \gamma_{ikls}) = \binom{\gamma_{ikls}}{y} \mu_{kls}^y (1-\mu_{kls})^{\gamma_{ikls}-y}. \tag{2.5}$$

In these three examples, γ_{ikls} represents the known heterogeneity across loci, whereas μ_{kls} and σ_{kls} are unknown parameters. For example, when using equation (2.3) or (2.4) in a ChIP-seq analysis with $S = 2$ states, we can estimate γ_{ikl1} using data from the control samples so that the ChIP sample read counts in the background state scale with the control sample data [e.g., as in Zuo and Keleş (2014)], and assume $\gamma_{ikl2} = 1$ for the enriched states. Equation (2.5) can be used to analyze allele-specific binding data, where γ_{ikls} is the total read counts from both paternal and maternal alleles and is constant across s. Application with the binomial distribution also requires that $\mu_{kls} \sum_{i=1}^{I} \gamma_{ikls}$, $\forall k, l$, is strictly increasing in s for model identification.

The MBASIC can be easily extended to other classes of parametric distributions and estimation, for these distributions follow the same Expectation-Maximization skeleton. While Section 3 relies on these three distributions to describe the model and the estimation algorithms, the second real data example in Section 4 utilizes a more complex parametrization, which demonstrates the wide applicability of the MBASIC framework. Furthermore, we consider the following degenerate distribution:

$$f_s(y|\mu_{kls}, \sigma_{kls}, \gamma_{ikls}) = I(y = s), \tag{2.6}$$

where $I(\cdot)$ denotes the indicator function. This degenerate form corresponds to the situation where the states, θ_{ik}'s, are directly observed rather than inferred from Y_{ikl}'s. We utilize this parametrization for comparing MBASIC to alternative two-step analysis approaches in Section 3.5. Parameter estimation for this case follows a slightly modified procedure from the nondegenerate cases, which is described in Section 3.

2.2. *State-space clustering.* Equation (2.2) models the distribution of θ_{ik} as a mixture of multiple distributions. To illustrate this model, we introduce additional variables. The goal is to identify J clusters from the set of observation units $1 \leq i \leq I$. Let $b_i = I$ (unit i does not belong to any cluster) and $z_{ij} = I$ (unit i belongs to cluster j). The b_i variables entertain the possibility that some observations are "singletons," that is, they do not cluster with any other observational units. With these additional variables, the distribution in equation (2.2) can be hierarchically decomposed as follows:

- $b_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\zeta)$;
- $(z_{i1}, z_{i2}, \dots, z_{iJ}) \overset{\text{i.i.d.}}{\sim} \text{MultiNom}(1, (\pi_1, \pi_2, \dots, \pi_J))$;
- Conditional on b_i and z_{ij}, θ_{ik}'s are independent samples from $\mathscr{S}$, with sampling probabilities $P(\theta_{ik} = s | b_i = 1) = p_{is}$, $P(\theta_{ik} = s | b_i = 0, z_{ij} = 1) = w_{jks}$.

In this setup, although the singleton state-space probabilities p_{is} are assumed to be constant across conditions, that is, $P(\theta_{ik} = s) = p_{is}$, $\forall k$, this assumption is mildly restrictive since it accommodates $(P(\theta_{ik} = 1), \dots, P(\theta_{ik} = S))$ to follow an arbitrary prior distribution [e.g., $(P(\theta_{ik} = 1), \dots, P(\theta_{ik} = S)) \sim \text{Dirichlet}(\alpha, \dots, \alpha), \forall k$] as long as it leads to the same marginal distribution for θ_{ik} for all k.

It is worth noting that this hierarchical structure essentially seeks a low-rank representation for the matrix $\Theta = (\theta_{ik})_{1 \le i \le I, 1 \le k \le K}$. To illustrate this, we introduce additional matrices $\Theta_s = (I(\theta_{ik} = s))_{1 \le i \le I, 1 \le k \le K}$, $W_s = (w_{jks})_{1 \le j \le J, 1 \le k \le K}$, $Z = (z_{ij})_{1 \le i \le I, 1 \le j \le J}$, and vectors $p_s = (p_{is})_{1 \le i \le I}$, $B = (b_i)_{1 \le i \le I}$. Then the conditional expectation of Θ_s is

$$E(\Theta_s | Z, B) = (ZW_s) \circ \big((1 - B)1_K^T\big) + (p_s \circ B)1_K^T, \tag{2.7}$$

where "$\circ$" denotes the Hadamard product. We note that $E(\Theta_s | Z, B)$ is a matrix of rank $J + 1$, which is usually much smaller than the dimension of the matrix Θ_s. Similar models for low-rank representation of discrete matrices were considered in Lee, Huang and Hu (2010), and turned out to be challenging both theoretically and computationally. The row-clustering structure for the matrices $E(\Theta_s | Z, B)$ in MBASIC is more restrictive than the general low-rank structure. Such additional restrictions not only reduce the difficulty in parameter estimation but also enable the flexibility in many useful ways. For example, while Lee, Huang and Hu (2010) can only estimate one matrix at a time and thus is only applicable when $S = 2$, MBASIC can be applied to arbitrary values of S.

3. Model estimation and selection.

3.1. *Likelihood functions.* In the MBASIC model, the likelihood function for both the observed random variables Y_{ikl}'s and the unobserved θ_{ik}'s, z_{ij}'s, b_i's, that is, the full data likelihood, is given by

$$\begin{aligned} &l(\mu, \sigma, \pi, p, \zeta, w; y, \theta, z, b) \\ &\quad = \prod_{i=1}^{I} \zeta^{b_i}(1 - \zeta)^{1-b_i} \cdot \prod_{i=1}^{I} \prod_{k=1}^{K} \prod_{s=1}^{S} p_{is}^{I(\theta_{ik}=s)b_i} \cdot \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_j^{z_{ij}} \end{aligned} \tag{3.1}$$

$$\times \prod_{i=1}^{I}\prod_{k=1}^{K}\prod_{s=1}^{S}\left[\prod_{l=1}^{n_k} f_s(y_{ikl}|\mu_{kls},\sigma_{kls},\gamma_{ikls})\right]^{I(\theta_{ik}=s)}$$
$$\times \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K}\prod_{s=1}^{S} w_{jks}^{I(\theta_{ik}=s)(1-b_i)z_{ij}}.$$

For nondegenerate distributions, we can show that the marginal likelihood is

$$\begin{aligned} &l(\mu,\sigma,\pi,p,\zeta,w;y)\\ &\quad=\prod_{i=1}^{I}\Biggl\{\zeta\prod_{k=1}^{K}\left[\sum_{s=1}^{S}p_{is}\prod_{l=1}^{n_k}f_s(y_{ikl}|\mu_{kls},\sigma_{kls},\gamma_{ikls})\right]\\ &\qquad+(1-\zeta)\sum_{j=1}^{J}\pi_j\prod_{k=1}^{K}\left[\sum_{s=1}^{S}w_{jks}\prod_{l=1}^{n_k}f_s(y_{ikl}|\mu_{kls},\sigma_{kls},\gamma_{ikls})\right]\Biggr\}. \end{aligned} \tag{3.2}$$

Equation (3.2) is easily interpretable. Conditional on b_i and z_{ij}, the joint distribution for each Y_{ikl}, $1 \le l \le n_k$ is a mixture of S components, where the weight on the sth component is either p_{is} (when $b_i = 1$) or w_{jks} (when $b_i = 0$ and $z_{ij} = 1$). This yields the expressions in the square brackets. Integrating out b_i and z_{ij}, the joint distribution for Y_{ikl} of fixed i is a mixture of $J + 1$ components, with probability ζ of being a singleton and probability $(1 - \zeta)\pi_j$ of belonging to cluster j.

For the degenerate case, by substituting (2.6) into (3.2), it can be shown that the marginal likelihood is

$$\begin{aligned} &l(\mu,\sigma,\pi,p,\zeta,w;\theta)\\ &\quad=\prod_{i=1}^{I}\left\{\zeta\prod_{k=1}^{K}\prod_{s=1}^{S}p_{is}^{I(\theta_{ik}=s)}+(1-\zeta)\sum_{j=1}^{J}\pi_j\prod_{k=1}^{K}\prod_{s=1}^{S}w_{jks}^{I(\theta_{ik}=s)}\right\}. \end{aligned} \tag{3.3}$$

3.2. *An Expectation and Maximization* (*E-M*) *algorithm.* The hierarchical structure of MBASIC naturally fits in the Expectation-Maximization algorithm Dempster, Laird and Rubin (1977), which maximizes the marginal likelihood [equations (3.2) or (3.3)] by iteratively maximizing the complete data log-likelihood function. We let ϕ denote a vector including all unknown parameters μ, σ, π, p, ζ, w, and let $\hat{\phi}^{(t)}$ denote the parameter estimates at the tth iteration. The complete data log-likelihood function is

$$\begin{aligned} Q(\phi|\hat{\phi}^{(t-1)}) &= \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{s=1}^{S}\left[\sum_{l=1}^{n_k}\log f_s(y_{ikl}|\mu_{kls},\sigma_{kls},\gamma_{ikls})\right]E\bigl[I(\theta_{ik}=s)|\hat{\phi}^{(t-1)}\bigr]\\ &\quad+\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{s=1}^{S}\log p_{is}E\bigl[I(\theta_{ik}=s)b_i|\hat{\phi}^{(t-1)}\bigr] \end{aligned} \tag{3.4}$$

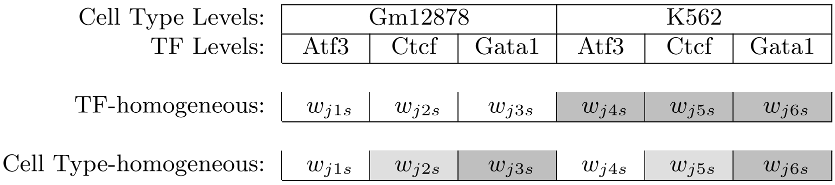
$$
\begin{aligned}
&+\sum_{i=1}^{I}\sum_{j=1}^{J}\log\pi_j E\big[z_{ij}(1-b_i)|\hat{\phi}^{(t-1)}\big]\\
&+\sum_{i=1}^{I}\big\{\log\zeta E\big[b_i|\hat{\phi}^{(t-1)}\big]+\log(1-\zeta)\big(1-E\big[b_i|\hat{\phi}^{(t-1)}\big]\big)\big\}\\
&+\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{s=1}^{S}E\big[I(\theta_{ik}=s)z_{ij}(1-b_i)|\hat{\phi}^{(t-1)}\big]\log w_{jks}.
\end{aligned}
$$

The E-M algorithm for MBASIC is outlined by Algorithm 1. Computational details for variable updates in each iteration are provided in Section 1.1 of the Supplementary Material [Zuo et al. (2016)]. Our parametrization allows for closed-form updates in the E-Step, which is important for computational speed. The M-Step updates are distribution dependent. We provide the updating formulum for distributions (2.3)–(2.6). Derivations for the updating formulum are provided in Section 1.2 of the Supplementary Material [Zuo et al. (2016)].

3.3. *Estimating structured clusters.* In integrative functional genomics studies, the set of experimental conditions usually consists of interactions of multiple experimental factors; hence, it is often important to identify clusters, states of which are homogeneous across the levels of one or more factors. For example, in a typical transcription factor network analysis, experimental conditions include the combination of different cell types and TFs. It is often desirable to separate loci groups whose states are homogeneous within each cell type from those with cell type-specific states for the purpose of cell type comparison. Depending on the cell types involved, such comparison can yield insights on cell development, pathology, and/or cell-specific functions. We refer to clusters with homogeneous states within each cell type as *TF-homogeneous*. Another example is encountered in comparative functional genomics studies across different species where experimental conditions range across both species and TFs. Clusters of loci, states of which are homogeneous across species conditional on each TF, constitute conserved functional modules. The *TF-homogeneous* clusters in this context represent

Algorithm 1 Expectation-Maximization (EM)

```
for $t = 1, 2, \ldots$ until convergence do
    Expectation Step: Compute the conditional expectations $E[I(\theta_{ik} = s)|\hat{\phi}^{(t-1)}]$, $E[b_i|\hat{\phi}^{(t-1)}]$, $E[I(\theta_{ik} = s)b_i|\hat{\phi}^{(t-1)}]$, $E[z_{ij}(1 - b_i)|\hat{\phi}^{(t-1)}]$, $E[I(\theta_{ik} = s)z_{ij}(1 - b_i)|\hat{\phi}^{(t-1)}]$;
    Maximization Step: Update estimates for parameters $\mu_{kls}$, $\sigma_{kls}$, $\zeta$, $\pi_j$, $w_{jks}$, $p_{is}$ as maximizers for (3.4).
end for
```



Cell Type Levels:	Gm12878			K562		
TF Levels:	Atf3	Ctcf	Gata1	Atf3	Ctcf	Gata1
TF-homogeneous:	w_{j1s}	w_{j2s}	w_{j3s}	w_{j4s}	w_{j5s}	w_{j6s}
Cell Type-homogeneous:	w_{j1s}	w_{j2s}	w_{j3s}	w_{j4s}	w_{j5s}	w_{j6s}

FIG. 1. *A graphical description for a parametrization with structural constraints. Interactions of* 2 *cell types and* 3 *TFs result in six experimental conditions. Parameters with homogeneous values are shaded by the same color.*

the marginal effect of the species factor, and play a central role in understanding the evolutionary relationships.

To estimate a cluster with homogeneity for a particular experimental factor, MBASIC allows structural constraints on its state-space parameters. Recall that the parameters of cluster j are represented by $w_{j.s} = (w_{j1s}, w_{j2s}, \dots, w_{jKs})$. Marginalizing the effect of this factor, the K experimental conditions can be partitioned into M sets, $\{1, 2, \dots, K\} = T_1 \cup T_2 \cup \cdots \cup T_M$, where conditions within each set differ only in the levels of this factor. The parameters of this cluster satisfy the following constraints:

$$w_{jk_1s} = w_{jk_2s}, \qquad \text{if } \exists\, m \text{ s.t. } k_1, k_2 \in T_m. \tag{3.5}$$

A pictorial depiction with six experimental conditions due to full interaction between 2 cell types and 3 TFs is depicted in Figure 1. Estimating structured clustering models follows the previous E-M algorithm with a slight modification. A constrained maximizer for w_{jks} subject to constraint (3.5) is computed as

$$\hat{w}_{jks}^{(t)} = \frac{\sum_{k':k' \in T_m} \sum_{i=1}^{I} E[I(\theta_{ik'} = s) z_{ij}(1 - b_i) | \hat{\phi}^{(t-1)}]}{\#\{T_m\} \sum_{s=1}^{S} \sum_{i=1}^{I} E[I(\theta_{ik} = s) z_{ij}(1 - b_i) | \hat{\phi}^{(t-1)}]}, \qquad k \in T_m.$$

MBASIC requires that such structural constraints must be specified a priori and remain fixed during model fitting. MBASIC incorporates a model selection procedure to compare models with different hypothesized structural constraints and numbers of clusters. We next describe the details of this model selection procedure.

3.4. *Model selection.* The MBASIC framework so far assumes that the total number of clusters J is known a priori. In practice, models with varying values of J need to be fitted independently and compared with each other according to some information criterion to determine the best value of J. Since the E-M algorithm aims to maximize the data likelihood function, AIC and BIC criteria can be utilized with MBASIC. The degrees of freedom for a model with J clusters is $df = F_1 S \sum_{k=1}^{K} n_k + (S - 1)I + J + F_2$, where $F_1 = 2$ for distributions (2.3) and (2.4), $F_1 = 1$ for (2.5), and F_2 is the total number of free variables among w_{jks}'s. If there are no structured clusters, we have $F_2 = JK(S - 1)$.

When there is no prior information available, both the total number of clusters and the number of clusters following structural constraints have to be determined. This results in a prohibitively large number of candidate models, and computing the information criterion for each of them is not practical. In such cases we incorporate the following two-phase strategy to limit the number of candidate models:

1. Evaluate models with a varying total number of clusters without any structural constraints. Select J_{opt} according to the minimal AIC or BIC value.
2. Evaluate models with the fixed number of J_{opt} clusters while varying the number of clusters following each structural constraint. Select the number of clusters following each structural constraint based on the minimal AIC or BIC value.

We acknowledge that the above two-step strategy is only a practical compromise to restrict the space of candidate models and does not guarantee finding the best model that globally minimizes the information criterion. However, we have conducted extensive simulation studies which illustrated that the proposed two-phase strategy performs well in a wide variety of settings.

3.5. *Simulation studies.* We conducted 6 model-based simulation studies to investigate the performance of MBASIC in various settings as summarized in Table 1. Each simulation study has multiple settings that vary the distributional assumptions, size of the state-space (S), proportion of singletons (ζ), number of units (I), number of clusters (J), and number of conditions (K). We provide the details of these simulation studies in Section 2 of the Supplementary Material and highlight the overall conclusions in this section.

Data in Simulation Studies 1–2 were simulated according to MBASIC's distributional assumptions. In Simulation Study 1, we emphasized the two most important features of MBASIC: the joint estimation procedure of all model parameters and the inclusion of a singleton cluster. We derived six alternative algorithms (Supplementary Table 1) to benchmark MBASIC's performance in various settings.

TABLE 1
Design of the simulation studies. S: size of the state-space; ζ: proportion of singletons; I: number of units; J: number of clusters; K: number of experimental conditions

Study	**Distribution**	S	ζ	I	(J, K)	**Model selection**
1	LN, NB, Bin	2, 3, 4	0, 0.1, 0.4	4000	(20, 30)	No
2	LN, NB, Bin	2	0.1, 0.4	4000	(20, 30)	Yes
3	iASeq	3	0, 0.1, 0.4	4000	(10, 20), (20, 30)	Yes
4	Cormotif	2	0	10,000	(4, 4), (5, 8), (5, 10)	Yes
5	Cormotif	2	0, 0.1, 0.4	4000	(10, 20)	Yes
6	LN	2	0, 0.33	4120, 4600, 6120	(8, 30)	Yes

Three of the algorithms (SE-HC, SE-MC, PE-MC) use two-stage procedures for model estimation, decoupling either the estimation of the state-space variables or the distributional parameters from the mixture modeling of clustering analysis. The other three algorithms are created as variations on these by excluding the singleton feature (SE-MC0, PE-MC0, MBASIC0). These benchmark algorithms are in spirit analogous to procedures in many applied genomic data analyses where the association between observational units are estimated separately from the estimation of individual dataset-specific parameters [e.g., Gerstein et al. (2012), Wei, Tenzen and Ji (2015), Wei et al. (2012)].

Supplementary Figures 2–4 summarize the performance comparisons in Simulation Study 1. We observed that MBASIC's joint estimation feature improved the inference for both the clustering structure and the individual states. In the presence of many singletons, the inclusion of their idiosyncratic state-space profiles was essential for robust clustering. In Simulation Study 2, we evaluated the effect of using BIC to select the number of clusters as well as the structural constraints within each cluster. Supplementary Tables 2 and 3 indicate that MBASIC was always able to select models with similar structures with the simulated truth.

In Simulation Studies 3 to 5, we simulated data according to the models proposed by iASeq [Wei et al. (2012)] and Cormotif [Wei, Tenzen and Ji (2015)]. These models allow heterogeneous distributional parameters within the same state, a potential advantage over MBASIC in specific data analysis such as differential expression or allele-specific binding. Comparison to these two models is intended to enable investigation of whether MBASIC is robust against such within-state heterogeneity. In Simulation Study 3, we showed that MBASIC with the binomial distribution could directly handle data generated under the iASeq framework and achieve competitive performance (Supplementary Figure 5). In Simulation Study 4, we inherited the simulation settings from Wei, Tenzen and Ji (2015), where distributions from different states were weakly separable, but the individual states were completely deterministic from the clustering. We explored more dynamic settings in Simulation Study 5, where we had easier separation between different states, but randomness among the states within the same cluster. We showed that a preprocessing step homogenizing the within-state units followed by MBASIC leads to comparable performance to Cormotif in Simulation Study 4 (Supplementary Figure 6) and much better performance in Simulation Study 5 (Supplementary Figure 7).

Wei, Tenzen and Ji (2015) discuss an interesting point that when the clustering model does not accommodate singletons, small clusters tend to be merged together to form spurious clusters, estimated state-space patterns of which are the averages among several true clusters. In order to investigate whether such a phenomena exists for MBASIC, we conducted Simulation Study 6, where we simulated data with two large clusters and six small clusters, and compared the performance of MBASIC and MBASIC0 to highlight the effect of including a singleton cluster.

We found that compared to MBASIC0, MBASIC was significantly less aggressive in merging small clusters. Overall, it captured large clusters and allocated the small cluster units as singletons (Supplementary Figures 10 and 11, Supplementary Tables 6, 7, and 8). This study highlighted the utility of a singleton cluster as a potential remedy for the merging of small clusters.

Combining results from all of our simulation studies, we conclude that MBASIC is a powerful model for both state-space estimation and clustering structure recovery. Its adaptability to singletons, effectiveness in model selection, and robustness against within-state heterogeneity strongly support its applicability for real datasets.

4. Applications of MBASIC to genome research problems.

4.1. *Transcription factor enrichment network.* Regulation of gene expression relies heavily on the context-specific combinatorial activities of TFs. Gene clustering analysis based on TF occupancy data, that is, ChIP-seq, aims to identify combinatorial patterns of TF occupancy and group genes based on such patterns. The ENCODE consortium [ENCODE Project Consortium (2012)] has generated TF ChIP-seq datasets for over 100 TFs across multiple cell types, and has motivated several integrative studies for learning regulation patterns [Gerstein et al. (2012), Wang et al. (2012)]. In this study, we applied MBASIC to the analysis of such data. Specifically, we focused on the TF enrichment patterns at the promoter regions, that is, -5000 bps and $+1000$ bps the transcription start site, of the 10,290 genes that had significant expression, as measured by RNA-seq, in either the Gm12878 or the K562 cells. The input data to MBASIC were the mapped numbers of reads at these promoter regions from the uniformly processed ChIP-seq data by Gerstein et al. (2012). We chose the cell types Gm12878 and K562 because they had the largest numbers of TF ChIP-seq experiments. The final dataset utilized included ChIP-seq data for $I = 10{,}290$ observational units over 30 TFs corresponding to $K = 60$ experimental conditions (cell type × TF) with a total of 166 replicate experiments.

We fitted MBASIC with $S = 2$ states and used log normal distributions as in equation (2.3). $s = 1$ corresponded to the unenriched state, and we let $\gamma_{ikl1} = \log(1 + x_{ik})$, where x_{ik} is the count from the matching control experiment at unit i. $s = 2$ corresponded to the enrichment state, and we let $\gamma_{ikl2} = 1$ for all loci.

We followed the two-phase procedure using BIC from Section 3.4 to select both the number of clusters and the structure of each cluster. In Phase 1, we selected the number of clusters as 24. In Phase 2, we considered two types of structural constraints for each cluster, referred to by *TF-homogeneity* and *cell type-homogeneity* and defined as $w_{jk_1s} = w_{jk_2s}$ if k_1 and k_2 corresponded to the same TF or cell type. We found that imposing cell type-homogeneity to any cluster would cause that cluster to be degenerate (i.e., no unit was assigned to that cluster). Therefore, we chose the final model among those with TF-homogeneity structures. The BIC and

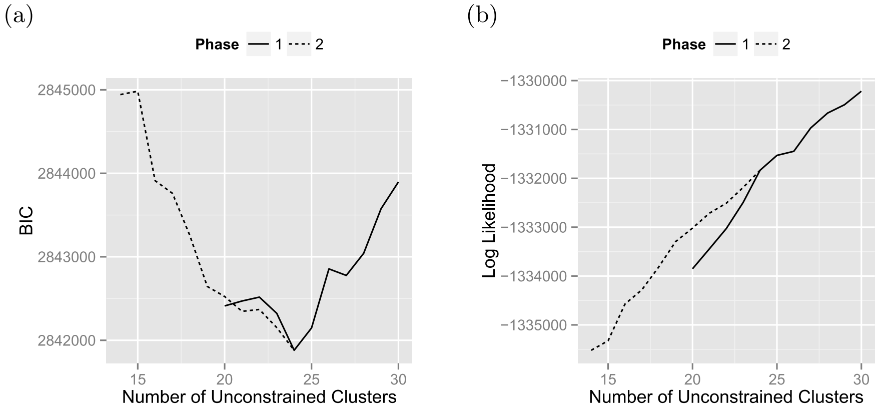


FIG. 2. (a) *BIC and* (b) *log-likelihood values for models with different structures. All the clusters are unstructured in the Phase* 1 *models and the x-axis denotes the total number of clusters. The total number of clusters is* 24 *for Phase* 2 *models and the x-axis denotes the number of unconstrained clusters. The remaining clusters have TF-homogeneity.*

log-likelihood values for different models fitted in both phases are shown in Figure 2. The final model had 24 unconstrained clusters, consisting of $1 - \zeta = 89.8\%$ of the 10,290 loci. The ranges of the estimated distribution parameters among replicates within the same cell type-TF combination are shown in Supplementary Figure 12. We notice that these parameters can be substantially different among replicated experiments. This provides further support for our replicate-specific parametrization.

To compare the normalized data and the predicted enrichment probability for each cluster, we computed the normalized signals[2] and compared them to the estimated cluster parameters. Figure 3 depicts such normalized signals from five randomly selected loci within each predicted cluster [Figure 3(a)], as well as the predicted enrichment probabilities at the corresponding condition and cluster (w_{jk2}'s) [Figure 3(b)]. We observe that the estimated enrichment probabilities at the cluster level capture the commonality among loci within each cluster. In addition, each loci cluster exhibits distinct combinatorial patterns of activity across all cell type-TF combinations. The cell type-TF combination enriched within each cluster is listed in Supplementary Table 9.

Our clustering results are consistent with the existing literature on the TF enrichment networks. For example, cooperating TFs tend to be enriched at the same loci. This pattern can be observed in Figure 3(b) between Bcl3 and Bclaf1. Pol2

[2] The normalized signal for unit i and condition k is

$$\tilde{\theta}_{ik} = \frac{\prod_{l=1}^{n_k} f_s(y_{ikl}|\hat{\mu}_{kl2}, \hat{\sigma}_{kl2}, \gamma_{ikl1})}{\prod_{l=1}^{n_k} f_s(y_{ikl}|\hat{\mu}_{kl2}, \hat{\sigma}_{kl2}) + \prod_{l=1}^{n_k} f_s(y_{ikl}|\hat{\mu}_{kl1}, \hat{\sigma}_{kl1}, \gamma_{ikl2})}.$$

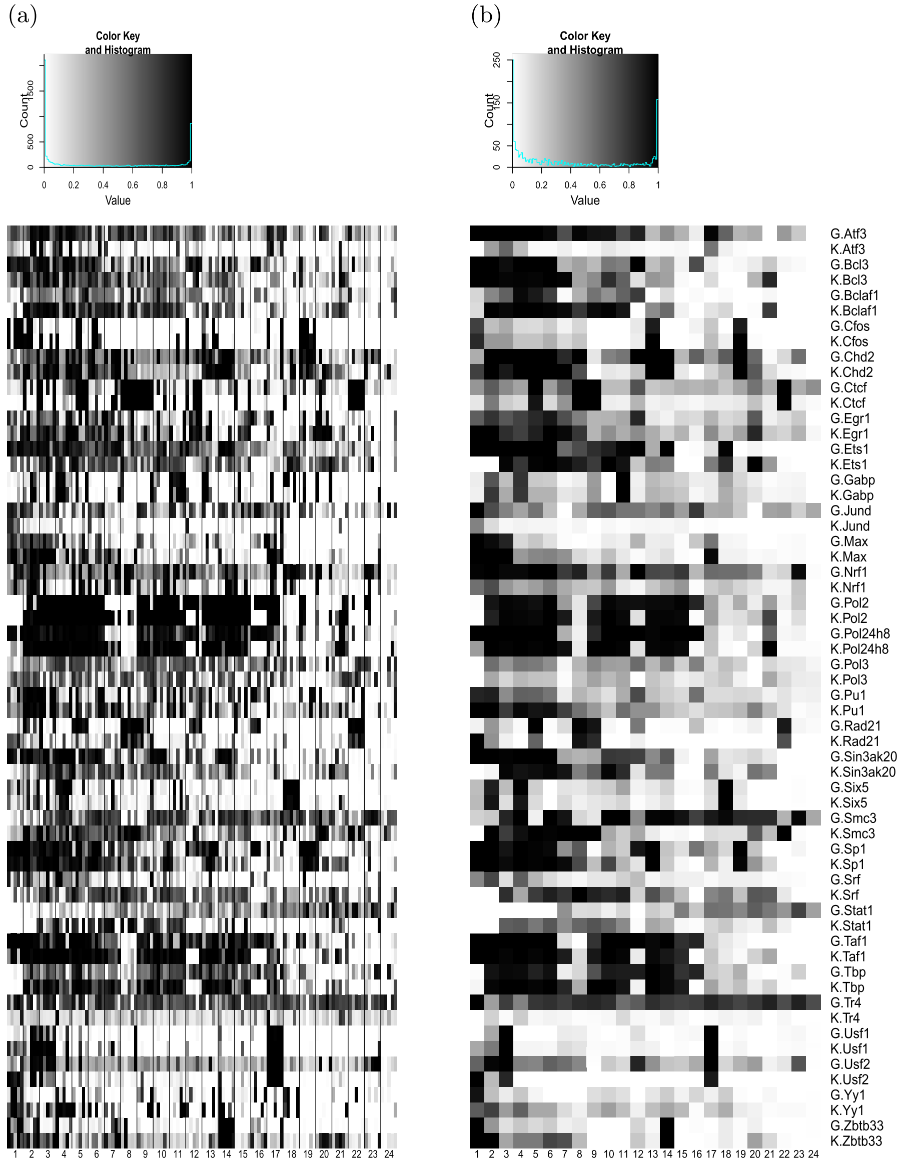


FIG. 3. (a) *Normalized data for each cell-TF combination at five subsampled loci within each cluster.* (b) *Estimated enrichment probability at each cell-TF combination for each cluster.*

and Pol24h8 represent Pol2 experiments with different antibodies. As expected, we observe enrichment at the same loci for these two different versions of Pol2 experiments. Moreover, pairs of TFs that have similar binding motifs have similar enrichment probabilities over the clusters. For example, Wang et al. (2012) discovered the UA1 motif as common to both Chd2 and Ets1 and the USF motif for

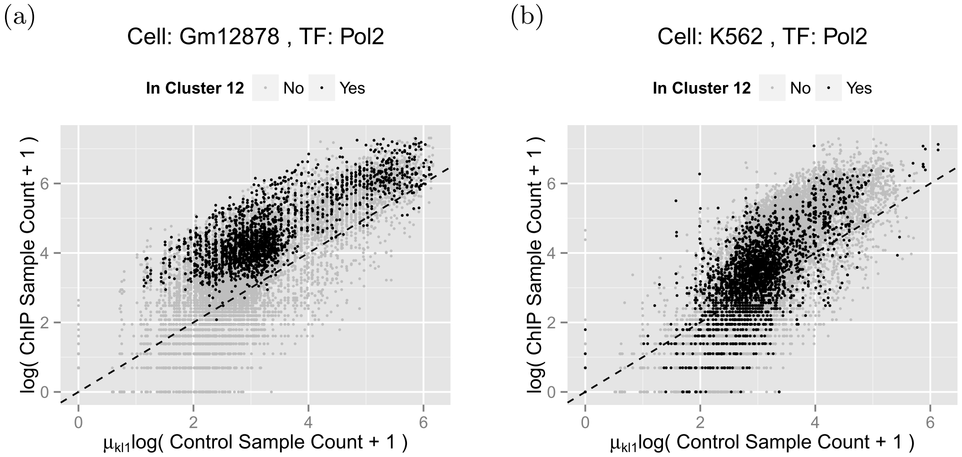


FIG. 4. (a), (b) *Plots of the transformed Pol*2 *ChIP sample read counts against the transformed control sample read counts for all units in* (a) *Gm*12878 *and* (b) *K*562 *cells. Data from unenriched units are expected to reside around the* 45 *degree dashed line.*

Max, Usf1, and Usf2. Interactions between Taf1 and Tbp have also been studied by Anandapadamanaban et al. (2013). Similar enrichment probabilities of these TFs across clusters can be observed in Figure 3(b). In addition to these observations that are consistent with the literature, our results illustrate how the genome-wide TF association patterns can be attributed to specific clusters. We explored the loci clusters with distinct patterns between cell types (e.g., Pol2 in Cluster 12, Figure 4), TFs from the same families (e.g., Bcl3 vs. Bclaf1 in Cluster 3, Supplementary Figure 13), and TFs with similar genome-wide enrichment (e.g., Max vs. Usf1 in Cluster 2, Supplementary Figure 14) using raw data. We further evaluated each cluster of genes for their KEGG pathway enrichment [Subramanian et al. (2005)], and identified 8 KEGG pathways that are significantly enriched in individual clusters (Table 2). Three of our clusters (Clusters 7, 9, and 19) have more

TABLE 2
Significantly enriched KEGG pathways across the 24 *clusters*

KEGG.name	# Genes overlapped	Z score	Cluster	Cluster size
Protein processing in endoplasmic reticulum	156	5.652	7	391
Fatty acid elongation in mitochondria	7	7.518	8	133
B cell receptor signaling pathway	74	6.016	9	146
Lysine biosynthesis	3	6.53	9	146
D-Glutamine and D-glutamate metabolism	3	5.548	12	184
Vitamin B6 metabolism	4	5.28	14	156
Nonhomologous end-joining	12	7.539	17	213
Lysosome	116	5.402	19	187

than half of their genes in one single pathway. Since KEGG pathways curate the known knowledge of molecular interaction systems, these clusters may be driven by unknown biological processes that warrant further investigation.

MBASIC infers the clustering structure based on its own estimates of the state-space profiles. The ENCODE consortium provides the estimated enrichment regions (i.e., *peaks*) for each experiment in this study. Then a natural question is whether MBASIC reveals more information compared to clustering of genes based on ENCODE-estimated binary enrichment profiles of TFs. To address this, we created a binary vector for each gene by overlapping its promoter with the ENCODE peaks. Then we applied the state-of-the-art MClust model [Fraley and Raftery (2002)] to cluster the 10,290 promoter regions based on these peak profiles. MClust selected 90 clusters based on BIC. Supplementary Figure 15 displays cluster-level estimated enrichment probabilities of TFs across the conditions considered. Compared to Figure 3, we can see that many of the MClust clusters have very similar enrichment profiles. For example, Clusters 51, 7, 8, 32, and 54 contained almost no enrichment for any TFs, but are classified as distinct clusters. The association between units across these clusters are thus nontrivial to interpret. In addition, we found that, for some conditions, the enrichment states predicted by MBASIC are quite different than those from the ENCODE peak profiles (e.g., Figure 5). This is because the ENCODE peaks are identified by whole genome-wide analysis and may not reflect the differences between the ChIP and control samples at the local promoter regions. MBASIC attains larger raw data fidelity by directly modeling the counts at each unit rather than inheriting results from existing analyses.

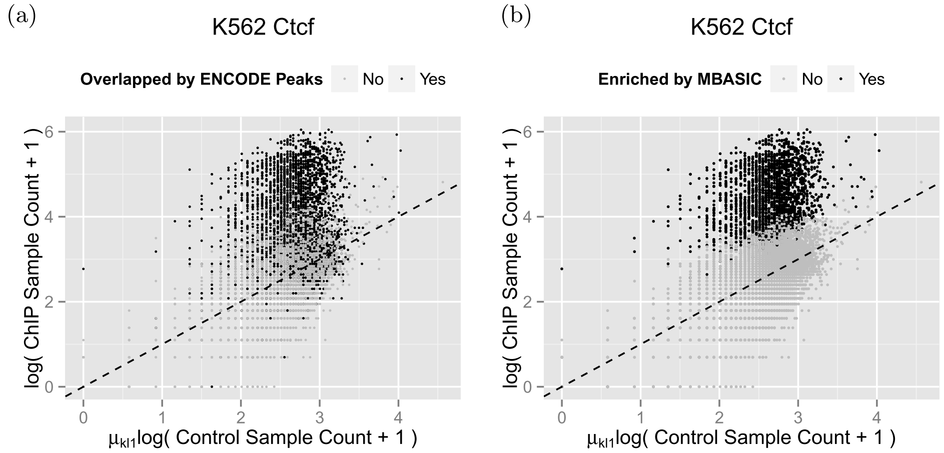


FIG. 5. (a), (b) *Transformed ChIP versus control sample read counts from a Gm*12878-*Ctcf dataset. Enrichment states are annotated by* (a) *ENCODE peak profiles and* (b) *MBASIC estimation. In MBASIC, an observational unit is estimated to be enriched if its enrichment probability satisfies* $P(\theta_{ik} = 2|Y) > 0.5$.

4.2. *Genome-wide identification of* +9.5*-like composite elements*. Johnson et al. (2012) and Gao et al. (2013) described the requirement of the intronic +9.5 site, an Ebox-GATA composite element located at chr6: 88143884–88157023 in the mouse genome (genome version mm9), to establish the hematopoietic stem/progenitor cell (HSC) compartment in the fetal liver and for hematopoietic stem cell genesis in the aorta–gonad–mesonephros (AGM), respectively. Furthermore, Johnson et al. (2012) and Hsu et al. (2013) showed that heterozygous +9.5 mutations cause a human immunodeficiency associated with myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML). Because the +9.5 site is the only known *cis*-element deletion which depletes fetal liver HSCs and is lethal at E13-14 of embryogenesis, identifying additional loci that have similar functionality is extremely important for establishing mechanisms that enable GATA factor-bound regions with nonredundant activity and have the potential to reveal novel targets for therapeutic modulation of hematopoiesis. In this application, we identified 4803 genomic regions with the Ebox-GATA motif (CATCTG-N[7-9]-AGATAA where N[7-9] denotes a variable size spacer of 7 and 9 nucleotides) in the human genome (genome version hg19). We considered a 150 bps window anchored at each of the 4803 composite elements as the observational unit. To analyze the TF occupancy activities at these units and identify a group of composite elements with occupancy profiles similar to that of the +9.5 composite element, we downloaded all ChIP-seq data for the Huvec and K562 cells from Gerstein et al. (2012). In total, the data set contained 224 replicates spanning $K = 84$ experimental conditions and 77 TFs.

We used negative binomial distributions with $S = 2$ states, where $s = 1$ denoted the unenriched (unoccupied) state, in the MBASIC framework. We chose $\gamma_{ikl1} = 1 + x_{ik}$, where x_{ik} is the count for unit i from the matching control experiment for condition k, to incorporate data from the accompanying control experiments of the ChIP samples. For $s = 2$, we utilized the following mixture distribution to account for the heavy tails observed in the raw data:

$$\begin{aligned} Y_{ikl} - 3 | \theta_{ik} &= 2 \overset{\text{ind.}}{\sim} \nu_{ikl} \, \mathrm{NB}(\mu_{kl2}, \sigma_{kl2}) + (1 - \nu_{ikl}) \, \mathrm{NB}(\mu_{kl3}, \sigma_{kl3}), \\ \nu_{ik} &\overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(\upsilon_{kl}). \end{aligned}$$

Here, the constant 3 represents the minimum count threshold for enrichment estimation. The use of mixture distributions to capture heavy-tailed count data was previously considered by Zuo and Keleş (2014). We note that an alternative approach to capture heavy tailed counts would be to fit a model using $S = 3$ states, with $s = 2, 3$ representing two distinct enrichment components. Such an approach would differ from the proposed approach in a subtle yet important way. In this alternative approach, allocation of each unit to different enrichment components would affect the clustering estimation, while, in our approach, clustering is only determined by the enrichment status of the individual unit regardless of which enrichment component it follows. The E-M algorithm for this setting requires a slight modification as discussed in Section 1.2 of the Supplementary Material [Zuo et al. (2016)].

Following the two-phase model selection procedure using BIC, we selected the model with 3 clusters, 2 of which were cell type-homogeneous. The ranges of the estimated distribution parameters among replicates within the same condition are displayed in Supplementary Figures 16–17. The three clusters (denoted by C1, C2, and C3) included 332, 837, and 157 composite elements, respectively, and the remaining 3477 composite elements were identified as singletons. A heatmap for the enrichment probability of each unit under each cell type-TF combination across the three clusters is shown in Figure 6. The +9.5 element is a member of cluster C3 which consists of a total of 157 +9.5-like composite elements. A detailed genomic annotation of these elements are provided in Supplementary Table 10. Notably, 46% of the C3 elements reside in intronic regions and 42% of these are within the first intron. Only 15% of the cluster are located up to 10 Kb upstream of transcription start sites.

A detailed analysis of Figure 6 reveals that cluster C3 is driven by several transcription factors with known associations to GATA2. First, we note that a large fraction of the C3 loci are bound by BRG1. The chromatin remodeler BRG1 is involved in GATA1-mediated chromatin looping [Kim, Bresnick and Bultman (2009), Kim et al. (2009)] and co-localizes with GATA1 at some chromatin sites [Hu et al. (2011)]. BRG1 has broad functions in many cell types; however, conditional knockouts of BRG1 reveal its importance in specific cell and tissue contexts [Holley et al. (2014)]. Another factor that clearly stands out as having a GATA2-like profile in cluster C3 is ETS1. Our prior work identified the propensity of occupied GATA motifs to reside near Ets motifs [Linneman et al. (2011)] and Doré et al. (2012) has highlighted GATA2-ETS co-localization.

We next performed an alternative naive analysis by utilizing the list of peaks provided by the ENCODE project. As in the case of the Transcription Factor Enrichment Network example of Section 4.1, these peaks, provided by the ENCODE consortium, were identified by analyzing each dataset individually with ENCODE's uniform ChIP-seq processing pipeline. Supplementary Figure 18 displays the ENCODE peak profiles for our cell type-TF conditions. For each of the 4803 composite elements, we constructed a *peak profile*, which is a binary vector indicating whether the element overlaps with the ENCODE peaks for each cell type-TF combination. We then computed the peak profile-based similarity between the +9.5 site and each the of the composite elements using the `R` function `dist.binary` with the "Jaccard index" option. For comparison, we computed *pseudo-binary similarities* between each element and the +9.5 site using the MBASIC estimated enrichment probabilities across all conditions.[3] We then

[3] The pseudo-binary similarity between two units i_1 and i_2 is calculated as

$$s(i_1, i_2) = \frac{\sum_k P\{\theta_{i_1k} = 1|Y\}P\{\theta_{i_2k} = 1|Y\}}{\sum_k P\{\theta_{i_1k} = 1|Y\} + P\{\theta_{i_2k} = 1|Y\} - P\{\theta_{i_1k} = 1|Y\}P\{\theta_{i_2k} = 1|Y\}}.$$

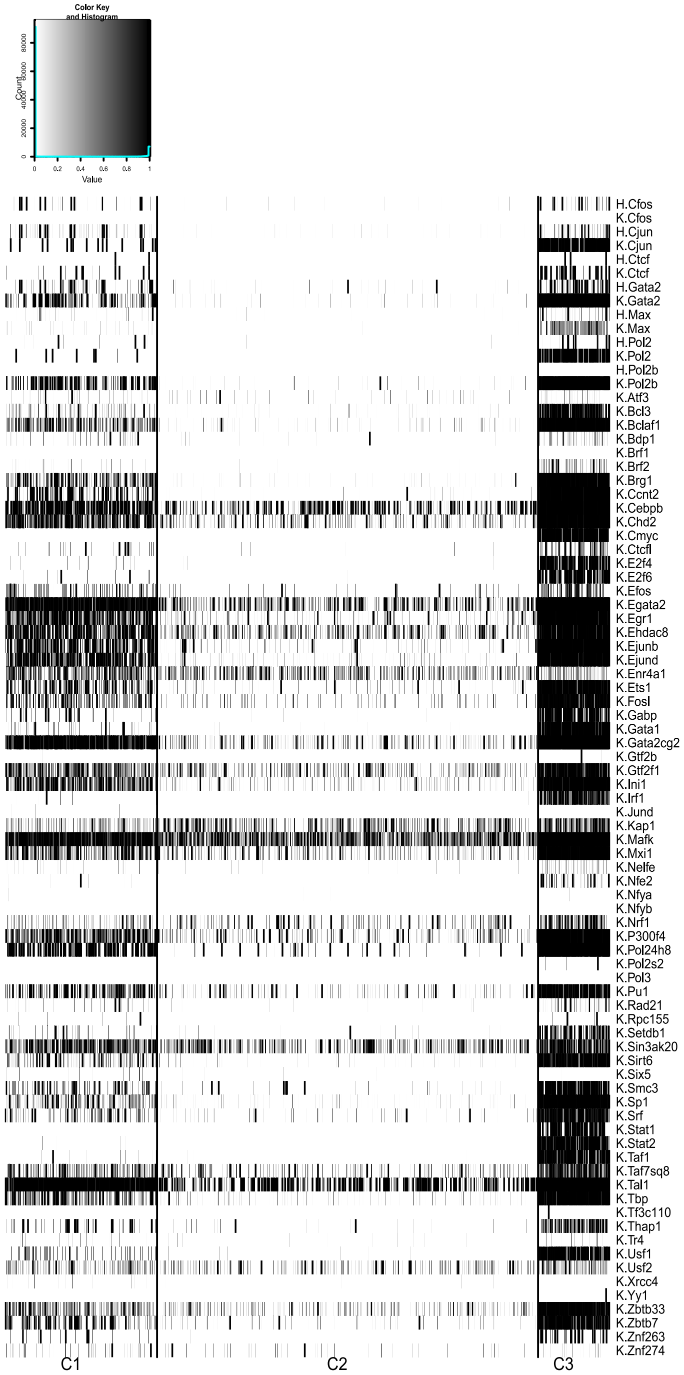


FIG. 6. *Posterior enrichment probability* [*i.e.*, $P(\theta_{ik} = 2|Y)$] *for all units in the three clusters. The rightmost column of the* $C3$ *cluster corresponds to the* $+9.5$ *element.*

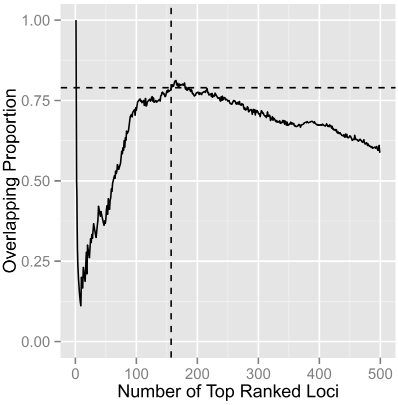


FIG. 7. *Proportion of overlap between the top ranked* +9.5*-like composite elements identified by MBASIC and ENCODE peak profiles. The overlap proportion is calculated by considering the same number of top ranked units* (x*-axis*) *in both the ENCODE-based and MBASIC-based similarities to the* +9.5 *site. The dashed lines mark that* 78.3% *of the C*3 *units are ranked in the top* 157 *based on the ENCODE peak profiles.*

ranked the composite elements based on both ENCODE and MBASIC estimated similarities. Figure 7 provides a comparison of the two lists as a function of top ranking composite elements. Overall, we observe that the rankings based on MBASIC estimation are consistent with the rankings based on the ENCODE peak profiles.

Although the rankings of the composite elements with respect to their +9.5 similarity using both the ENCODE peak profiles and MBASIC estimation were quite similar, the two approaches resulted in different enrichment estimation at the individual TF-cell combination level. Figure 8(a) compares the estimated cluster-level enrichment probabilities of each cell type-TF combination for cluster C3 against their average ENCODE peak profiles and highlights the difference between the two procedures. To further investigate these differences, we plotted the raw data for individual replicates and compared the composite elements that were estimated to be enriched by the two methods. An example using data from K562-Chd2 is displayed in Figure 8(b) and (c). Although many elements have significantly higher counts in the ChIP sample compared to the control sample, they are not identified as occupied by Chd2 in K562 according to ENCODE peak annotation. Another example using a replicate from K562-Yy1 is shown in Supplementary Figure 19, where several elements with zero ChIP count are overlapped by ENCODE peaks. These results indicate that MBASIC provides a grouping of the Ebox-GATA composite elements that is more consistent with the raw data compared to grouping based on ENCODE peak annotation.

5. Conclusions and discussion. Clustering analysis based on an underlying state-space is a common problem for many genomic and epigenomic studies where

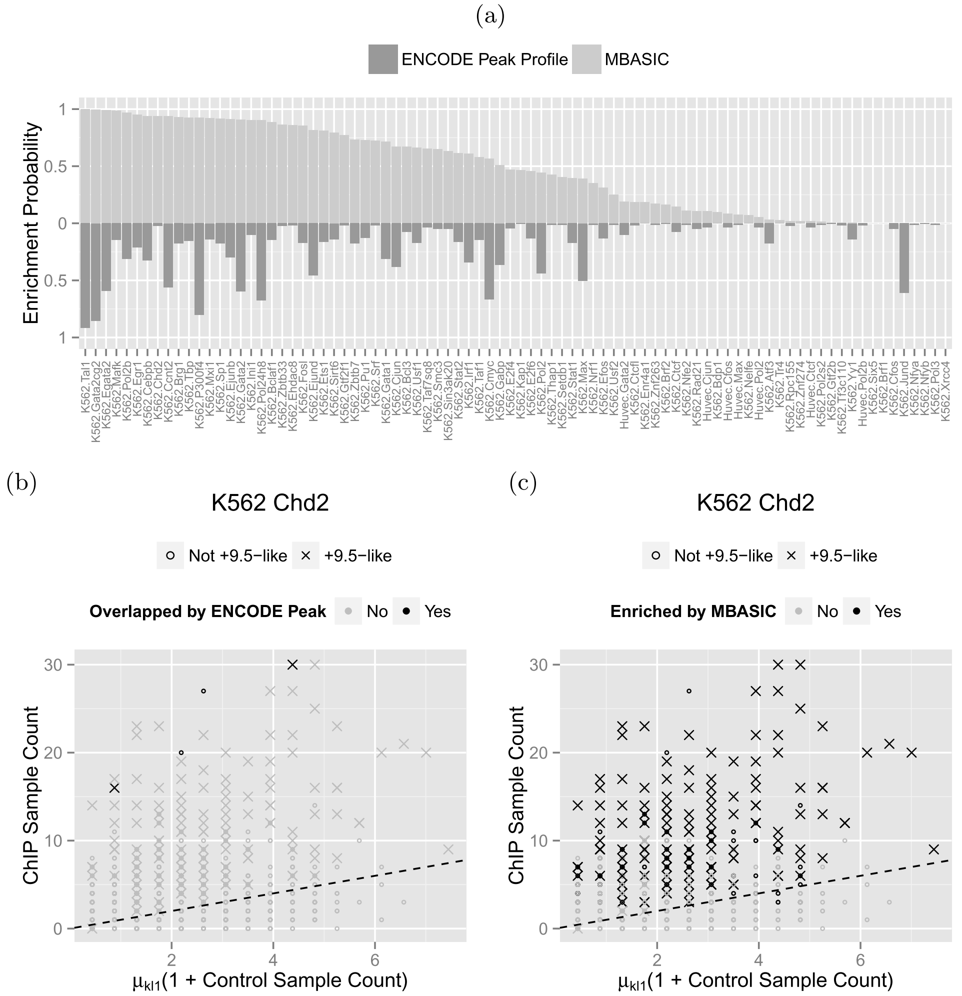


FIG. 8. (a) *Top half: Enrichment probabilities for the C*3 *units across all experimental conditions estimated by MBASIC. Bottom half: Proportion of C*3 *units that are overlapped by the ENCODE peaks for each condition.* (b), (c) *ChIP sample read counts against normalized control sample read counts for one replicate of the K*562*-Chd*2 *dataset. Enrichment status is annotated by* (a) *the ENCODE peak profiles and* (c) *MBASIC prediction.*

multiple datasets over many observational units are integrated. In this paper, we developed a unified statistical framework, called MBASIC, for addressing this class of problems. MBASIC simultaneously projects the observations onto a hidden state-space and infers clustered units in this space. The hierarchical structure of MBASIC enables the information of the state-space clusters to be fed back into the projection of the raw data, thus it reinforces the accuracy of predicting the state-space states of individual units. The MBASIC framework offers flexibility in a number of aspects of experimental design, such as different numbers of replicates under individual experimental conditions and missing values. Additionally, it is applicable to many parametric distributions. Our computational studies highlighted good operating characteristics of MBASIC and the two genomic applications illustrated how large numbers of ChIP-seq datasets can be integrated for addressing specific problems. In both of the applications, the MBASIC algorithm converged within 20 minutes for a fixed model on a 64 bit machine with an Intel Xeon 3.0 GHz processor and 64 GB of RAM. For model selection, we utilized the `R` package `snow` to implement the 2-phase procedure with parallel fitting of different candidate models using an 8-core 64 bit, 64 GB RAM machine with 8 Intel Xeon 3.0 GHz processors. These runs were completed under 2 hours. The computational efficiency of our model depends on the simple, closed-form updates in our E-M algorithm. Such a mathematical form is due, at least in part, to our modeling assumption that the rows of our state-space matrix are clustered. We have argued that this assumption, as compared to the PCA-type model structures, offers easier interpretation and is well suited for many genomic applications. MBASIC is available as an `R` package `mbasic` at https://github.com/chandlerzuo/mbasic.

SUPPLEMENTARY MATERIAL

Supplement to "A hierarchical framework for state-space matrix inference and clustering" (DOI: 10.1214/16-AOAS938SUPP; .pdf). Supplementary methods, simulation studies, tables, and figures.

REFERENCES

ANANDAPADAMANABAN, M., ANDRESEN, C., HELANDER, S., OHYAMA, Y., SIPONEN, M. I., LUNDSTRÖM, P., KOKUBO, T., IKURA, M., MOCHE, M. and SUNNERHAGEN, M. (2013). High-resolution structure of TBP with TAF1 reveals anchoring patterns in transcriptional regulation. *Nat. Struct. Mol. Biol.* **20** 1008–1014.

ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* **11** R106.

CHENG, C., YAN, K.-K., HWANG, W., QIAN, J., BHARDWAJ, N., ROZOWSKY, J., LU, Z. J., NIU, W., ALVES, P., KATO, M., SNYDER, M. and GERSTEIN, M. (2011). Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput. Biol.* **7**.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DORÉ, L. C., CHLON, T. M., BROWN, C. D., WHITE, K. P. and CRISPINO, J. D. (2012). Chromatin occupancy analysis reveals genome-wide GATA factor switching during hematopoiesis. *Blood* **119** 3724–3733.

ENCODE PROJECT CONSORTIUM (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74.

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635

GAO, X., JOHNSON, K. D., CHANG, Y.-I., BOYER, M. E., DEWEY, C. N., ZHANG, J. and BRESNICK, E. H. (2013). Gata2 cis-element is required for hematopoietic stem cell generation in the mammalian embryo. *J. Exp. Med.* **210** 2833–2842.

GERSTEIN, M. B., KUNDAJE, A., HARIHARAN, M., LANDT, S. G., YAN, K.-K., CHENG, C., MU, X. J., KHURANA, E., ROZOWSKY, J., ALEXANDER, R., MIN, R., ALVES, P., ABYZOV, A., ADDLEMAN, N., BHARDWAJ, N., BOYLE, A. P., CAYTING, P., CHAROS, A., CHEN, D. Z., CHENG, Y., CLARKE, D., EASTMAN, C., EUSKIRCHEN, G., FRIETZE, S., FU, Y., GERTZ, J., GRUBERT, F., HARMANCI, A., JAIN, P., KASOWSKI, M., LACROUTE, P., LENG, J., LIAN, J., MONAHAN, H., O'GEEN, H., OUYANG, Z., PARTRIDGE, E. C., PATACSIL, D., PAULI, F., RAHA, D., RAMIREZ, L., REDDY, T. E., REED, B., SHI, M., SLIFER, T., WANG, J., WU, L., YANG, X., YIP, K. Y., ZILBERMAN-SCHAPIRA, G., BATZOGLOU, S., SIDOW, A., FARNHAM, P. J., MYERS, R. M., WEISSMAN, S. M. and SNYDER, M. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature* **489** 91–100.

HOLLEY, D. W., GROH, B. S., WOZNIAK, G., DONOHOE, D. R., SUN, W., GODFREY, V. and BULTMAN, S. J. (2014). The BRG1 chromatin remodeler regulates widespread changes in gene expression and cell proliferation during B cell activation. *J. Cell. Physiol.* **229** 44–52.

HSU, A. P., JOHNSON, K. D., FALCONE, E. L., SANALKUMAR, R., SANCHEZ, L., HICKSTEIN, D. D., CUELLAR-RODRIGUEZ, J., LEMIEUX, J. E., ZERBE, C. S., BRESNICK, E. H. and HOLLAND, S. M. (2013). GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome. *Blood* **121** 3830–3837.

HU, G., SCHONES, D. E., CUI, K., YBARRA, R., NORTHRUP, D., TANG, Q., GATTINONI, L., RESTIFO, N. P., HUANG, S. and ZHAO, K. (2011). Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res.* **21** 1650–1658.

JI, H., LI, X., WANG, Q. and NING, Y. (2013). Differential principle component analysis of ChIP-seq. *Proc. Natl. Acad. Sci. USA* **110** 6789–6794.

JOHNSON, K. D., HSU, A. P., RYU, M.-J., WANG, J., GAO, X., BOYER, M. E., LIU, Y., LEE, Y., CALVO, K. R., KELES, S., ZHANG, J., HOLLAND, S. M. and BRESNICK, E. H. (2012). Cis-element mutation in a GATA-2-dependent immunodeficiency syndrome governs hematopoiesis and vascular integrity. *J. Clin. Invest.* **10** 3692–3704.

KIM, S.-I., BRESNICK, E. H. and BULTMAN, S. J. (2009). BRG1 directly regulates nucleosome structure and chromatin looping of the α globin locus to activate transcription. *Nucleic Acids Res.* **37** 6019–6027.

KIM, S.-I., BULTMAN, S. J., KIEFER, C. M., DEAN, A. and BRESNICK, E. H. (2009). BRG1 requirement for long-range interaction of a locus control region with a downstream promoter. *Proc. Natl. Acad. Sci. USA* **106** 2259–2264.

KUNARSO, G., CHIA, N.-Y., JEYAKANI, J., HWANG, C., LU, X., CHAN, Y.-S., NG, H.-H. and BOURQUE, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42** 631–634.

LEE, S., HUANG, J. Z. and HU, J. (2010). Sparse logistic principal components analysis for binary data. *Ann. Appl. Stat.* **4** 1579–1601. MR2758342

LIANG, K. and KELES, S. (2012). Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics* **28** 121–122.

LINNEMAN, A. K., O'GEEN, H., KELEŞ, S., FARNHAM, P. J. and BRESNICK, E. H. (2011). Genetic framework for GATA factor function in vascular biology. *Proc. Natl. Acad. Sci. USA* **108** 13641–13646.

NEPH, S., STERGACHIS, A. B., REYNOLDS, A., SANDSTROM, R., BORENSTEIN, E. and STAMATOYANNOPOULOS, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150** 1274–1286.

ROY, S., WAPINSKI, I., PFIFFNER, J., FRENCH, C., SOCHA, A., KONIECZKA, J., HABIB, N., KELLIS, M., THOMPSON, D. and REGEV, A. (2013). Arboretum: Reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Res.* **23** 1039–1050.

SCHMIDT, D., WILSON, M. D., BALLESTER, B., SCHWALIE, P. C., BROWN, G. D., MARSHALL, A., KUTTER, C., WATT, S., MARTINEZ-JIMENEZ, C. P., MACKAY, S., TALIANIDIS, I., FLICEK, P. and ODOM, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328** 1036–1040.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. and MESIROV, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102** 15545–15550.

WALTMAN, P., KACMARCZYK, T., BATE, A. R., KEARNS, D. B., REISS, D. J., EICHENBERGER, P. and BONNEAU, R. (2010). Multi-species integrative biclustering. *Genome Biol.* **11** R96.

WANG, J., ZHUANG, J., IYER, S., LIN, X., WHITFIELD, T. W., GREVEN, M. C., PIERCE, B. G., DONG, X., KUNDAJE, A., CHENG, Y., RANDO, O. J., BIRNEY, E., MYERS, R. M., NOBLE, W. S., SNYDER, M. and WENG, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22** 1798–1812.

WEI, Y., TENZEN, T. and JI, H. (2015). Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics* **16** 31–46. MR3365409

WEI, Y., LI, X., WANG, Q. and JI, H. (2012). iaseq: Integrative analysis of allele-specificity of protein-dna interactions in multiple chip-seq datasets. *BMC Genomics* **13** 1–19.

ZENG, X., SANALKUMAR, R., BRESNICK, E. H., LI, H., CHANG, Q. and KELEŞ, S. (2013). jMOSAiCS: Joint analysis of multiple ChIP-seq datasets. *Genome Biol.* **14** R38.

ZUO, C., CHEN, K., HEWITT, K. J., BRESNICK, E. H. and KELEŞ, S. (2016). Supplement to "A hierarchical framework for state-space matrix inference and clustering." DOI:10.1214/16-AOAS938SUPP.

ZUO, C. and KELEŞ, S. (2014). A statistical framework for power calculations in ChIP-seq experiments. *Bioinformatics* **30** 753–760.

C. ZUO
K. CHEN
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN–MADISON
1300 UNIVERSITY AVENUE
MADISON, WISCONSIN 53706
USA
E-MAIL: zuo@stat.wisc.edu
kchen@stat.wisc.edu

K. J. HEWITT
E. H. BRESNICK
DEPARTMENT OF CELL AND
REGENERATIVE BIOLOGY
UNIVERSITY OF WISCONSIN–MADISON
1300 UNIVERSITY AVENUE
MADISON, WISCONSIN 53706
USA
E-MAIL: kjhewitt@wisc.edu
ehbresni@wisc.edu

S. KELEŞ
DEPARTMENT OF BIOSTATISTICS AND
MEDICAL INFORMATICS
UNIVERSITY OF WISCONSIN–MADISON
425 HENRY MALL
MADISON, WISCONSIN 53706
USA
E-MAIL: keles@stat.wisc.edu