

A MAD-Bayes Algorithm for State-Space Inference and Clustering with Application to Querying Large Collections of ChIP-Seq Data Sets

Chandler Zuo, Kailei Chen, and Sündüz Keleş^(✉)

Department of Statistics, Department of Biostatistics and Medical Informatics,
University of Wisconsin-Madison, Madison, WI, USA
`keles@stat.wisc.edu`

Abstract. Current analytic approaches for querying large collections of chromatin immunoprecipitation followed by sequencing (ChIP-seq) data from multiple cell types rely on individual analysis of each dataset (i.e., peak calling) independently. This approach discards the fact that functional elements are frequently shared among related cell types and leads to overestimation of the extent of divergence between different ChIP-seq samples. Methods geared towards multi-sample investigations have limited applicability in settings that aim to integrate 100s to 1000s of ChIP-seq datasets for query loci (e.g., thousands of genomic loci with a specific binding site). Recently, [1] developed a hierarchical framework for state-space matrix inference and clustering, named MBASIC, to enable joint analysis of user-specified loci across multiple ChIP-seq datasets. Although this versatile framework both estimates the underlying state-space (e.g., bound vs. unbound) and also groups loci with similar patterns together, its Expectation-Maximization based estimation structure hinders its applicability with large numbers of loci and samples. We address this limitation by developing a MAP-based Asymptotic Derivations from Bayes (MAD-Bayes) framework for MBASIC. This results in a K-means-like optimization algorithm which converges rapidly and hence enables exploring multiple initialization schemes and flexibility in tuning. Comparisons with MBASIC indicates that this speed comes at a relatively insignificant loss in estimation accuracy. Although MAD-Bayes MBASIC is specifically designed for the analysis of user-specified loci, it is able to capture overall patterns of histone marks from multiple ChIP-seq datasets similar to those identified by genome-wide segmentation methods such as ChromHMM and Spectacle.

Keywords: Small-variance asymptotics · MAD-Bayes · Unified state-space inference and clustering · ChIP-seq

1 Introduction

Many large consortia (e.g., ENCODE [2], REMC [3]) as well as investigator-initiated projects generated large collections of ChIP-seq data profiling multiple

proteins and histone modifications across a wide variety of systems. Most current approaches for analyzing data from multiple cell types perform initial analyses such as peak calling in ChIP-seq independently in each cell/tissue/condition type. This approach ignores the fact that functional elements are frequently shared between related cell types, and leads to an over estimation of the extent of functional divergence between the conditions. Although the uniform processing pipelines developed by data-generating consortia and the resulting analysis of consortia data enable easy access to these data, joint analysis approaches that take advantage of the inherent relationships between datasets and cell types are required. Joint inference for ChIP-seq datasets can be formulated as inferring for each locus whether or not it exhibits ChIP-seq signal in a given condition and also grouping loci based on their profile similarity across multiple samples.

It is now widely accepted that joint analysis of these types of data can uncover signals that are otherwise too small to detect from a single experiment [4,5]. Among the available joint analysis methods, jMOSAiCS [6] builds on ChIP-seq peak-caller MOSAiCS [7] and incorporates a multi-layer hidden states model that governs the relationship of enrichment among different samples. [8] utilizes a one-dimensional Markov random field (MRF) model to account for spatial dependencies along the genome while modeling individual components by mixtures of Zero Inflated Poisson or Negative Binomial models. dCaP [9] uses a three-step log-likelihood ratio test to jointly identify binding events in multiple experimental conditions. ChromHMM [10] and Segway [11] are two commonly adopted approaches for segmenting the genome into chromatin states based on histone ChIP-seq and rely on hidden Markov models and Bayesian Networks, respectively. Recently, Spectacle [12] provided a transformative improvement of ChromHMM by utilizing spectral learning for parameter estimation in HMMs. hiHMM [13] uses a Bayesian non-parametric formulation of the HMMs while taking into account species-specific biases.

Overall, available strategies for considering multiple ChIP-seq datasets simultaneously can be broadly classified based on (i) whether or not they can deal with only TFs [14,15], only histone modifications [10–12,16,17], or both [5,6] types of ChIP-seq data; (ii) whether or not they rely on a priori analysis of individual datasets [10,12,14,15,17], (iii) whether or not they focus on differential occupancy and can handle very few numbers of conditions [14,18,19], (iv) whether or not they can scale up to 100s to 1000s of datasets. These approaches, with the potential exception of [12], do not scale up to 100s to 1000s of datasets since they, to a large extent, utilize variants of hidden Markov models and/or implement variants of the Expectation-Maximization (EM) algorithm [20] for parameter estimation. Furthermore, none of these approaches accommodate querying of multiple datasets for *selected* loci. Their analysis results serve to “annotate” user-specified loci without any notion of uncertainty.

We recently introduced MBASIC [1] as a probabilistic method for querying multiple ChIP-seq datasets jointly for user-specified loci. When multiple ChIP-seq datasets (multiple TFs profiled in different cell/tissue types under a variety of conditions) are available, the key inference encompasses both identifying peaks

in individual datasets (*state-space mapping*) as well as identifying groups of loci that cluster across different experiments (*state-space clustering*). At the core of MBASIC are biologically validated and commonly adapted models for measurements from individual experiments (e.g., read data models from [7, 21] for state-space mapping) and a mixture model for clustering of the loci with similar state-space mapping. Parameter estimation in this versatile model is based on the EM algorithm and hence does not scale up with large numbers of user-specified loci and ChIP-seq datasets. In this paper, we adopt a small-variance asymptotics framework for MBASIC and derive a K-means-like MAD-Bayes algorithm [22]. This alternative estimation framework for MBASIC targets large-scale datasets and genomic loci. Specifically, we consider a mixture of Log-normal distributions for state-specific observations with a Chinese Restaurant Process (CRP) [23, 24] as the clustering prior. Small-variance asymptotics for maximizing the posterior distribution leads to a K-means like objective function with a key penalty term for the number of clusters. Extensive comparisons with MBASIC indicate that this approach can significantly speed up model estimation without significant impact on the estimation performance. Although methods like ChromHMM and Spectacle inherently have a different purpose than MAD-Bayes MBASIC, we compared the three on histone ChIP-seq data from GM12878 cells. This comparison indicated that MAD-Bayes MBASIC can capture the overall patterns that these segmentation methods identify.

2 Method

We begin our exposition with an overall description of the Bayesian MBASIC model (Fig. 1) and then derive the MAD-Bayes algorithm. Some key aspects of our approach are model initialization and tuning parameter selection. Although these aspects arise in all of the above mentioned joint analysis methods, they are typically not well studied because of the computational costs.

2.1 The Bayesian MBASIC Model

We consider I genomic loci of interest, indexed from $i = 1, \dots, I$, from the reference genome with observations from K different experimental conditions. We use the notion of loci loosely in the sense that these loci could correspond to promoter regions of genes (all or members of specific pathways), locations of genome with a specific transcription factor (TF) binding motif, or peaks from a specific ChIP-seq experiment. The K conditions denote different TFs and cell/tissue types. Then, the key inference concerns analyzing I loci based on these K experiments. To further motivate the circumstances this inference problem arises, we consider an example from GATA-factor biology. In [25], we were interested in an overall analysis of all the E-box-GATA composite elements based on all the ENCODE ChIP-seq data to identify sites similar to the functional E-box-GATA composite element at the +9.5 loci which is causal for MonoMAC disease (a rare genetic disorder associated with myelodysplasia, cytogenetic abnormalities,

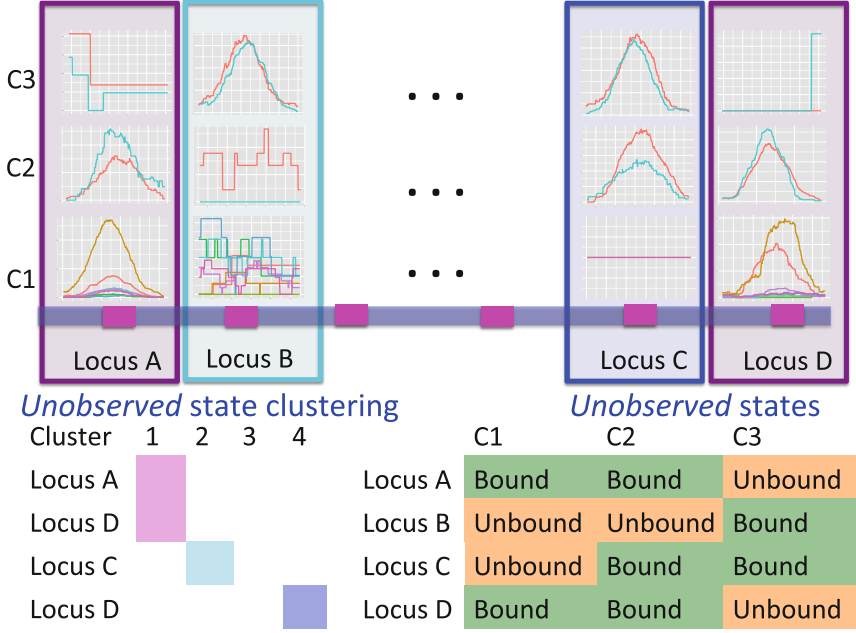


Fig. 1. Overview of the MBASIC modeling framework. Curves within each panel depict different replicates under the experimental conditions C1, C2, and C3. Loci A and D are in the same cluster.

and myeloid leukemias) [26]. The E-box-GATA composite elements are represented by CANNTGN{6-14}AGATAA oligonucleotides, where N denotes any nucleotide and N{6-14} denotes any nucleotide sequence of length 6 to 14 bps and are found abundantly in the genome, e.g., hg19 harbors $\sim 102\text{K}$ of them. Joint analysis of these loci over, for example, all the available ENCODE TF ChIP-seq datasets (~ 880 based on <https://www.encodeproject.org>) to identify groups of loci that are similar to the +9.5 element represents one potential application. In the MBASIC framework, the binding states are governed by a clustering structure, which groups genomic loci with similar overall binding states across experiments together. For the E-box-GATA composite elements example, in addition to the binding states for each candidate loci across experiments, MBASIC also reports a clustering of loci based on the binding states. The cluster with the +9.5 loci harbors candidate E-box-GATA elements to follow up [25].

Let n_k denote the number of experimental replicates for the k -th condition. We denote the observation for the i -th locus under condition k for the l -th replicate by Y_{ikl} , for $1 \leq i \leq I$, $1 \leq k \leq K$, and $1 \leq l \leq n_k$. We assume that a latent state is associated with the i -th locus and the k -th condition. θ_{iks} is the indicator for the state to be s , where s takes values in a discrete state-space $\{1, \dots, S\}$. In a ChIP-seq experiment, we typically have $S = \{1, 2\}$, where $\theta_{ik1} = 1$ or $\theta_{ik2} = 1$ indicates that the i -th locus is unenriched (unbound) or

enriched (bound) under condition k , respectively. Our model consists of two key components. The first component, *state-space mapping*, assumes the following distribution of Y_{ikl} conditional on θ_{ik} :

$$(Y_{ikl}|\theta_{iks} = 1) \stackrel{i.i.d.}{\sim} f_s(\cdot|\mu_{kls}, \sigma_{kls}, \gamma_{ikls}),$$

where f_s is a density function with parameters μ_{kls} , σ_{kls} , and γ_{ikls} denotes covariates encoding known information for locus i . Note that γ_{ikls} carries information related to how the counts for unenriched loci arise (when $\theta_{ik} = 0$), i.e., data from control Input experiments, GC content, and mappability [21]. In this paper, we take f_s to be Log-normal distribution to represent ChIP-seq read counts after potential normalization for mappability and GC content:

$$(\log(Y_{ikl} + 1)|\theta_{iks} = 1) \stackrel{i.i.d.}{\sim} N(\mu_{kls}\gamma_{ikls}, \sigma_{kls}^2), \quad (1)$$

where we utilize conjugate priors $\mu_{kls} \sim N(\xi, \tau^2)$ and $\sigma_{kls}^2 \sim \text{Gamma}(\omega, \nu)$.

The second part of the Bayesian MBASIC model is *state-space clustering*. We assume that the loci can be clustered into J groups denoted by C_1, \dots, C_J , i.e., $\{1, 2, \dots, I\} = C_1 \cup \dots \cup C_J$. Let $z_{ij} = 1$ if the i -th locus belongs to cluster j and 0 otherwise. The states for the loci within the same cluster follow a product multinomial distribution:

$$(\theta_{iks})_{s=1}^S | z_{ij} = 1 \stackrel{i.i.d.}{\sim} \text{Multinomial}(1, (w_{jks})_{1 \leq s \leq S}), \quad \sum_{s=1}^S w_{jks} = 1, \quad (2)$$

with non-informative prior $(w_{jks})_{1 \leq s \leq S} \sim \text{Dir}(1, 1, \dots, 1)$. We further assume a Chinese Restaurant Process [24] as a prior for the number of clusters J . Let α be a hyper-parameter of the model. The first locus forms C_1 at the start and each locus gets assigned to a cluster recursively. Suppose we have assigned loci $1, \dots, i-1$ to J' clusters. The i -th locus is then assigned to $C_{j'}$, $j' \leq J'$ with probability proportional to the size of $C_{j'}$. It can also form a new cluster $C_{j'+1}$ with probability proportional to α . Then, the prior density for a partition with J clusters is

$$f(z_{ij}, i = 1, \dots, I, j = 1, \dots, J) = \alpha^{J-1} \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha + I)} \prod_{j=1}^J \left(\sum_{i=1}^I z_{ij} - 1 \right)!. \quad (3)$$

With these specifications, we can derive the posterior density of the model for parameter estimation. Although the resulting posterior density leads to a Gibbs sampling algorithm, such a Gibbs sampling scheme requires excessive computational time for mixing (data not shown). Therefore, we derive MAD-Bayes algorithm by utilizing small-variance asymptotics.

2.2 MAD-Bayes Algorithm

We further make the following small-variance assumptions for the MBASIC model:

Assumption 1. All data sets have equal variance: $\sigma_{kls}^2 = \sigma^2 \rightarrow 0$.

Assumption 2. For a given cluster and condition, one of the hidden states dominates with $w_{jks} \in \{1 - (S - 1)e^{-\lambda_w/\sigma^2}, e^{-\lambda_w/\sigma^2}\}$ for $\lambda_w > 0$.

Assumption 3. $\alpha = e^{-\lambda_w \lambda_r / 2\sigma^2} \xrightarrow{\sigma^2 \rightarrow 0} 0$ for $\lambda_w, \lambda_r > 0$.

Proposition 1. Under 1, 2, 3, and as $\sigma^2 \rightarrow 0$, the posterior density reduces to

$$\begin{aligned}
& -2\sigma^2 \log \mathbb{P}(\theta, z, \mu, \sigma, w, J|Y) \\
&= \sum_{i=1}^I \sum_{k=1}^K \sum_{l=1}^{n_k} \sum_{s=1}^S \theta_{iks} [\log(y_{ikl} + 1) - \mu_{kls} \gamma_{ikls}]^2 \\
&+ \lambda_w \sum_{i=1}^I \sum_{j=1}^J z_{ij} \left[\sum_{k=1}^K \sum_{s=1}^S (\theta_{iks} - w_{jks})^2 \right] + \lambda_w \lambda_r (J - 1) + \text{Constant} + o(1).
\end{aligned} \tag{4}$$

This proposition implies that the MAP estimate of the MBASIC framework with CRP and Log-normal mixture model is asymptotically equivalent to the solution of the following optimization problem:

$$\begin{aligned}
& \min_{\mu, z, \theta, w, J} \sum_{i=1}^I \sum_{k=1}^K \sum_{l=1}^{n_k} \sum_{s=1}^S \theta_{iks} [\log(y_{ikl} + 1) - \mu_{kls} \gamma_{ikls}]^2 \\
&+ \lambda_w \sum_{i=1}^I \sum_{j=1}^J z_{ij} \left[\sum_{k=1}^K \sum_{s=1}^S (\theta_{iks} - w_{jks})^2 \right] + \lambda_w \lambda_r (J - 1),
\end{aligned} \tag{5}$$

where the objective function can be viewed as a weighted loss function that integrates the state inference error from Log-normal density as the first term, the clustering error as the second term, and the cost for creating new clusters as the third term. Here, $\lambda_w > 0$ and $\lambda_r > 0$ are tuning parameters that ensure that the cluster assignments are non-trivial. The equal variance assumption is inherently quite strong for ChIP-seq data; however, it was recently shown to work well as a first approximation in a differential ChIP-seq analysis context [19]. We next derive the MAD-Bayes algorithm to generate a local solution for this minimization problem (Algorithm 1).

We note that each step of this algorithm does not increase the objective function in Eq. (5), and the updates for w_{jks} 's and μ_{kls} 's minimize the objective function for a fixed configuration of θ_{iks} 's and z_{ij} 's. Moreover, there are finite number of combinations for θ_{iks} 's and z_{ij} 's such that no cluster is empty and all clusters are distinct from one another. With such observations, we conclude the convergence of this algorithm.

Proposition 2. Algorithm 1 converges after a finite number of iterations to a local minimum of the objective function in Eq. (5).

Algorithm: The MAD-Bayes algorithm for the Bayesian MBASIC model.

repeat

1. Update the cluster labels z_{ij} 's. For each $i = 1, \dots, I$, compute the distance between locus i and each existing cluster $j = 1, \dots, J$ as:

$$t_j = \sum_{k=1}^K \sum_{s=1}^S (\theta_{iks} - w_{jks})^2$$

and find the minimal $j_0 = \arg \min t_j$. If $t_{j_0} < \lambda_r$, assign $z_{ij_0} = 1$. Otherwise, generate a new cluster $J + 1$ with a single locus i .

2. Assign the states θ_{iks} 's. For $i = 1, \dots, I$, $k = 1, \dots, K$, and $s = 1, \dots, S$, let

$$s_0 \leftarrow \arg \min_s \sum_{l=1}^{n_k} [\log(y_{ikl} + 1) - \mu_{kls} \gamma_{ikls}]^2 \\ + \lambda_w \sum_{j=1}^J z_{ij} \left[(1 - w_{jks})^2 + \sum_{s' \neq s} w_{jks'}^2 \right]$$

and let $\theta_{iks_0} = 1$, $\theta_{iks} = 0$ for $s \neq s_0$.

3. Update the Log-normal mean parameters μ_{kls} 's. For $k = 1, \dots, K$, $l = 1, \dots, n_k$, and $s = 1, \dots, S$,

$$\mu_{kls} \leftarrow \frac{\sum_{i=1}^I \theta_{iks} \log(y_{ikl} + 1) \gamma_{ikls}}{\sum_{i=1}^I \theta_{iks} \gamma_{ikls}}.$$

4. Update the Multinomial parameters w_{jks} 's. For $j = 1, \dots, J$, $k = 1, \dots, K$, and $s = 1, \dots, S$,

$$w_{jks} \leftarrow \frac{\sum_{i=1}^I z_{ij} \theta_{iks}}{\sum_{i=1}^I z_{ij}}.$$

until *Convergence*;

Algorithm 1. The MAD-Bayes algorithm for the Bayesian MBASIC model.

2.3 Model Initialization

Similar to the EM algorithm variants for HMMs, the MAD-Bayes algorithm for MBASIC also converges to a local solution and hence can be sensitive to initial starting values. We present a guided two-stage initialization strategy for the states and clusters to attenuate the impact of initialization. We start from initialization of the states by minimizing the state inference error (the first term in Eq. (5)), which has a degenerate form if $\lambda_w = 0$:

$$\min_{\mu, \theta} \sum_{i=1}^I \sum_{k=1}^K \sum_{l=1}^{n_k} \sum_{s=1}^S \theta_{iks} [\log(y_{ikl} + 1) - \mu_{kls} \gamma_{ikls}]^2. \quad (6)$$

Therefore, we use Algorithm 1 by setting $\lambda_w = 0$ to initialize θ_{iks} 's and μ_{kls} 's.

We utilize these initial values of θ_{iks} 's and consider three options for the cluster initialization (i.e., z_{ij} 's and w_{iks} 's): K-means, K-means++, and Adaptive K-means++, where the first two require a pre-determined number of clusters J which we discuss in Sect. 2.4. The K-means option runs hard K-means algorithm on the θ_{iks} 's; while the K-means++ option assigns a cluster label to each unit i with probability inversely proportional to its distance to the current clusters $d_i = \sum_{j=1}^J z_{ij} \sum_{k=1}^K \sum_{s=1}^S (\theta_{iks} - w_{jks})^2$. The adaptive K-means initialization uses a K-means++ style, but increases the number of clusters from $J = 1$, until the value of the function in Eq. (7) does not decrease.

2.4 Selecting the Tuning Parameters

We note that the CRP prior for the number of clusters and the small-variance asymptotics assumptions introduce tuning parameters for the MAD-Bayes algorithm (Algorithm 1). Even for the models with one tuning parameter, [22] acknowledged the difficulty in choosing their appropriate values in practice. Hence, we propose an empirically-motivated method for tuning parameter selection. In practice, we don't expect our small-variance assumption $e^{-\lambda_w/\sigma^2} \rightarrow 0$ as $\sigma^2 \rightarrow 0$ to hold rigidly for real data; however, we expect $e^{-\lambda_w/\sigma^2}$ to be small since it represents the prior probability of enrichment. To maintain the relative small value of $e^{-\lambda_w/\sigma^2}$, we set λ_w as $2\hat{\sigma}^2$ with $\hat{\sigma}^2$ obtained by optimization of the first term in Eq. (5):

$$\hat{\sigma}^2 = \min_{\mu, \theta} \sum_{i=1}^I \sum_{k=1}^K \sum_{l=1}^{n_k} \sum_{s=1}^S \theta_{iks} [\log(y_{ikl} + 1) - \mu_{kls} \gamma_{ikls}]^2.$$

Our computational experiments (data not shown) indicate that varying λ_w in the order of $\hat{\sigma}^2$ does not impact model estimation. The λ_r parameter mediates between the clustering error and the cost of the number of clusters for fixed λ_w . We choose a set of candidate λ_r values by considering the conjugacy between λ_r and J . Suppose J is a global minimum of the objective function in Eq. (5), then fixing the θ_{iks} 's, λ_w , λ_r , J minimizes

$$\sum_{i=1}^I \sum_{j=1}^{J'} z_{ij} \left[\sum_{k=1}^K \sum_{s=1}^S (\theta_{iks} - w_{jks})^2 \right] + \lambda_r (J - 1). \quad (7)$$

Therefore, we let

$$L(J') = \min_{z, w} \left\{ \sum_{i=1}^I \sum_{j=1}^{J'} z_{ij} \left[\sum_{k=1}^K \sum_{s=1}^S (\theta_{iks} - w_{jks})^2 \right] \right\},$$

with $L(J) - L(J + 1) \leq \lambda_r \leq L(J - 1) - L(J)$ (Fig. 4). Algorithm 2 applies this idea to choose a list of candidate λ_r values up to the square root of total number of instances.

Algorithm: Choosing m candidate values of the tuning parameter λ_r .

1. Compute the surrogate values of $L(J')$ for $1 \leq J' \leq \lfloor \sqrt{I} \rfloor := J_{\max}$,
2. Let $\lambda'_j = (L(j - 1) - L(j + 1))/2$ for $2 \leq j \leq J_{\max} - 1$
3. Choose $\frac{1}{m+2}$ -th, $\frac{1}{m+2}$ -th, \dots , $\frac{m}{m+2}$ -th quantile in the $\{\lambda'_j\}$ as candidate values.
4. Given a selected λ_r , choose the initial number of clusters as $J \leftarrow \arg \min_j |\lambda'_j - \lambda_r|$.

Algorithm 2. Algorithm for choosing m candidate λ_r values.

Finally, we use the Silhouette score [27], which has been successfully used for evaluating goodness of fit in clustering, across these values of the tuning parameters.

3 Results

3.1 Computational Experiments

We designed computational experiments to evaluate MAD-Bayes MBASIC in settings where the underlying truth is known. In our experiments, we considered I user-specified loci (e.g., promoters from I genes, binding sites of a transcription factor, or peaks from a ChIP-seq experiment). Given multiple simulated ChIP-seq datasets, there are different “baseline” methods for performing these loci-focused analysis. Therefore, in addition to MBASIC, we considered such alternative approaches that practitioners might adopt.

- **MBASIC:** The EM algorithm on the full MBASIC model, where singleton, i.e., unclusterable loci, are also taken into account.
- **SE-HC:** A two-stage method with first **State Estimation** on individual datasets (i.e., conventional peak calling), and then combining the results by hierarchical clustering on the posterior probabilities of the states $\hat{\theta}_{iks} = P(\theta_{iks} = 1|Y)$ from the first stage.
- **SE-MC:** A two-stage method with first **State Estimation** on individual datasets (i.e., conventional peak calling), and then combining the results by mixture clustering on the binarized results $\theta_{iks_0}^* = 1$, where $s_0 = \arg \max_s P(\theta_{iks} = 1|Y)$ from the first stage.

- **PE-MC**: A two stage method with first **P**arameter **E**stimation on individual datasets to determine the state-specific observations distributions (e.g., distributions of the read counts), and then combining the results by simultaneous state inference and mixture clustering. This is essentially similar to MBASIC, except that state-specific densities are fixed and not updated at every iteration.

The alternatives to MBASIC 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 state-space clustering. For example, SE-HC corresponds to overlapping user-loci with the peak sets from the ENCODE project and generating and clustering the binary overlap or peak confidence profiles of the loci. In contrast, PE-MC is analogous to estimating the distributional parameters of state-space for each individual experiment separately and then clustering with these fixed distributions as in [6, 28]. 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 data set specific parameters [28–30].

For the MAD-Bayes algorithm, we evaluated all the three clustering initializations: Adaptive K means, K means, and Kmeans++. The MAD-Bayes algorithm automatically selects the number of clusters. We used the Silhouette score for SE-HC to accommodate hierarchical clustering and used Bayesian Information Criterion for the other methods. The experiments utilized $I = 4,000$ genomic loci, $J = 10$ clusters, and $K = 20$ experimental conditions. For each condition, the number of replicates, n_k , were drawn from 1 to 3 with probabilities (0.3, 0.5, 0.2). The clustering concentration parameter was simulated from non-informative prior $\alpha \sim \text{Dir}(0.1, \dots, 0.1)$. The state probabilities, w_{jks} s, were simulated from $\text{Dir}(1, \dots, 1)$. The Log-normal parameters were set as follows: the mean was simulated from $N(2*s, 0.05^2)$, where s represented the state label; and the standard error was set to 0.5. We considered four scenarios by varying the number of states S between 2 and 4, and the proportion of singleton loci as $\zeta = 0, 0.4$. Here, singletons represented loci with overall ChIP-seq enrichment profile different than the clusters, i.e., unclusterable locus, and introduced noise to the model. Results for each setting were summarized over 10 simulated datasets. We compared the algorithms in terms of run-time, state-space inference (identifying whether or not each locus is bound), and also the clustering structure via the adjusted Rand index [31].

Figure 2(a) displays run-time comparisons of the methods and indicates that all three implementations of the MAD-Bayes algorithm are about 100 times faster than the EM on full MBASIC and the PE-MC algorithm, and about 10 times faster than the two-step SE-HC and SE-MC algorithms. This speed improvement is significant and makes it possible for the MBASIC framework to scale up. For example, MAD-Bayes can process $I = 100,000$ and $K = 2000$ (e.g., 100,000 DNase accessible regions in the genome across all the available

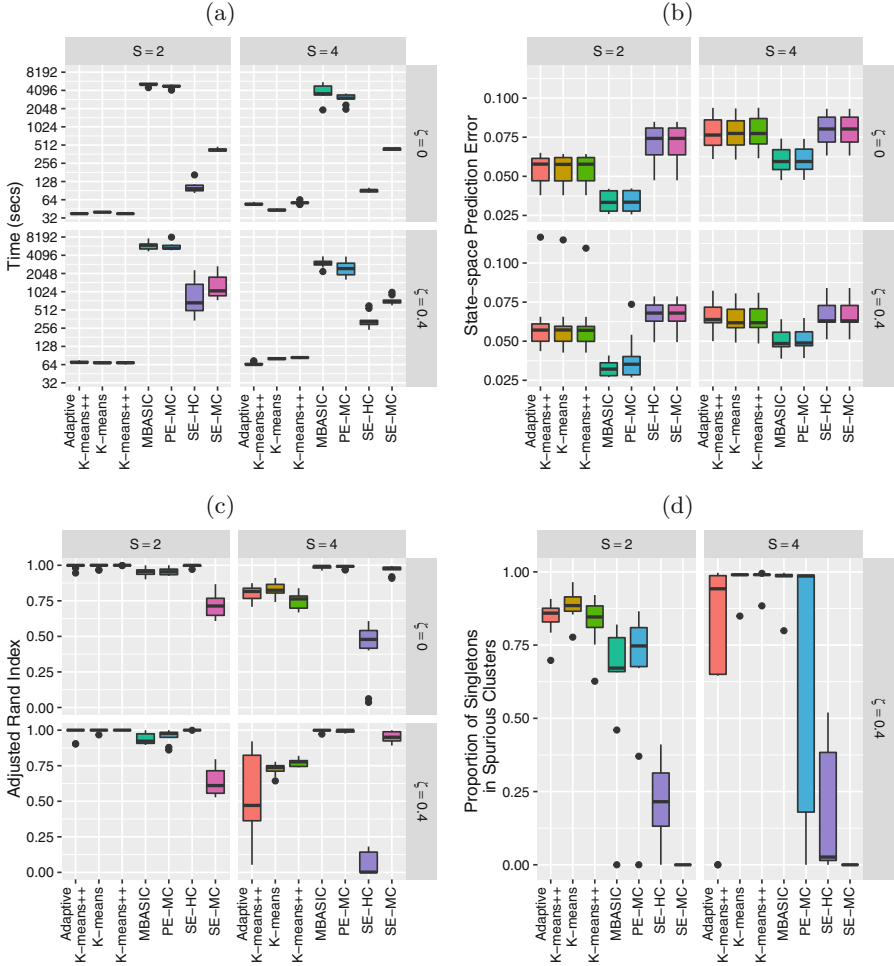


Fig. 2. (a) Run-time comparisons on a 64 bit machine with Intel Xeon 3.0 GHz processor and 64GB of RAM and 8 cores. (b) State-space prediction error. (c) Clustering accuracy based on the adjusted Rand index. (d) Clustering assignments of the singletons when $\zeta = 0.4$.

ENCODE ChIP-seq data) in about 6 hours while the EM algorithm on full MBASIC requires more than a week.

We also observe that speed up in run time does not come at a significant loss in accuracy. Figure 2(b) compares state-space prediction errors of the algorithms and indicates that while MAD-Bayes MBASIC does not perform as accurately as the EM algorithm on full MBASIC and PE-MC, it performs significantly better than SE-HC and SE-MC algorithms, both of which would be the baseline choices for many practitioners. Existence of singleton genomic loci deteriorate

performance of all the algorithms. When there are no singletons, MAD-Bayes with varying cluster initializations perform the best (Fig. 2(c)). When $\zeta = 0.4$ indicating that 40 % genomic loci do not belong to any cluster, the MAD-Bayes algorithm tends generate extra, i.e., spurious, clusters for such loci (Fig. 2(d)) instead of forcing them into other clusters. As a result, the true clusters are largely preserved and less polluted by singletons (Fig. 5) compared to other methods which do not handle singletons (PE-MC, SE-HC, SE-MC).

3.2 Application to Histone ChIP-Seq Data from GM12878 Cells

The key inference question for the MBASIC framework is identifying the enrichment patterns for a given set of user-specified loci across large sets of ChIP-seq datasets and grouping these loci to elucidate similarities and differences. From this point of view, the MBASIC framework is more loci-focused and not directly comparable with any of the available joint analysis methods that can handle large datasets. However, to get a general sense of how MBASIC would compare with ChromHMM [10] and its computationally efficient version Spectacle [12], we analyzed ChIP-seq data of 8 histone marks (H3k4me1, H3k4me2, H3k4me3, H3k9ac, H3k27ac, H3k27me3, H3k36me3, and H4k20me1 from GM12878 cells) from the ENCODE project. Raw data and peak calls for these marks are available at <https://www.encodeproject.org/>. We used the 9038 peaks on chr 18 from the ENCODE uniform processing pipeline as the input loci to MAD-Bayes MBASIC and fixed the number of clusters as 20 since Spectacle identified robust number of chromatin states across multiple chromatin modification datasets as 20. As a result, we also set the number of emission states in chromHMM as 20.

We then performed pairwise comparisons of all the three approaches by matching their clusters/states via maximizing the sum of Jaccard index [32]. We reordered the cluster/state labels of MAD-Bayes and Spectacle according to their agreement with ChromHMM. For example, MAD-Bayes cluster “C1” and Spectacle emission state “E1” are both matched to ChromHMM emission state “E1”; however, this does not necessarily indicate that these two are the best matches between MAD-Bayes and Spectacle.

Figure 3(a) displays that the overall agreements between MAD-Bayes vs. Spectacle and MAD-Bayes vs. ChromHMM follow the same trend with the degree of agreement between Spectacle vs. ChromHMM, which we think of as the baseline agreement since they are both HMM based. In particular, for the emission states with agreement between Spectacle vs. ChromHMM, the corresponding MAD-Bayes clusters also have higher agreement with these. When there is large discrepancy between Spectacle vs. ChromHMM, the MAD-Bayes clusters tend to agree with results from one of the methods. For example, MAD-Bayes “C2” agrees better with Spectacle, and MAD-Bayes “C18” overlaps better with ChromHMM. Figure 3(b) and (c) display comparisons of MAD-Bayes MBASIC to ChromHMM and Spectacle, respectively. We observe that some of MAD-Bayes clusters are distributed over multiple clusters of ChromHMM and Spectacle, e.g., MAD-Bayes cluster “C5” overlaps with the “E12”, “E13”, “E14” of both ChromHMM and Spectacle. This overall agreement indicates that the

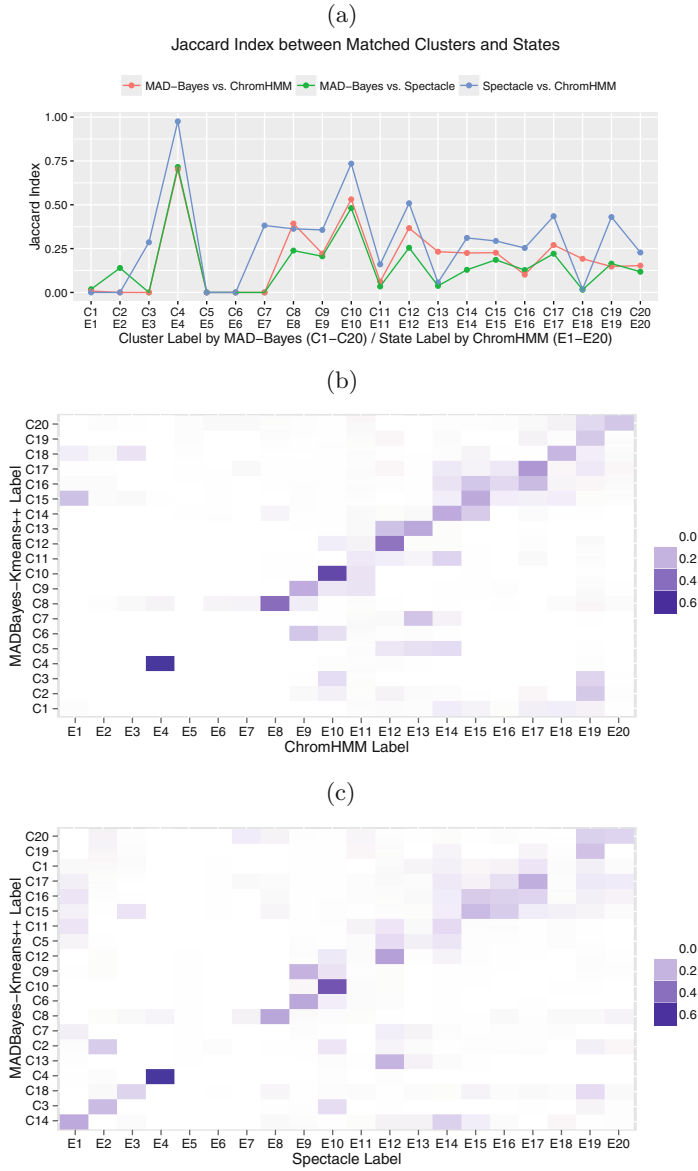


Fig. 3. (a) Comparison of clusters and state labels between MAD-Bayes, Spectacle, and ChromHMM. (b) Jaccard index between MAD-Bayes clusters and ChromHMM states. (c) Jaccard index between MAD-Bayes clusters and Spectacle states. The diagonal blocks indicate agreement between clusters and states; MAD-Bayes clusters and Spectacle states are ordered according to their overlap with the ChromHMM states.

clustering task of MAD-Bayes on the histone marks is reasonable even though it is using selected loci and is not accounting for local dependencies inherent among genomic loci with broad histone marks.

4 Discussion

In this paper, we derived a MAD-Bayes algorithm by developing a Bayesian version of the MBASIC model. Our evaluations indicated that MAD-Bayes MBASIC significantly improves the computational time without sacrificing accuracy. We also observed that even though MAD-Bayes MBASIC does not have a built-in mechanism for singletons (unclusterable loci), it groups singletons as additional clusters and minimizes their effect on other more coherent clusters.

We developed MAD-Bayes MBASIC as a fast method for querying large sets (1000s) of ChIP-seq data with user-specified large sets of loci. This represents the first application of the MAD-Bayes framework in a large scale genome regulation context. From a practical point of view, we showed that this approach is both more efficient and powerful than using individual analysis of each datasets and clustering them with an off-the-shelf method such as hierarchical clustering or finite mixture models. From an algorithmic point of view, we developed an empirical method for selecting tuning parameters. This improves the current state-of-the-art for MAD-Bayes implementations since they lack principled methods for tuning parameter selection. The MBASIC framework offers flexibility in a number of aspects of experimental design, such as different numbers of replicates under individual experimental conditions. This is a relatively important point because many peak callers will operate separately on individual peaks sets or handle two jointly [33] leaving the reconciliation of peaks over multiple replicates to the user. Our current derivation of the MAD-Bayes relied on Log-normal distribution; however, it can be extended to larger class of exponential family distributions via the Bregman divergence [34]. Such extensions are likely to foster its use with other genomic data types such as RNA-seq, DNase-seq, and Methyl-seq, where both state-space estimation and clustering of similar loci pose significant challenges.

Appendix

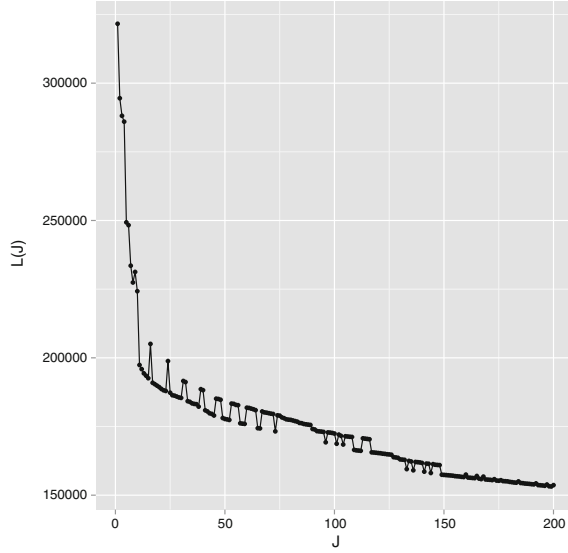


Fig. 4. A graphical interpretation of the conjugacy between λ_r and J . We use the K-means initialization to compute surrogate values for $L(J)$ for a large collection of $J \geq 1$. The λ_r value that can yield J clusters in the global solution must satisfy: $\sup_{J' > J} \frac{L(J) - L(J')}{J - J'} \leq \lambda_r \leq \inf_{J' > J} \frac{L(J') - L(J)}{J' - J}$. When λ_r satisfies this condition, a line with slope $-\lambda_r$ passing through $(J, L(J))$ on the graph should be tangent to the trace of all $L(J)$ values. Although using the surrogate $L(J)$ values can lead to the curve connecting the $L(J)$ values to be con-convex, making the solution for λ_r not hold for some J , we can use a convex approximation to the trace of $L(J)$ so that so that a λ_r exists for each J . A simpler approach is to order the $L(J)$ from largest to smallest and require the following condition for λ_r . $L(J) - L(J + 1) \leq \lambda_r \leq L(J - 1) - L(J)$. Algorithm 2 essentially applies this idea to select the λ_r values. Each J corresponds to a λ_r of value $[L(J - 1) - L(J + 1)]/2$ that satisfies the conjugacy inequality. The algorithm essentially tries to identify the range of λ_r that leads up to \sqrt{I} number of clusters.

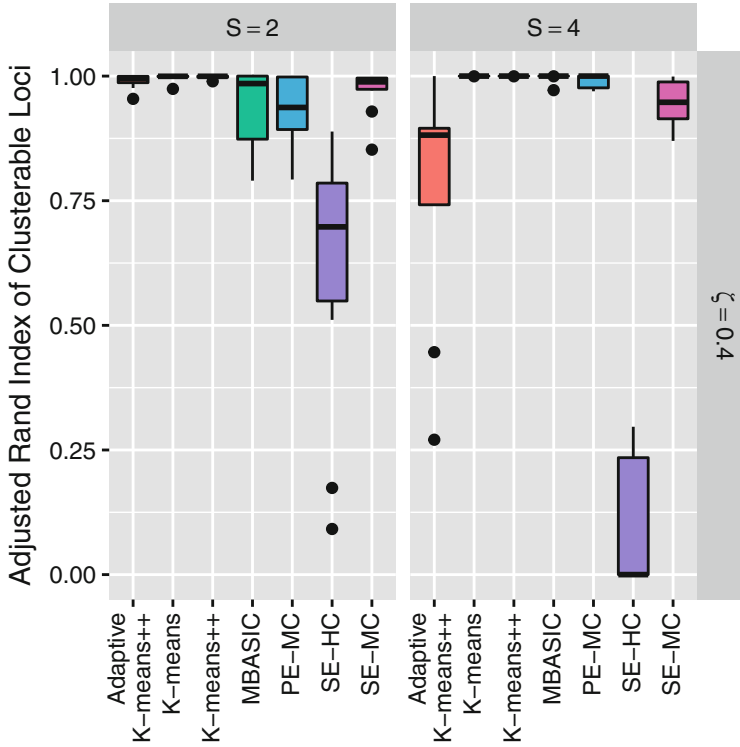


Fig. 5. Comparison of the clustering accuracy with the adjusted Rand index by excluding the singleton loci.

References

1. Zuo, C., Hewitt, K.J., Bresnick, E.H., Keleş, S.: A hierarchical framework for state-space matrix inference and clustering. *Ann. Appl. Stat.* (Revised)
2. The ENCODE project consortium: an integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012)
3. Roadmap epigenomics consortium: integrative analysis of 111 reference human epigenomes. *Nature* **518**(7539), 317–330 (2015)
4. Bardet, A.F., He, Q., Zeitlinger, J., Stark, A.: A computational pipeline for comparative ChIP-seq analyses. *Nat. Protoc.* **7**(1), 45–61 (2012)
5. Bao, Y., Vinciotti, V., Wit, E., AC't Hoen, P.: Accounting for immunoprecipitation efficiencies in the statistical analysis of ChIP-seq data. *BMC Bioinform.* **14**(1), 169 (2013)
6. Zeng, X., Sanalkumar, R., Bresnick, E.H., Li, H., Chang, Q., Keleş, S.: jMOSAICS: joint analysis of multiple ChIP-seq datasets. *Genome Biol.* **14**, R38 (2013). Highly accessed. An R package for joint analysis of multiple ChIP-seq datasets. Available in Bioconductor <http://bioconductor.org/packages/2.12/bioc/html/jmosaics.html>

7. Kuan, P.F., Chung, D., Pan, G., Thomson, J., Stewart, R., Keleş, S.: A statistical framework for the analysis of ChIP-Seq data. *J. Am. Stat. Assoc.* **106**, 891–903 (2011). Software available on Galaxy <http://toolshed.g2.bx.psu.edu/> and also on Bioconductor <http://bioconductor.org/packages/2.8/bioc/html/mosaics.html>
8. Bao, Y., Vinciotti, V., Wit, E., 't Hoen, P.: Joint modeling of ChIP-seq data via a Markov random field model. *Biostatistics* **15**(2), 296–310 (2014)
9. Chen, K.B., Hardison, R., Zhang, Y.: dCaP: detecting differential binding events in multiple conditions and proteins. *BMC Genomics* **15**(9), 1–14 (2014)
10. Ernst, J., Kellis, M.: Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**(8), 817–825 (2010)
11. Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., Noble, W.S.: Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012)
12. Song, J., Chen, K.C.: Spectacle: fast chromatin state annotation using spectral learning. *Genome Biol.* **16**(1), 33 (2015)
13. Sohn, K.A., Ho, J.W.K., Djordjevic, D., Jeong, H.H., Park, P.J., Kim, J.H.: hiHMM: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics*, bttv117 (2015)
14. Liang, K., Keleş, S.: Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics* **28**(1), 121–122 (2012). Available in Bioconductor (<http://www.bioconductor.org/packages/2.12/bioc/html/DBChIP.html>)
15. Mahony, S., Edwards, M.D., Mazzoni, E.O., Sherwood, R.I., Kakumanu, A., Morrison, C.A., Wichterle, H., Gifford, D.K.: An integrated model of multiple-condition ChIP-Seq data reveals predeterminants of Cdx2 binding. *PLoS Comput. Biol.* **10**(3), e1003501 (2014)
16. Song, Q., Smith, A.D.: Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics* **27**, 870–1 (2011)
17. Ferguson, J.P., Cho, J.H., Zhao, H.: A new approach for the joint analysis of multiple ChIP-seq libraries with application to histone modification. *Stat. Appl. Genet. Mol. Biol.* **11**(3), Article 1 (2012)
18. Taslim, C., Huang, T., Lin, S.: DIME: R-package for identifying differential ChIP-seq based on an ensemble of mixture models. *Bioinformatics* **27**(11), 1569–70 (2011)
19. Ji, H., Li, X., Wang, Q.F., Ning, Y.: Differential principal component analysis of ChIP-seq. *Proc. Nat. Acad. Sci. U.S.A.* **110**(17), 6789–6794 (2013)
20. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B Met.* **39**, 1–38 (1977)
21. Zuo, C., Keleş, S.: A statistical framework for power calculations in ChIP-seq experiments. *Bioinformatics* **30**(6), 853–860 (2014)
22. Broderick, T., Kulis, B., Jordan, M.: MAD-Bayes: MAP-based asymptotic derivations from Bayes. In: *Proceedings of the 30th International Conference on Machine Learning* (2013)
23. Blackwell, D., MacQueen, J.B.: Ferguson distributions via Polya urn schemes. *Ann. Stat.* **1**(2), 353–355 (1973)
24. Aldous, D.J.: Exchangeability and related topics. In: Hennequin, P.L. (ed.) *École de Probabilités de Saint-Flour XIII*, vol. 1117, pp. 1–198. Springer, Heidelberg (1983)
25. Hewitt, K.J., Kim, D.H., Devadas, P., Prathibha, R., Zuo, C., Sanalkumar, R., Johnson, K.D., Kang, Y.A., Kim, J.S., Dewey, C.N., Keleş, S., Bresnick, E.: Hematopoietic signaling mechanism revealed from a stem/progenitor cell cistrome. *Mol. Cell* **59**(1), 62–74 (2015)

26. Johnson, K.D., Hsu, A., Ryu, M.J., Boyer, M.E., Keleş, S., Zhang, J., Lee, Y., Holland, S.M., Bresnick, E.H.: Cis-element mutation in a GATA-2-dependent immunodeficiency syndrome governs hematopoiesis and vascular integrity. *J. Clin. Inv.* **10**(122), 3692–3704 (2012)
27. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987)
28. Wei, Y., Li, X., Wang, Q.F., Ji, H.: iASeq: integrative analysis of allele-specificity of protein-DNA interactions in multiple ChIP-seq datasets. *BMC Genomics* **13**, 681 (2012)
29. 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., Fietze, S., Fu, Y., Gertz, J., Grubert, F., Harman, 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., Snyder, M.: Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**(7414), 91–100 (2012)
30. Wei, Y., Tenzen, T., Ji, H.: Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics* **16**(1), 31–46 (2015)
31. Rand, W.M.: Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**(336), 846–850 (1971)
32. Tan, P.N., Steinbach, M., Kumar, V.: Cluster analysis: basic concepts and algorithms. In: *Introduction to Data Mining*, chap. 8 (2005)
33. Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al.: ChIP-seq guidelines and practices of the encode and modencode consortia. *Genome Res.* **22**(9), 1813–1831 (2012)
34. Banerjee, A.: Clustering with Bregman divergences. *J. Mach. Learn. Res.* **6**, 1705–1749 (2005)

<http://www.springer.com/978-3-319-31956-8>

Research in Computational Molecular Biology

20th Annual Conference, RECOMB 2016, Santa Monica,
CA, USA, April 17-21, 2016, Proceedings

Singh, M. (Ed.)

2016, XXXIV, 274 p. 64 illus., 41 illus. in color., Softcover

ISBN: 978-3-319-31956-8