SIMULTANEOUS ANALYSIS OF LARGE SCALE DATASETS IN DIFFERENT CHIP-SEQ PROBLEM SETTINGS

by

Kailei Chen

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON
2018

Date of final oral examination: 08/31/2018

The dissertation is approved by the following members of the Final Oral Committee:

Sündüz Keleş, Professor, Statistics
Michael A. Newton, Professor, Statistics
Sushmita Roy, Assistant Professor, Biostatistics and Medical Informatics
Colin Dewey, Professor, Biostatistics and Medical Informatics
Qiongshi Lu, Assistant Professor, Biostatistics and Medical Informatics

SIMULTANEOUS ANALYSIS OF LARGE SCALE DATASETS IN DIFFERENT CHIP-SEQ PROBLEM SETTINGS

Kailei Chen

Under the supervision of Professor Sündüz Keleş

At the University of Wisconsin-Madison

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a common genomics tool for studying regulation of transcription. Analyses of ChIP-seq data typically concern: i) binding state inference, which aims at detecting the DNA loci occupied by the transcription factor of interest; and ii) allele-specific binding/histone modification analysis, which incorporates the heterozygous Single Nucleotide Polymorphisms (SNPs) information in diploid organisms. Current approaches for both problems target at individual datasets independently.

Chapter 2 introduces a MAP-based Asymptotic Derivations from Bayes (MAD-Bayes) ([11]) method based on strong assumptions in MBASIC ([60]) framework. This results in a K-means-like optimization algorithm which converges rapidlyThe fast-converging nature enables exploring multiple initialization schemes and flexibility in tuning. Computational experiments and application shows that MAD-Bayes MBASIC improves computational efficiency without sacrificing accuracy in estimation performance.

In Chapter 3, I extends the MAD-Bayes MBASIC to the allele-specific analysis setting, via a variance-stabilizing transformation, enabling the method to apply discrete distributions in both binding state and allele-specific binding problems. Application to the allele-specific analysis of transcription factor binding displays higher power and accuracy of the joint approach compared to individual dataset level approach. The first systematic analysis of allelic imbalance of histone modifications reveals properties unique in allele-specific histone modification. By connecting allele-specific histone modification results to allele-specific expression data, we detected SNP loci as the potential candidate for further studies in the regulatory mechanisms.

ACKNOWLEDGMENTS

First, I would like to extend my deepest gratitude to my advisor, Professor Sündüz Keleş. This dissertation would be impossible without her support and guidance. She continually and convincingly conveyed a passion for research and sharpness to new ideas, which will always be inspiring me in lifetime.

I would also like to thank all our current and previous members in Keleş' Research Group. I benefited a lot from the generous suggestions and insightful questions. In particular, I acknowledge the valuable help from Chandler Zuo for the brilliance and effort in our collaborative works.

I appreciate all of our committee members' help. Professor Michael Newton, Professor Kam-Wah Tsui and Professor Sushmita Roy have made very enlightening comments to my work. Thanks to Professor Colin Dewey and Professor Qiongshi Lu for the precious time.

Last, I want to thank my family, for the love and encougagement. I am so lucky to have them.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Appendix
Figure Page

Appendix
Figure Page

ABSTRACT

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a common genomics tool for studying regulation of transcription. Analyses of ChIP-seq data typically concern: i) binding state inference, which aims at detecting the DNA loci occupied by the transcription factor of interest; and ii) allele-specific binding/histone modification analysis, which incorporates the heterozygous Single Nucleotide Polymorphisms (SNPs) information in diploid organisms. Current approaches for both problems target at individual datasets independently.

Chapter 2 introduces a MAP-based Asymptotic Derivations from Bayes (MAD-Bayes) ([11]) method based on strong assumptions in MBASIC ([60]) framework. This results in a K-means-like optimization algorithm which converges rapidlyThe fast-converging nature enables exploring multiple initialization schemes and flexibility in tuning. Computational experiments and application shows that MAD-Bayes MBASIC improves computational efficiency without sacrificing accuracy in estimation performance.

In Chapter 3, I extends the MAD-Bayes MBASIC to the allele-specific analysis setting, via a variance-stabilizing transformation, enabling the method to apply discrete distributions in both binding state and allele-specific binding problems. Application to the allele-specific analysis of transcription factor binding displays higher power and accuracy of the joint approach compared to individual dataset level approach. The first systematic analysis of allelic imbalance of histone modifications reveals properties unique in allele-specific histone modification. By connecting allele-specific histone modification results to allele-specific expression data, we detected SNP loci as the potential candidate for further studies in the regulatory mechanisms.

Chapter 1

Introduction

1.1 Background

The state-of-art technology, Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has revolutionized the detection of biological signals including protein-DNA interaction of transcription factors (TFs) and histone modifications with significantly accelerated data generation speed with affordable expenses. ChIP-seq helps to identify how chromatin-associated proteins influence phenotype-affecting mechanisms. Determining how proteins interact with DNA in DNA regulation is essential for fully understanding many biological processes and disease states.

Large consortia (e.g., ENCODE ([49]), REMC ([40])) as well as investigator-initiated projects generated large collections of ChIP-seq data profiling multiple transcription factors 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 overestimation 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.

Moreover, availability of large collections of ChIP-seq datasets enables integrative analysis of datasets under different biological conditions, i.e., TFs/histone modifications/treatments/tissues/cells. For example, [20] and [54] each studied human regulatory network by analyzing over 100 TF datasets. [57] identified clustered TF binding patterns in cancer cells by investigating 565 different TF datasets. A common feature of these studies is that genomic loci that exhibit signals (bound loci) are identified from individual datasets, and loci with the same combinatorial patterns across the experiments are clustered. Overall, two general problems emerge from the joint analysis of ChIP-seq datasets:

Binding State Inference. A variety of problems such as TF regulatory network and chromatin state annotation analysis fall into this category. For example, about 2,000 transcription factors in human genome collaborate in a diversity of ways to perform the complex functioning of regulation of 20,000 genes dynamically in different tissues ([51]). Such a combinatorial scheme makes it critical to elucidate the association patterns among different transcription factors. While the number of possible patterns grows exponentially with number of conditions, studies in epigenetic marks ([45]) show that actual combination patterns of epigenetic marks in real human data scales only as a polynomial with degree around 1 or 2, suggesting that a small handful of protein binding co-association patterns can actually account for the majority of human genome.

Allele-specific Binding/Histone modification Inference. Allele-specific Histone Modification (ASHM) and Allele-specific Binding (ASB) inference incorporate genetic variation due to SNPs in analysis of ChIP-seq datasets in diploid organisms and complement the standard analysis of peak calling ([36]). The key inference is to detect SNP loci that can perturb histone modifications or transcription factor (TF) binding and generate allelic imbalance between the two parental strands, and assign preferred parental alleles to each loci.

Allele-specific analysis of ChIP-seq data involves two key steps. The first is quantification of read counts for each allele ([53, 42, 27]). The second step concerns inferring the level of allelic imbalance from the allele specific counts. Binomial and Beta-Binomial tests are two common methods for count data in allele-specific event such as allele-specific expression or allele-specific

binding analysis for individual datasets ([42, 13, 38, 15]). iASeq ([56]) uses a Bayesian hierarchical model to learn correlation patterns of allele-specificity across conditions by focusing on co-occurrence of allelic-specificity across datasets.

Despite the vast availability of ChIP-seq datasets and the growing demand for more comprehensive and fast joint analysis method, only a few joint methods have emerged and most of the methods do not scale up to 100s or 1000s. Matrix Based Analysis for State-space Inference and Clustering (MBASIC) is a versatile framework both estimates the underlying state-space (e.g., bound vs. unbound) and also studies association between genomic loci by grouping loci with similar state patterns together, while its Expectation-Maximization based estimation structure hinders its applicability with large numbers of loci and samples. In this thesis, I present a fast method MAD-Bayes MBASIC for large scale datasets, which handles the two problems in a unified way.

1.2 Overview of the Chapters

Chapter 2 adopts a small variance asymptotics framework for MBASIC and derives a K-means-like MAD-Bayes algorithm under strong assumptions for binding state inference problem. The fast-converging nature enables exploring multiple initialization schemes and flexibility in tuning. Computational experiments and application shows that MAD-Bayes MBASIC improves computational efficiency without sacrificing accuracy in estimation performance.

In Chapter 3, I extends the MAD-Bayes MBASIC to the allele-specific analysis setting, by employing a variance-stabilizing transformation to maintain small variances. Application to the allele-specific analysis of transcription factor binding displays higher power and accuracy of the joint approach compared to individual dataset level approach. I also provides the first systematic analysis of allelic imbalance of histone modifications. Associating allele-specific histone modification results to allele-specific expression data, and we detected SNP loci as the potential candidate for further studies in the regulatory mechanisms.

Chapter 2

A MAD-Bayes Algorithm for State-Space Inference and Clustering with Application to Querying Large Collections of ChIP-Seq Data Sets [1]

2.1 Introduction

Many large consortia (e.g., ENCODE ([49]), REMC ([40])) 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 overestimation 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 ([9, 7]). Among the available joint analysis methods, jMOSAiCS ([58]) builds on ChIP-seq peak-caller MOSAiCS ([28]) and incorporates a

[1] The manuscript for this chapter is published in [61]. Method in this chapter is implemented in the `R` package `MBASIC` available at `http://github.com/KaileiChen/mbasic`.

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 ([14]) uses a three-step log-likelihood ratio test to jointly identify binding events in multiple experimental conditions. ChromHMM ([17]) and Segway ([22]) 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 ([45]) provided a transformative improvement of ChromHMM by utilizing spectral learning for parameter estimation in HMMs. hiHMM ([44]) 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 only TF ([30, 32]), only histone ([17, 22, 45, 46, 19]), or both ([7, 58]) types of ChIP-seq data (large amounts of both are awaiting in the public databases!), (ii) whether or not they rely on a priori analysis of individual datasets ([17, 45, 30, 32, 19]), (iii) whether or not they focus on differential occupancy and can handle very few numbers of conditions ([30, 48, 23]), (iv) whether or not they can scale up to 100s to 1000s of datasets. These approaches, with the potential exception of ([45]), 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 ([16]) for parameter estimation. Furthermore, none of these approaches accommodate querying of multiple datasets for *selected* loci. Their analysis results can only be used to ”annotate” user-specified loci without any notion of uncertainty.

We recently introduced MBASIC ([60]) 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 [28, 59] 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 ([11]). Specifically, we consider a mixture of Log-normal distributions for the state-specific observations with a Chinese Restaurant Process (CRP) ([10, 2]) 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.

2.2 Method

We begin our exposition with an overall description of the Bayesian MBASIC model (Fig.2.1) 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.

2.2.1 The MBASIC Model

We consider I genomic loci of interest, indexed from $i = 1, \cdots, I$, 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

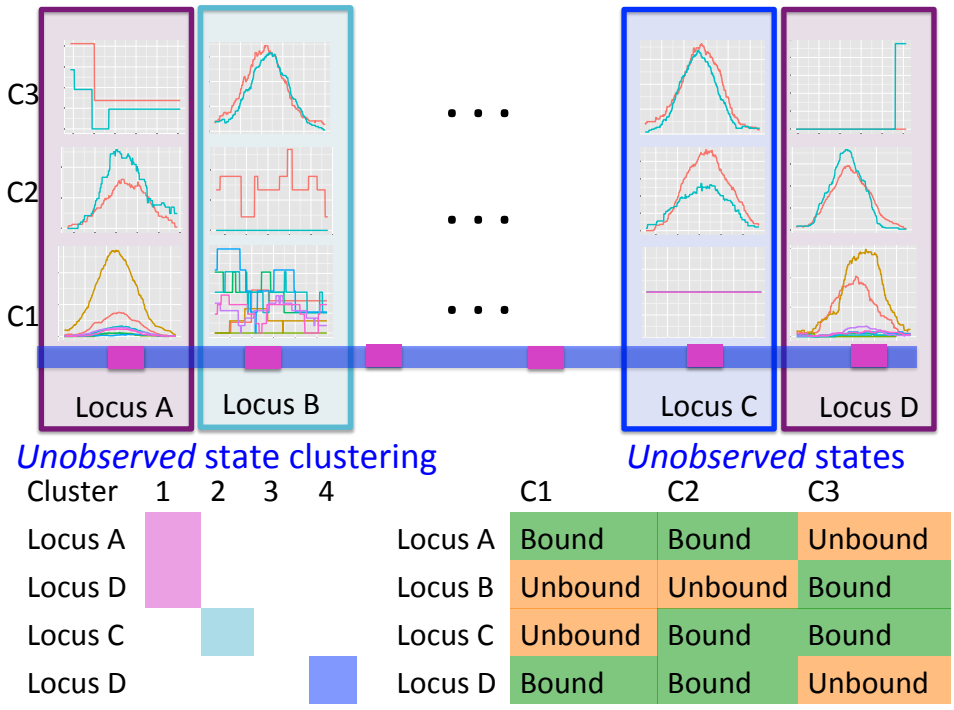


Figure 2.1: Overview of the MBASIC modeling framework. Curves within each panel depict different replicates under the experimental conditions C1, C2, and C3. Loci A and D are in the same cluster.

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 [21], 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, and myeloid leukemias ([24])). 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 ~ 102K of them. Joint analysis of these loci over, for example, all the available ENCODE TF ChIP-seq datasets ($\sim$ 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 [21].

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 $\{1, \cdots, S\}$. 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}:

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

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 ([59]). 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:

$$(\log(Y_{ikl}+1)|\theta_{iks}=1) \overset{\text{i.i.d.}}{\sim} N(\mu_{kls}\gamma_{ikls}, \sigma^2_{kls}), \tag{2.1}$$

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

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 $C_1, \cdots, C_J$, i.e., $\{1, 2, \cdots, I\} = C_1 \cup \cdots \cup C_J$. 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:

$$(\theta_{iks})_{s=1}^{S}|z_{ij}=1 \overset{\text{i.i.d.}}{\sim} \text{Multinomial}\left(1, (w_{jks})_{1\le s\le S}\right), \quad \sum_{s=1}^{S} w_{jks} = 1, \tag{2.2}$$

with non-informative prior $(w_{jks})_{1\le s\le S} \sim \text{Dir}(1,\ 1, \cdots,\ 1)$. We further assume a Chinese Restaurant Process (CRP, [2]) 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 $1,\ \cdots,\ i-1$ to J' clusters. The i-th locus is then assigned to $C_{j'},\ j' \le J'$ 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

$$f(z_{ij}, i=1,\cdots,I, j=1,\cdots,J) = \alpha^{J-1}\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+I)}\prod_{j=1}^{J}\left(\sum_{i=1}^{I} z_{ij} - 1\right)!. \tag{2.3}$$

With these specifications, we can derive the posterior density (2.4) of the model for parameter estimation.

$$\begin{aligned}\mathbb{P}(\theta, z, \mu, \sigma, w, J|Y) \propto & \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K}\prod_{s=1}^{S} w_{jks}^{\theta_{iks}z_{ij}} \times \alpha^(J-1)\prod_{j=1}^{J}(\sum_{i=1}^{I} z_{ij} - 1)! \\ & \times \prod_{k=1}^{K}\prod_{l=1}^{n_k}\prod_{s=1}^{S}\frac{1}{\tau}e^{-\frac{(\mu_{kls}-\xi)^2}{2\tau^2}}\frac{1}{\sigma_{kls}^{2(\omega+1)}}e^{-\frac{1}{\sigma^2_{kls}}} \\ & \times \prod_{i=1}^{I}\prod_{k=1}^{K}\prod_{s=1}^{S}\left[\prod_{l=1}^{n_k}\frac{1}{\sigma_{kls}}e^{-\frac{(log(y_{ikl}+1)-\mu_{kls}\gamma_{ikls})^2}{2\sigma^2_{kls}}}\right]^{\theta_{iks}}\end{aligned} \tag{2.4}$$

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.

2.2.2 MAD-Bayes MBASIC

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

1. All variances in the Log-normal distributions are small and similar in scale. Let $\sigma^2 = \max_{k,l,s} \sigma^2_{kls}$, then $\sigma^2 \to 0$, and

$$\frac{\sigma^2}{\min \sigma^2_{kls}} = O(1).$$

2. There is a dominant hidden state for each fixed cluster and condition, i.e. $w_{jks} \in \{1-(S-1)e^{-\lambda_w/\sigma^2}, e^{-\lambda_w/\sigma^2}\}$ for $\lambda_w > 0$.

3. Associating the degree of dominance in each cluster and condition to the CRP clustering prior, then reparameterize the concentration parameter as $\lambda = -2\sigma^2 \log(\alpha)$ and $\lambda = \lambda_w \lambda_r$, then $\alpha = e^{-\lambda_w \lambda_r / 2\sigma^2} \overset{\sigma^2 \to 0}{\longrightarrow} 0$ for $\lambda_w, \lambda_r > 0$.

Proposition 2.1. *Under assumptions (1)-(3), as $\sigma^2 \to 0$, the posterior density reduces to*

$$\begin{aligned} & -2\sigma^2 \log \mathbb{P}(\theta, z, \mu, \sigma, w, J|Y) \\ =& \sum_{i=1}^{I} \sum_{k=1}^{K} \sum_{l=1}^{n_k} \sum_{s=1}^{S} \theta_{iks} [\log(y_{ikl}+1) - \mu_{kls}\gamma_{ikls}]^2 \frac{\sigma^2}{\sigma^2_{kls}} \\ +& \lambda_w \sum_{i=1}^{I} \sum_{j=1}^{J} z_{ij} \left[\sum_{k=1}^{K} \sum_{s=1}^{S} (\theta_{iks} - w_{jks})^2 \right] + \lambda_w \lambda_r (J-1) + \text{Constant} + o(1). \end{aligned} \tag{2.5}$$

Proof. The log posterior density Eqn. (2.4) can be written as:

$$\begin{aligned}\log P(\theta,z,\mu,\sigma,w,J|y) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\sum_{s=1}^{S}\theta_{iks}[\log(y_{ikls}+1)-\mu_{kls}\gamma_{ikls}]^2\frac{\sigma^2}{\sigma^2_{kls}}\\ &+\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{s=1}^{S}\theta_{iks}z_{ij}\log w_{jks}+\log\alpha(J-1)-\frac{NS}{\nu\sigma^2}\frac{\sigma^2}{\sigma^2_{kls}}\\ &-[(\omega+1)NS+NI/2]\log\sigma^2-(\omega+1)\sum_{k=1}^{K}\sum_{l=1}^{n_k}\sum_{s=1}^{S}\log\left(\frac{\sigma^2_{kls}}{\sigma^2}\right)\\ &-\frac{1}{2}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{s=1}^{S}\theta_{iks}\sum_{l=1}^{n_k}\log\left(\frac{\sigma^2_{kls}}{\sigma^2}\right)+O(1),\end{aligned} \tag{2.6}$$

where $N=\sum_{k=1}^{K}n_k$ is the total number of replicates, and all terms unrelated to σ^2 ends in $O(1)$. Therefore, as $\sigma^2\to 0$,

$$\begin{aligned}-2\sigma^2\log P(\theta,z,\mu,\sigma,w,J|y) &= \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\sum_{s=1}^{S}\theta_{iks}[\log(y_{ikls}+1)-\mu_{kls}\gamma_{ikls}]^2\frac{\sigma^2}{\sigma^2_{kls}}\\ -2\sigma^2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{s=1}^{S}\theta_{iks}z_{ij}\log w_{jks} &-2\sigma^2\log\alpha(J-1)\\ +(\omega+1)\sum_{k=1}^{K}\sum_{l=1}^{n_k}\sum_{s=1}^{S}\sigma^2\log\frac{\sigma^2_{kls}}{\sigma^2} &+\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{s=1}^{S}\theta_{iks}\sum_{l=1}^{n_k}\sigma^2\log\left(\frac{\sigma^2_{kls}}{\sigma^2}\right)\\ +\frac{NS\sigma^2}{\nu}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\sum_{s=1}^{S}\frac{\sigma^2}{\sigma^2_{kls}} &+o(1)\end{aligned} \tag{2.7}$$

According to assumption 1, the last three terms end up as $o(1)$.

For the term including w_{jks}, note that by assumption 2, for each j,k, $w_{jks}=e^{-\frac{\lambda_w}{\sigma^2}}$ except for one s. Thus, when $\theta_{iks'}=1$ for $s'=\arg\max_s w_{jks}$, either

$$-2\sigma^2\sum_{s}\theta_{iks}\log w_{jks}=-2\sigma^2\log[1-(S-1)e^{-\frac{\lambda_w}{\sigma^2}}]=o(1),$$

$$\sum_{s}(\theta_{iks}-w_{jks})^2=S(S-1)e^{-\frac{2\lambda_w}{\sigma^2}}=o(1);$$

or,

$$-2\sigma^2 \sum_s \theta_{iks} \log w_{jks} = 2\lambda_w,$$

$$\sum_s (\theta_{iks} - w_{jks})^2 = S(S-1)e^{-\frac{2\lambda_w}{\sigma^2}} - 2Se^{\frac{\lambda_w}{\sigma^2}} + 2 = 2 + o(1).$$

In both cases, we have

$$-2\sigma^2 \sum_s \theta_{iks} \log w_{jks} = \lambda_w \sum_s (\theta_{iks} - w_{jks})^2 + o(1). \tag{2.8}$$

Note that $-2\sigma^2 \log(\alpha) = \lambda_w \lambda_r$, we prove Eqn. (2.5).

□

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:

$$\begin{aligned} \min_{\mu,z,\theta,w,J} & \sum_{i=1}^{I} \sum_{k=1}^{K} \sum_{l=1}^{n_k} \sum_{s=1}^{S} \theta_{iks} [\log(y_{ikl}+1) - \mu_{kls}\gamma_{ikls}]^2 \frac{\sigma^2}{\sigma^2_{kls}} \\ & + \lambda_w \sum_{i=1}^{I} \sum_{j=1}^{J} z_{ij} \left[\sum_{k=1}^{K} \sum_{s=1}^{S} (\theta_{iks} - w_{jks})^2 \right] + \lambda_w \lambda_r (J-1), \end{aligned} \tag{2.9}$$

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 ([23]). We next derive the MAD-Bayes algorithm to generate a local solution for this minimization problem (Algorithm. 2.1).

We note that each step of this algorithm does not increase the objective function in Eqn. (2.9), 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

Algorithm 2.1 The MAD-Bayes algorithm for MBASIC model.

repeat

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

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

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

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

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

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

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

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

Step 5. Update the Log-normal variance parameters σ_{kls}'s. For $k = 1, \cdots, K$, $l = 1, \cdots, n_k$, and $s = 1, \cdots, S$,

$$\sigma_{kls}^2 = \frac{\sum_{i=1}^{I} \theta_{iks}[\log(y_{ikl} + 1) - \mu_{kls}\gamma_{ikls}]^2}{\sum_{i=1}^{I} \theta_{iks}},$$

and $\sigma^2 = \max_{k,l,s} \sigma_{kls}^2$.

until Convergence.

cluster is empty and all clusters are distinct from one another. With such observations, we conclude the convergence of this algorithm.

Proposition 2.2. Algorithm 2.1 *converges after a finite number of iterations to a local minimum of the objective function in Eqn. (2.9).*

2.2.3 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 Eqn. (2.9)), which has a degenerate form if $\lambda_w = 0$:

$$\min_{\mu,\theta} \sum_{i=1}^{I} \sum_{k=1}^{K} \sum_{l=1}^{n_k} \sum_{s=1}^{S} \theta_{iks} [\log(y_{ikl} + 1) - \mu_{kls}\gamma_{ikls}]^2 \frac{\sigma^2}{\sigma_{kls}^2}. \tag{2.10}$$

Therefore, we repeat Step. 2, Step. 3 and Step 5 in Algorithm 2.1 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 Section 2.2.4. 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 Eqn. (2.11) does not decrease.

2.2.4 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 the MAD-Bayes algorithm (**Algorithm 2.1**). Even

for the models with one tuning parameter, [11] 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 $e^{-\lambda_w/\sigma^2} \to 0$ as $\sigma^2 \to 0$ to hold rigidly for real data; however, we expect $e^{-\lambda_w/\sigma^2}$ to be small since it represents the prior probability of not having a particular state. To maintain the relative small value of $e^{-\lambda_w/\sigma^2}$, we set λ_w as $2\hat{\sigma}^2$, where $\hat{\sigma}^2$ is obtained by letting $\sigma^2_{kls} = \sigma^2$ and optimizing the first term in Eqn. (2.9):

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

Our computational experiments 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 Eqn. (2.9), then fixing the θ_{iks}'s, λ_w, λ_r, J minimizes

$$\sum_{i=1}^{I} \sum_{j=1}^{J'} z_{ij} \left[\sum_{k=1}^{K} \sum_{s=1}^{S} (\theta_{iks} - w_{jks})^2 \right] + \lambda_r (J - 1). \tag{2.11}$$

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. A.1). **Algorithm 2.2** applies this idea to choose a list of candidate λ_r values up to the square root of total number of instances.

Algorithm 2.2 Algorithm for choosing m candidate λ_r values.

Step 1. Compute the surrogate values of $L(J')$ for $1 \leq J' \leq \lfloor \sqrt{I} \rfloor := J_{\max}$,

Step 2. Let $\lambda'_j = (L(j-1) - L(j+1))/2$ for $2 \leq j \leq J_{\max} - 1$

Step 3. Choose $\frac{1}{m+2}$-th, $\frac{1}{m+2}$-th, $\cdots$, $\frac{m}{m+2}$-th quantile in the $\{\lambda'_r\}$ as candidate values.

Step 4. Given a selected λ_r, choose the initial number of clusters as $J \leftarrow \arg\min_j |\lambda'_j - \lambda_r|$.

Finally, we use the Silhouette score ([41]), which has been successfully used for evaluating goodness of fit in clustering, across these values of the tuning parameters.

2.3 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.

1. Prob-HC: A two-stage method with first modeling on individual datasets, then passing the estimated **Prob**abilities of the states $\tilde{\theta}_{iks} = P(\theta_{iks} = 1|Y)$ into **H**ierarchical **C**lustering to combine the results.

2. State-MC: A two-stage method with first modeling on individual datasets, then passing the estimated **State**s $\theta^*_{iks_0} = 1$, where

   $s_0 = \arg\max_s P(\theta_{iks} = 1|Y)$ into **M**ixture **C**lustering to combine the results.

3. Param-SIMC: A two stage method with first modeling on individual datasets, then passing the estimated **Param**eters in the state-specific observations distributions (e.g., distributions of the read counts) into simultaneous **S**tate **I**nference and **M**ixture **C**lustering to combine the results. (Similar to MBASIC, except that state-specific densities are fixed and not updated at every iteration.)

4. MBASIC: The EM algorithm on the MBASIC model. The full model is robust, allowing singletons, i.e., unclusterable loci.

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, Prob-HC corresponds overlapping user-loci with the peak sets from the ENCODE project and generating and clustering the binary overlap profiles of

the loci. In contrast, Param-SIMC is analogous to estimating the distributional parameters of state-space for each individual experiment separately and then clustering with these fixed distributions as in [58, 56]. 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 ([56, 20, 55]).

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 Prob-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 $\alpha \sim \text{Dir}(0.1, \cdots, 0.1)$. The state probabilities, w_{jks}s, were simulated from $\text{Dir}(1, \cdots, 1)$. 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 ([39]).

Fig. 2.2(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 Param-SIMC algorithm, and about 10 times faster than the two-step Prob-HC and State-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 ENCODE ChIP-seq data) in about 6 hours while the EM algorithm on full MBASIC requires more than a week.

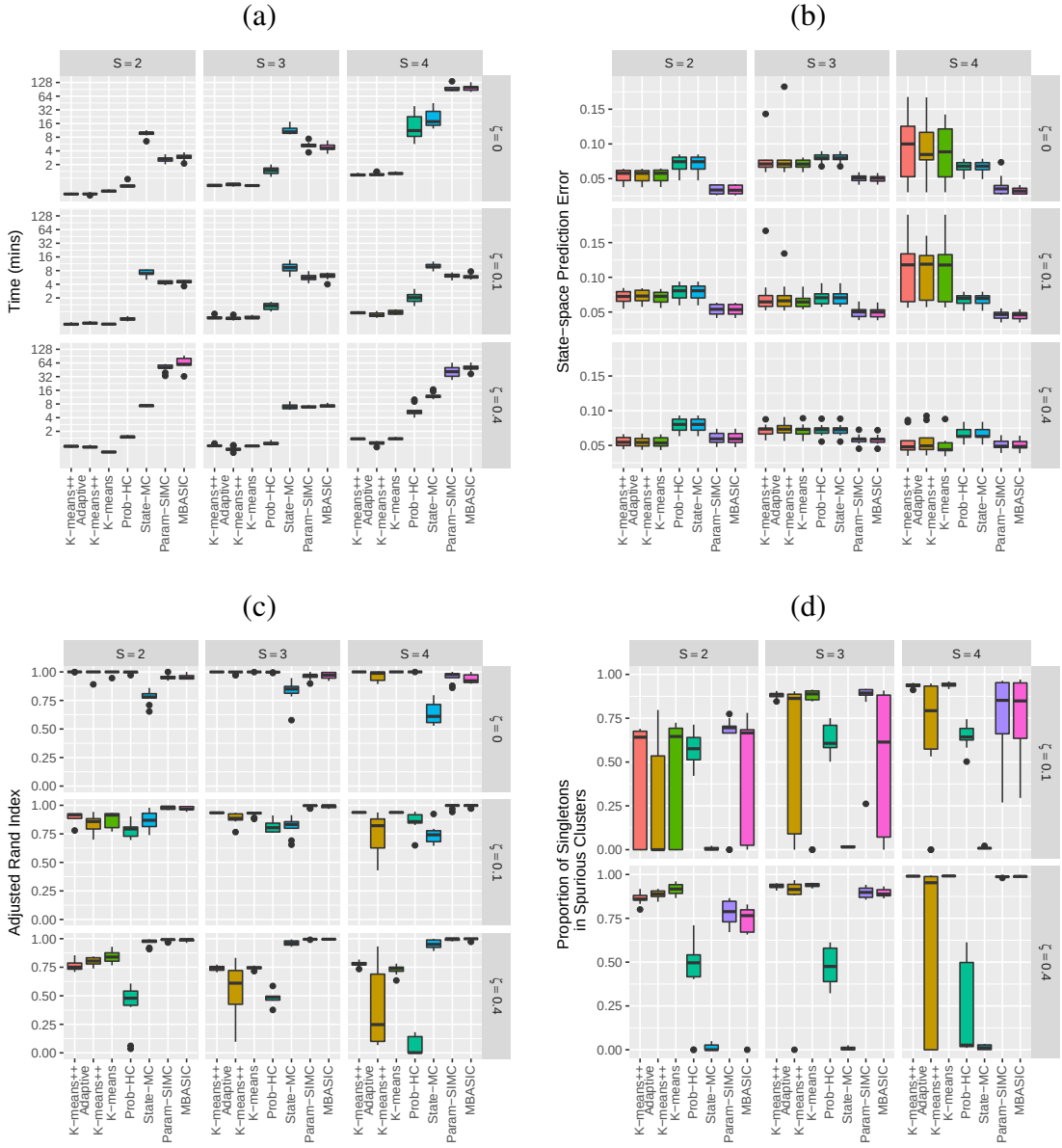


Figure 2.2: S: Number of potential states. ζ: Proportion of loci do not belong to any true cluster. (a) Run-time comparisons on a 64 bit machine with Intel Xeon 3.0GHz processor and 64GB of RAM and 8 cores. (b) State-space prediction error. (c) Clustering accuracy based on the adjusted Rand index. (d) Clustering assignments of unclusterable loci when 10% ($\zeta = 0.1$) and 40% ($\zeta = 0.4$) of the loci do not belong to a cluster.

We also observe that speed up in run time does not come at a significant loss in accuracy. Fig. 2.2(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 Param-SIMC, it performs significantly better than Prob-HC and State-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. 2.2(c) and Fig. 2.2(a)). 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. A.2(b)) instead of forcing them into other clusters. As a result, the true clusters are largely preserved and less polluted by singletons (Fig. 2.2 (d)) compared to other methods which do not handle singletons (Param-SIMC, Prob-HC, State-MC).

2.4 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 ([17]) and its computationally efficient version Spectacle ([45]), 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 ([47]). 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.

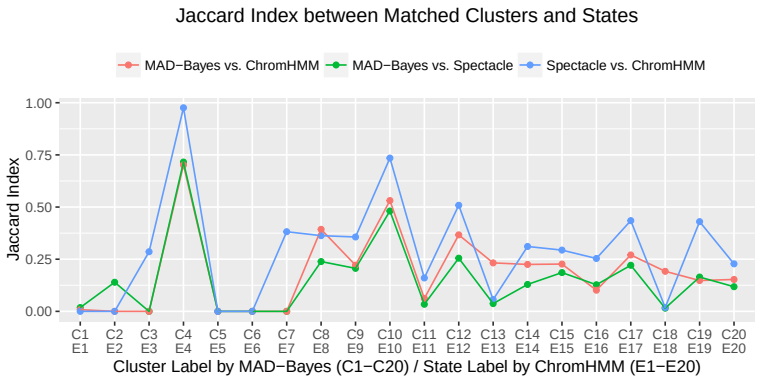
Fig. 2.3(a) displays that the overall agreement between MAD-Bayes vs. Spectacle and MAD-Bayes vs. ChromHMM follow the same trend with the degree of agreement between Spectacle vs. ChromHMM, which we think of as the baseline agreement since they are both HMM based. In particular, for the emission states 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. Figs. 2.3(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 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.

2.5 Conclusions and Discussion

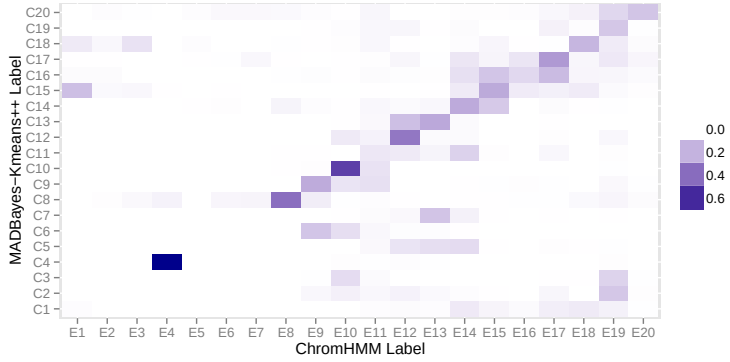
We derived a MAD-Bayes algorithm by developing a Bayesian version of the MBASIC model. Our evaluations indicated that MAD-Bayes MBASIC significantly improves the computational time without sacrificing accuracy. We also observed that even though MAD-Bayes MBASIC does not have a built-in mechanism for 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

(a)



(b)



(c)

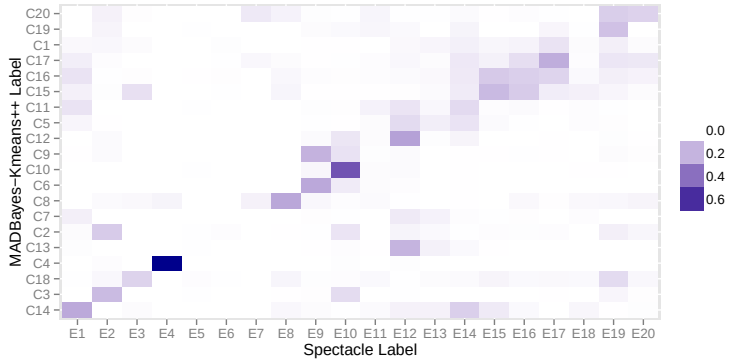


Figure 2.3: (a) Comparison of clusters and state labels between MAD-Bayes, Spectacle, and ChromHMM. (b) Jaccard index between MAD-Bayes clusters and ChromHMM states. (c) Jaccard index between MAD-Bayes clusters and Spectacle states. The diagonal blocks indicate agreement between clusters and states; MAD-Bayes clusters and Spectacle states are ordered according to their overlap with the ChromHMM states.

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 ([29]) 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 ([6]). 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.

Chapter 3

Allele-Specific ChIP-seq Analysis with MAD-Bayes MBASIC[1]

3.1 Introduction

Approaches such as the quantitative trait loci (QTL)-mapping have helped researchers to identify a large number of genetic loci with potential roles in gene regulation [31, 35]. However, we still have a limited understanding of regulatory mechanisms because the mere mapping of loci associated with quantitative traits or gene expression cannot provide inference at that level. Elucidating the causal variants and the regulatory elements that they impact is a critical challenge [18]. A reasonable strategy is to focus on some particular mediators, detect genetic variants capable of perturbing these mediators, and carry out studies to connect changes in mediators to variations in quantitative traits/expression (Fig 3.1).

Histone modifications constitute one of the widely studied class of mediators ([34]). Some functions of modifications include enabling or disrupting chromatin contacts, recruiting or prohibiting binding of transcription factors, and utilizing enzymatic activities to affect biological processes such as repair, replication, and recombination. At least 8 distinct classes of modifications have been detected on the four core histones, at over 60 different residues ([26]). The ever-growing list of histone modifications create a vast number of potential functional responses. The fact that histone modifications function in a combinatorial way brings in an extra level of complexity; hence, analysis of histone modifications often involves the joint analysis of multiple datasets ([58, 17, 22, 45, 44]). The emergence of Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has revolutionized the study of transcription factors and histone modifications globally

[1] The manuscript for this chapter is to be submitted.

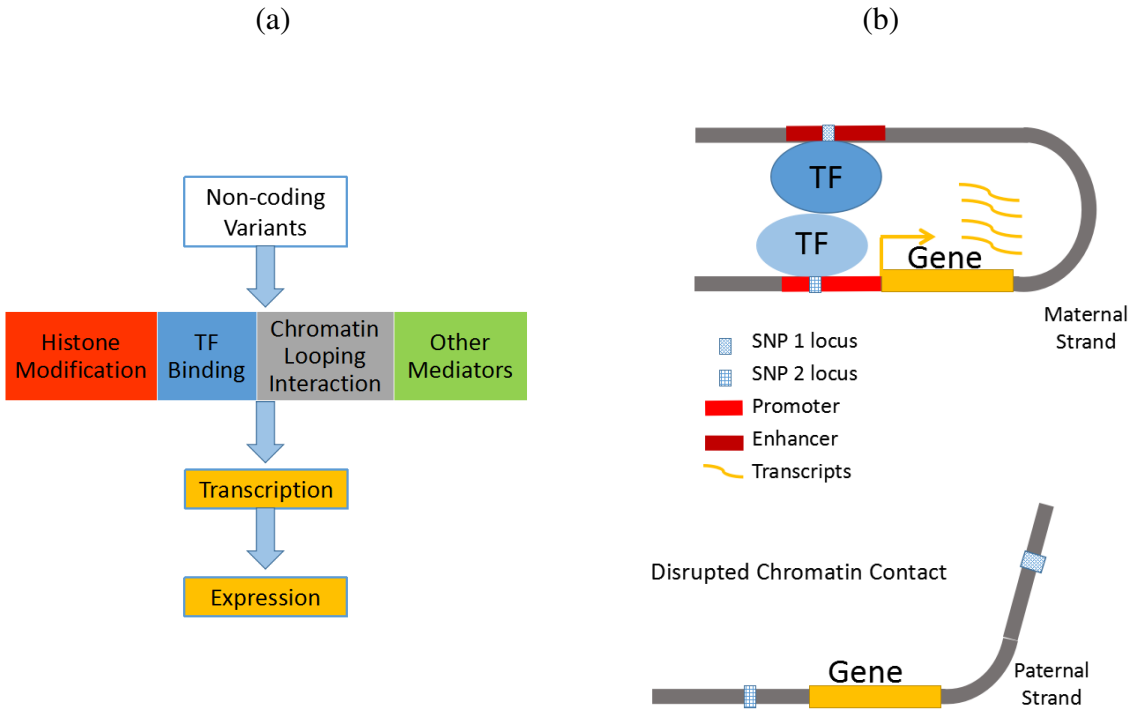


Figure 3.1: (a) Genetic variants impact gene expression through multiple mediators. (b) Pictorial illustration of how an enhancer or promoter SNP can disrupt transcription factor gene interaction. SNPs in enhancer or promoter regions may disrupt required histone modifications and/or transcription factor binding, leading to loss of enhancer-promoter interaction in the paternal strand.

with significantly accelerated data generation speed and affordable expenses. We now have access to thousands of histone modification datasets from large consortium such as ENCODE ([50]) and REMC ([40]).

Allele-specific Histone Modification (ASHM) and Allele-specific Binding (ASB) inference incorporate genetic variation due to SNPs in analysis of ChIP-seq datasets in diploid organisms and complement the standard analysis of peak calling ([36]). The key inference is to detect SNP loci that can perturb histone modifications or transcription factor (TF) binding and generate allelic imbalance between the two parental strands, and assign preferred parental alleles to each loci.

Allele-specific analysis of ChIP-seq data involves two key steps. The first is quantification of read counts for each allele ([53, 42, 27]). The second step concerns inferring the level of allelic imbalance from the allele specific counts. Binomial and Beta-Binomial tests are two common methods for count data in allele-specific event such as allele-specific expression or allele-specific binding analysis for individual datasets ([42, 13, 38, 15]). iASeq ([56]) uses a Bayesian hierarchical model to learn correlation patterns of allele-specificity across conditions by focusing on co-occurrence of allelic-specificity across datasets.

[60] recently introduced MBASIC as a hierarchical framework for querying multiple ChIP-seq datasets jointly for user-specified loci. The joint inference is formulated by both identifying hidden states (e.g. maternal allele-specific, maternal allele-specific or neutral) in individual datasets (*state-space mapping*) as well as identifying groups of loci that cluster across different experiments (*state-space clustering*) based on their profile similarity (Fig. 3.2). To enable application to large number of datasets, [61] developed a small-variance asymptotics framework for MBASIC and derived a K-means-like MAD-Bayes algorithm ([11]) for Log-Normal distribution. By utilizing variance-stabilizing transformation, we adopted this algorithm to discrete distributions such as Binomial, Poisson and enabled joint allele-specific binding or histone modification inference to accommodate multiple datasets. Our computational experiments illustrates that MAD-Bayes MBASIC is computationally efficient at the expense of a small loss in inference and clustering accuracy. Comparing to iASeq, MAD-Bayes MBASIC performs better in terms of clustering accuracy when number of clusters increases.

We applied MAD-Bayes MBASIC to the allele-specific analysis of transcription factor datasets of ETV6, IRF3, and TRIM22 and histone datasets of H3K4me1, H3K4me3 and H3K27ac. Validation of ChIP-seq datasets analysis with in silico impact data of SNPs on transcription factor binding revealed higher power and accuracy of the joint approach compared to individual dataset level analysis. For histone modification data, we connected ASHM results to Allele-specific Expression (ASE) data and detected SNP loci as the potential candidate for further studies in the causal order of regulatory mechanisms. Furthermore, our study provides the first systematic analysis of allelic imbalance of histone modifications, and their integration with the allele-specific expression.

3.2 Method

We start with a brief overview of the MBASIC framework for allele-specific histone modification or transcription factor binding (Fig. 3.2), and then discuss the small variances asymptotics method of MAD-Bayes MBASIC and its limitations in the allele-specific analysis. We extend the MAD-Bayes MBASIC algorithm to accommodate model of violations via a variance-stabilizing transformation.

3.2.1 The MBASIC Model for allele-specific analysis

Taking SNP loci as observational unit of interest, indexed from $i = 1, \cdots, I$, and different histones, TFs, treatments and cell/tissue types as experimental conditions, indexed from $k = 1, \cdots, K$, we assume allelic-specificity as a latent state associated with the i-th locus and the k-th condition. Let θ_{iks} be the indicator for the state to be s, where s takes values in a discrete state-space $\{1, \cdots, S\}$. The state-space $\{1, 2, 3\}$ represents whether the binding event or histone modification is maternal-specific, neutral (no difference between two strands), or paternal-specific. Let n_k denote the number of experimental replicates for the k-th condition. We denote the observed maternal strand counts for the i-th locus under condition k for the l-th replicate by Y_{ikl}, and total counts by X_{ikl}, for $1 \leq i \leq I$, $1 \leq k \leq K$, and $1 \leq l \leq n_k$.

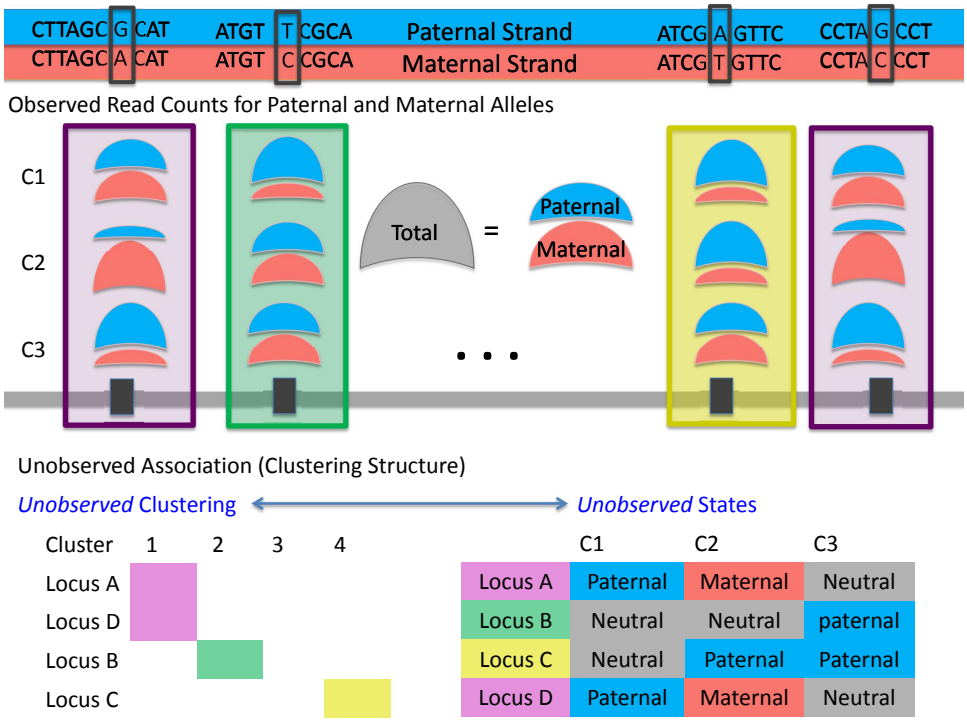


Figure 3.2: Overview of the MBASIC modeling framework for allele-specific histone modification or transcription factor binding. Each panel depicts different datasets under the experimental conditions C1, C2, and C3. Input to MBASIC are loci-level read counts for the maternal and paternal alleles.

For *state-space mapping*, we assume the following mixture distribution of Y_{ikl} conditional on θ_{ik}:

$$(Y_{ikl}|\theta_{iks}=1, X_{ikl}) \overset{\text{i.i.d.}}{\sim} f_s(\cdot|\mu_{kls}, \sigma_{kls}),$$

where f_s is a density function with parameters μ_{kls} and σ_{kls}, denotes covariates encoding known information for locus i.

For *state-space clustering*, we keep using the CRP process as prior.

Before adapting the MAD-Bayes MBASIC to allele-specific analysis setting, we introduce another step for the discrete distributions.

3.2.2 Variance-stabilizing Transformation

While the distribution of normalized read counts from ChIP-seq experiments can be modeled by Log-Normal distribution relatively well ([5, 43]), in the allele-specific binding/histone modification analysis, the count on one strand (paternal or maternal) out of total count (paternal strand count + maternal strand count) is usually assumed to be Binomially distributed. Furthermore, many binding state inference methods model read counts with a Poisson or Zero-Inflated Poisson distribution ([28, 8]). Since assumption 1 about small variances is violated in these models, we extend MBASIC MAD-Bayes to accommodate the two discrete distribution by a variance-stabilizing transformation (VST). Specifically, we utilized the square root transformation of [12] for Poisson read counts. For $Y \sim \text{Poisson}(m)$, $\tilde{Y} = \sqrt{Y + 0.25}$ reduces the variance to

$$\text{Var}(\tilde{Y}) = \frac{1}{4} - \frac{1}{32}m^{-1} + \frac{3}{64}m^{-2} + O(m^{-3}). \tag{3.1}$$

For Binomial read counts, we utilized the inverse Sine transformation of [3]. For $Y' \sim \text{Binomial}(n, p)$, $\tilde{Y'} = \sqrt{n + 0.5}\sin^{-1}\left(\sqrt{\frac{Y+0.375}{n+0.75}}\right)$ reduces the variance to

$$\text{Var}(\tilde{Y'}) = \frac{1}{4} + O(n^{-2}). \tag{3.2}$$

According to Eqn. 3.1 and Eqn. 3.2, the small variances can be maintained for large mean parameter for Poisson and large size parameter for Binomial distributions. We observed that under

the criteria $m \geq 2$ for Poisson distribution and $n \geq 8$ (Fig.) when number of loci in each state is large enough (≥ 30), transformed data is approximately normal with small variances by Central Limit Theorem. After the transformation, the MAD-Bayes algorithm is applied with the transformed read counts.

3.3 Computational Experiments

Similar to the Log-Normal model for binding dtate inference, we compared MAD-Bayes MBASIC to the following methods in the application to simulated data where true model is known.

1. Prob-HC: A two-stage method with first modeling on individual datasets, then passing the estimated **Prob**abilities of the states $\tilde{\theta}_{iks} = P(\theta_{iks} = 1|Y)$ into **H**ierarchical **C**lustering to combine the results.

2. State-MC: A two-stage method with first modeling on individual datasets, then passing the estimated **State**s $\theta^*_{iks_0} = 1$, where

   $s_0 = \arg\max_s P(\theta_{iks} = 1|Y)$ into **M**ixture **C**lustering to combine the results.

3. Param-SIMC: A two stage method with first modeling on individual datasets, then passing the estimated **Param**eters in the state-specific observations distributions (e.g., distributions of the read counts) into simultaneous **S**tate **I**nference and **M**ixture **C**lustering to combine the results. (Similar to MBASIC, except that state-specific densities are fixed and not updated at every iteration.)

4. MBASIC: The EM algorithm on the MBASIC model. The full model is robust, allowing singletons, i.e., unclusterable loci.

The three two-stage methods can still mimic procedures where the association between experimental units are estimated separately from the estimation of individual dataset specific parameters ([56, 20, 55]). For example, Prob-HC is similar to taking p-values from Beta-Binomial tests in ASB analyses and clustering; while Param-SIMC is analogous to estimating the Beta-Binomial

parameters of state-space for each individual experiment separately, and then clustering with these fixed distributions as in [56]. We used the Silhouette score for MAD-Bayes MBASIC and Prob-HC, and Bayesian Information Criterion for the other methods to select the number of clusters,

Generalizing the problem settings, we considered I user-specified loci (e.g., SNP loci or potential binding sites) and nine settings by varying the number of states S between 2, 3 and 4, ($S = 3$ mirrors the real ASB/ASHM problem; we tried different number of states for generalization.) and the proportion of unclusterable loci as $\zeta = 0, 0.1(10\%), 0.4(40\%)$. Here, unclusterable loci represented outlier SNP loci with overall ASB/ASHM patterns different than the clusters and introduced additional noise to the model. We compared the methods in terms of run-time, state-space inference (allele-specific inference), and also the clustering structure via the adjusted Rand index ([39]) with 10 simulated datasets (Fig. 3.3). Let $I = 4,000$ genomic loci, $J = 10$ clusters, and $K = 20$ experimental conditions. For each condition, the number of replicates, n_k, were randomized from 1 to 3 with probabilities $(0.3, 0.5, 0.2)$. The clustering concentration parameter followed a non-informative prior $\alpha \sim \text{Dir}(0.1, \cdots, 0.1)$. The state assignment probabilities of each cluster, w_{jks}'s, were simulated from $\text{Dir}(1, \cdots, 1)$.

The Binomial distribution size parameters (equivalent to total counts in ASB/ASHM when $S = 3$) are drown from a Poisson distribution with mean 10, the probabilities (of maternal-specific) was simulated from $\text{Beta}(3(2s-1), 3(2(S-s)+1))$, where s represented the state label.

MAD-Bayes MBASIC is about 100 times faster than the EM on full MBASIC and the Param-SIMC method, and about 10 times faster than the two-step Prob-HC and State-MC procedures (Fig. 3.3(a)). The significant speeding-up enables the MBASIC framework to simultaneously handle 100s or 1000s of datasets. We tested MAD-Bayes MBASIC on the extreme setting of $I = 100,000$ and $K = 2,000$ (e.g., $100,000$ SNP loci for $2,000$ TFs or histone modifications in different experiments) in about 6 hours while the EM algorithm could not fit MBASIC model in a week.

We also observe that speed up in run time does not come at a significant loss in accuracy. In terms of state inference, while MAD-Bayes MBASIC does not perform as accurately as the

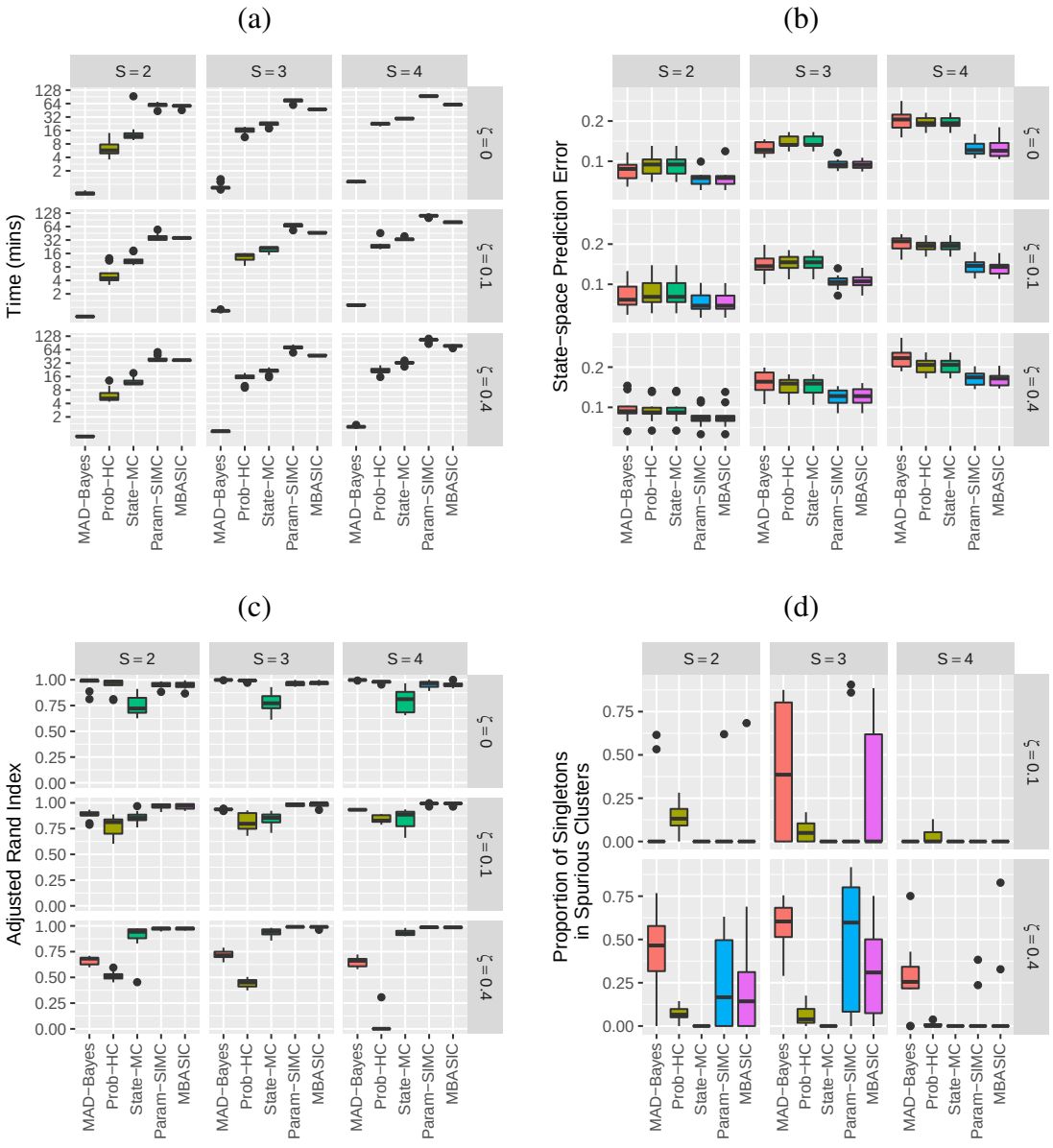


Figure 3.3: Computational experiments with Binomial distribution. S: Number of potential states. ($S = 3$ represents the ASHM/ASB inference setting.) ζ: Proportion of loci do not belong to any true cluster. (a) Run-time comparisons on a 64 bit machine with Intel Xeon 3.0GHz processor and 64GB of RAM and 8 cores. (b) State-space prediction error. (c) Clustering accuracy based on the adjusted Rand index. (d) Clustering assignments of unclusterable loci when 10% ($\zeta = 0.1$) and 40% ($\zeta = 0.4$) of the loci do not belong to a cluster.

EM algorithm on full MBASIC and Param-SIMC, it performs significantly better than Prob-HC and State-MC methods, both of which would be the baseline choices for many practitioners (Fig. 3.3(b)). When there are no unclusterable loci ($\zeta = 0$), MAD-Bayes MBASIC performs the best in clustering (Fig. 3.3(c)). Existence of unclusterable loci impacts performances of all the methods. When $\zeta = 0.1$ and $\zeta = 0.4$, indicating that 10% or 40% genomic loci do not belong to any cluster, while MAD-Bayes MBASIC was not designed to handle unclusterable loci, it tend to generate extra, i.e., spurious, clusters for such loci (Fig. 3.3(d)) instead of forcing them into other clusters. As a result, the true clusters are largely characterized by true member loci and less polluted by model noise, compared to Prob-HC, State-MC and Param-SIMC which do not handle singletons ().

We also tried the MAD-Bayes MBASIC in Poisson models (binding site inference). Given unbound state, from realistic sense of ChIP-seq data, we use a Zero-One-Inflated Poisson with three components to avoid small Poisson means so that the small variance assumption hold for transformed data.

$$Y_{ikl1}|\theta_{ik1} = 1 \sim p^0_{kl} I\{Y = 0\} + p^1_{kl} I\{Y = 1\} + p^*_{kl}\text{Poisson}(\mu_{kl1}).$$

The inflated probabilities were simulated from $p^0 \sim \Gamma(2.27, 0.08)$, $p^1 \sim \Gamma(1.768, 0.058)$ and $p^* = 1 - p^0 - p^1$ (parameters were the average of the fitted zero-one Inflated models of 244 ChIP-seq datasets from ENCODE); mean parameters $\mu_{kl1} \in 2, 3, \cdots, 7$ with probabilities $(0.280, 0.284, 0.320, 0.108, 0.004, 0.004)$, and $\mu_{kls} = \mu_{kl1} + 12(s - 1) + 3\text{Uniform}[-1, 1]$, for $s \geq 2$. We achieved similar results to the ones from Binomial model (Fig. B.1).

3.4 Application to ChIP-seq datasets of GM12878 cells

We analyzed ChIP-seq data of transcription factors, ETV6, IRF3, and TRIM22 of GM12878 cells from ENCODE ([50]) and ChIP-seq data of histone modifications, H3K4me1, H3K4me3 and H3K27ac also of GM12878 cells from [25]. Maternal and paternal genomes were attained from the 1000 Genomes Project ([1]). We followed the AlleleSeq pipeline ([42]) to obtain maternal and paternal strand read counts at all the SNP loci. Since transcription factor datasets had relative low

sequencing depths, we pooled the replicates for each transcription factor but kept the replicates of the histone datasets as separate.

3.4.1 Allele-specific Transcription Factor Binding Analysis

We preprocessed the SNP loci by overlapping them with the union of peaks sets identified by ENCODE from the ChIP-seq experiments. We further filtered SNPs with total counts less than 8 to avoid the small size in the Binomial distribution (Eqn. 3.2), and obtained 3,416 SNPs loci as input to MAD-Bayes MBASIC.

The state-space was set as {"M":Maternal-specific, "N":Neutral, "P":Paternal-specific}. Using Silhouette score as a criteria, MAD-Bayes MBASIC clustered the 3,416 loci into 27 clusters where each cluster coincided with one of the patterns in {"M", "N", "P"}^{3}. We compared the inference results to Binomial ([42]) and Beta-Binomial tests ([13]) followed by FDR control at level 0.1 (denoted as Binomial and Beta-Binomial, respectively). Overall, the MAD-Bayes method recovered almost all the significant allele-specific SNP loci identified by Binomial and Beta-Binomial tests (Table 3.1). According to the three TF datasets, the Beta-Binomial test performed most conservatively while MAD-Bayes MBASIC identified the largest number of allele-specific binding sites. A direct comparison of the states inferred from MAD-Bayes MBASIC and the Binomial and Beta-Binomial test indicated higher sensitivity of MAD-Bayes to identify events with smaller effect size as shown by the ratio of maternal read count to total read count (Fig. 3.4, Fig. B.2, Fig. B.3). Next we focused on the comparison between Binomial test and MAD-Bayes MBASIC since the Beta-Binomial test performed very conservatively by detecting only few number of significant SNP loci for IRF3.

We validated the results of our analysis by a mutational motif analysis tool atSNP ([62]), which utilizes the degree of match between a local sequence and motif position weight matrices (PWMs) to evaluate the allelic-specificity of a binding event in silico. Fig. 3.5 (a) displays a SNP (A/G) locus at Chr 1:149,899,885, which resides within an IRF3 peak. atSNP detected allelic-specificity there in favor of maternal strand sequence, which was also detected by MAD-Bayes method, but not by the Binomial or the Beta-Binomial test.

Table 3.1: Comparison of significant allele-specific binding events detected by MAD-Bayes MBASIC, Binomial and Beta-Binomial tests with FDR < 0.1. Each cell represents number of SNP loci with specific pattern for: (a) ETV6, (b) IRF3, and (c) TRIM22.

(a) IRF3

	MAD-Bayes					MAD-Bayes			
Binomial	"M"	"N"	"P"	Total	Beta-Binomial	"M"	"N"	"P"	Total
"M"	**68**	0	0	68	"M"	**5**	0	0	5
"N"	748	**1,710**	862	3,320	"N"	811	**1,710**	887	3,408
"P"	0	0	**28**	28	"P"	0	0	**3**	3
Total	816	1,710	890	3,416	Total	816	1,710	890	3,416

(b) ETV6

	MAD-Bayes					MAD-Bayes			
Binomial	"M"	"N"	"P"	Total	Beta-Binomial	"M"	"N"	"P"	Total
"M"	**266**	0	0	266	"M"	**46**	0	0	46
"N"	631	**1,694**	571	2,896	"N"	851	**1,694**	786	3,331
"P"	0	0	**254**	254	"P"	0	0	**39**	39
Total	897	1,694	825	3,416	Total	897	1,694	825	3,416

(c) TRIM22

	MAD-Bayes					MAD-Bayes			
Binomial	"M"	"N"	"P"	Total	Beta-Binomial	"M"	"N"	"P"	Total
"M"	**88**	1	0	89	"M"	**11**	0	0	11
"N"	659	**1,685**	926	3,270	"N"	736	**1,686**	967	3,396
"P"	0	0	**57**	57	"P"	0	0	**16**	16
Total	747	16,86	983	3,416	Total	747	16,86	983	3,416

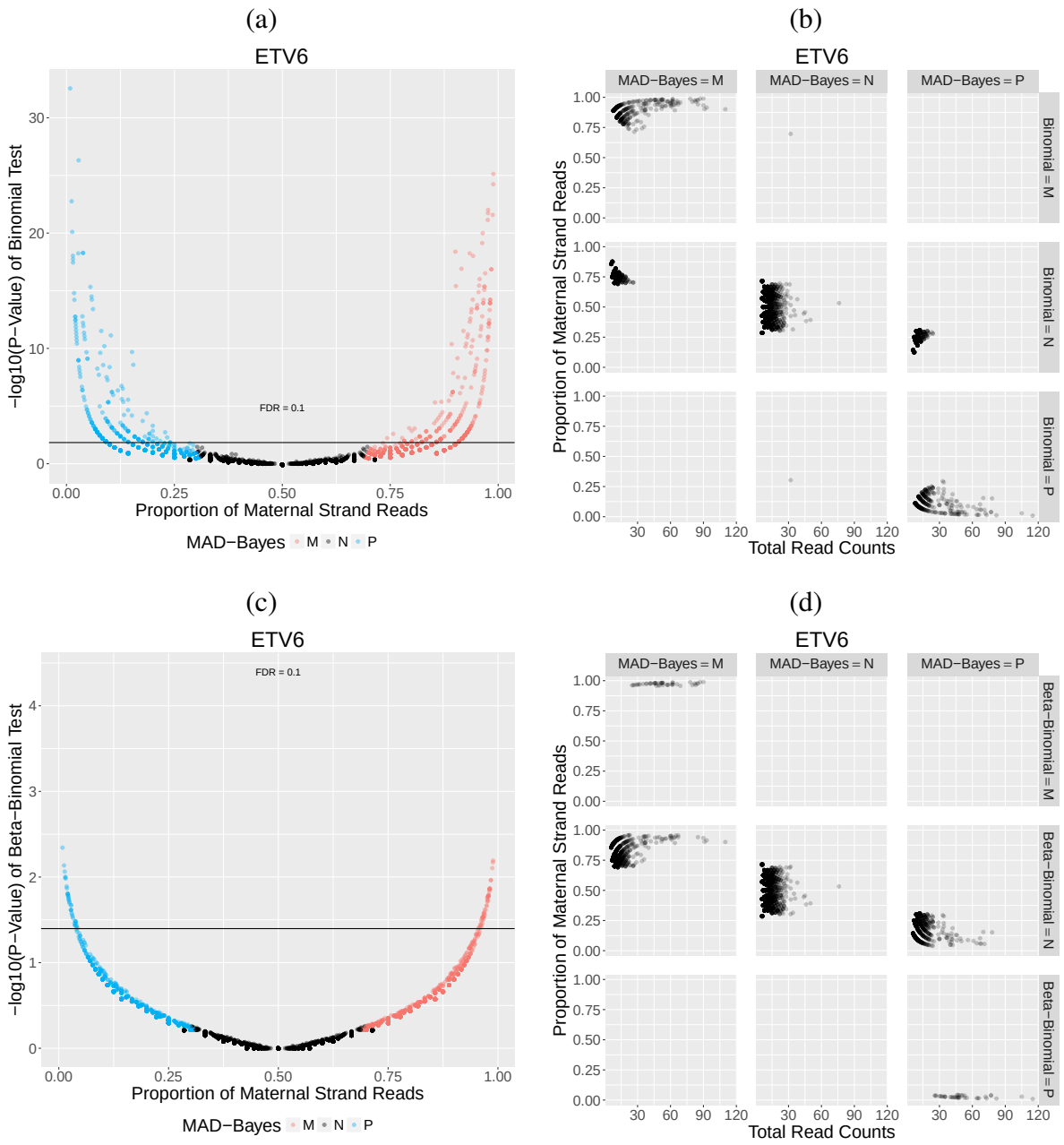
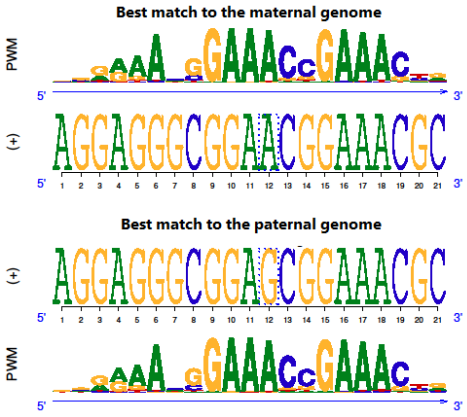
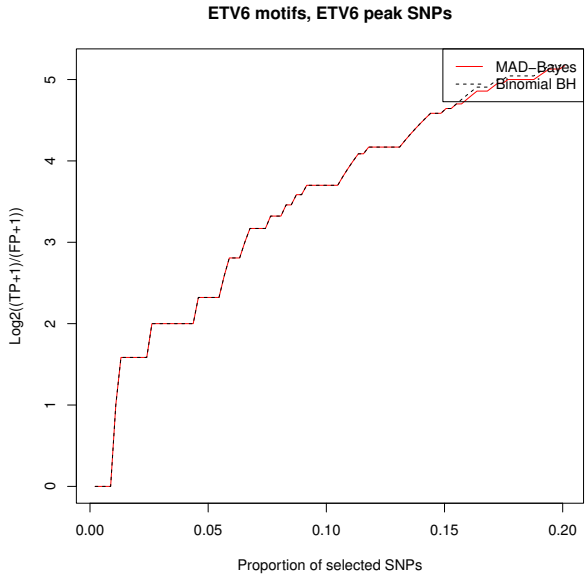


Figure 3.4: (a) and (b): Comparison between the results of Binomial test and MAD-Bayes for ETV6. (c) and (d): Comparison between the results of Beta-Binomial test and MAD-Bayes for ETV6.

(a)



(b)



(c)

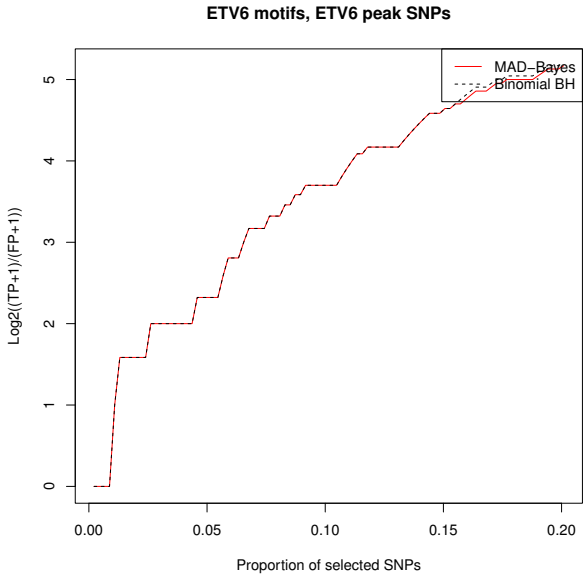


Figure 3.5: (a) Composite logo plot for IRF3 PWM (IRF3_2 from ENCODE) and SNP at chr1:149,899,885. This locus had maternal count proportion of with a total count of It is inferred as exhibiting maternal-specific IRF3 binding by MAD-Bayes MBASIC, as supported by the better match to the PWM with the maternal allele A compared to paternal allele G. In contrast, Binomial and Beta-Binomial inferred this locus as neutral. (b) Accuracy ($\log_2\left(\frac{TP+1}{FP+1}\right)$) for fixed proportion of SNPs in peaks after filtering: IRF3. (c) Accuracy ($\log_2\left(\frac{TP+1}{FP+1}\right)$) for fixed proportion of SNPs in peaks after filtering: ETV6.

We validated the allele-specific SNPs impacting ETV6 and IRF3 binding with atSNP by matching target SNP loci to corresponding position weight matrices of the two TFs at FDR level of 0.1 (No PWMs available for TRIM22 in JASPAR ([33]) or ENCODE ([50]).). A winning strand in {"M":Maternal strand; "P":Paternal strand} was defined as the strand with a better match for each SNP-motif pair . We then filtered: i) SNPs where the two motifs displayed opposite signal, i.e., both motifs labeled a SNP as significant but with different winning strands; and ii) SNPs where the read count data exhibited opposite signal of the atSNP winning strand, i.e., if the maternal count $> 75\%$ of total count, while motif match was in favor of paternal strand, and vice versa. In the first case, allele-specificity was ambiguous in terms of motif matching; while in the latter case, given the threshold of 25% and 75% on the proportions, methods based on read count proportions did not support allele-specific binding on the opposite strand. This process result in a set of SNPs, with allelic-specificity as a gold standard. We next compared the inferred states of MAD-Bayes and Binomial test to winning strands (Table 3.2). For SNPs within the peaks of the evaluated TF, we found that MAD-Bayes led to equal or more correct allele-specific inferences, and eliminated false positives.

Next we extended the definition of winning strand from SNPs inferred significant allele-specific by atSNP to all target SNPs within corresponding peaks and ranked them by their significance from atSNP. The SNPs underwent similar filtering for opposite motif signals and inconsistency between motif signals and read count proportions, resulting in 242 SNPs for IRF3 and 162 SNPs for ETV6. We labeled a SNP as a potential True Positive (TP) if the winning strand based on motif matching was the same as inferred state; a potential False Positive (FP) otherwise. Fig. 3.5(b) and (c) display as well or better accuracy of MAD-Bayes for both transcription factors. Both MAD-Bayes and Binomial test inferred most of SNPs as neutral in terms of transcription factor binding. Potential reasons include: a higher degree of matching between motif and DNA sequences not necessarily implying a higher probability of binding event due to in silico nature of atSNP and the relatively low total read counts at SNP loci introducing noise and attenuating signal strength.

Table 3.2: Comparison between atSNP winning strands and inferred states of MAD-Bayes MBASIC and Binomial test of FDR < 0.1, for SNP loci predicted to lead to allele-specific for corresponding TFs by atSNP (FDR < 0.1). Each cell represents number of SNPs loci with specific allelic pattern for (a) IRF3, and (b) ETV6.

(a):IRF3		Within Peaks			Not in Peaks		
		atSNP			atSNP		
		"M"	"P"	Total	"M"	"P"	Total
MAD-Bayes	"M"	**4**	0	4	**15**	12	27
	"N"	13	8	21	42	37	79
	"P"	0	**3**	3	16	**26**	42
	Total	17	11	28	73	75	148
Binomial	"M"	**1**	0	1	**2**	0	2
	"N"	16	11	27	71	75	146
	"P"	0	**0**	0	0	**0**	0
	Total	17	11	28	73	75	148

(b):ETV6		Within Peaks			Not in Peaks		
		atSNP			atSNP		
		"M"	"P"	Total	"M"	"P"	Total
MAD-Bayes	"M"	**5**	0	5	**57**	1	58
	"N"	1	3	4	115	9	205
	"P"	0	**3**	3	7	**44**	51
	Total	6	6	12	179	135	314
Binomial	"M"	**5**	0	5	**46**	1	47
	"N"	1	3	4	132	113	245
	"P"	0	**3**	3	1	**21**	22
	Total	6	6	12	179	135	314

3.4.2 Allele-specific Histone Modification Analysis

We processed the SNP set by (1) keeping SNP loci that resided in the union of peak set of the three marks; (2) subsetting loci where at least one of the 6 datasets (2 replicate$\times$3 histone marks) had enough read depth (total read count of maternal and paternal strand $\geq$ 8); and (3) removing loci where the maternal count proportion of the two replicates from same modification show inconsistent allelic preference. (In defining this inconsistency, we considered loci with read depth $\geq$ 20 and used the 80% quantile of absolute difference of maternal proportion between 2 replicates as our threshold to filter all SNP loci.) this resulted in 23,836 SNP loci as our input set of MAD-Bayes MBASIC.

MAD-Bayes MBASIC assigned the allele-specificity at each loci to the state-space: $\{$"M":Maternal-specific, "N": Neutral, "P": Paternal-specific, "-": not within peak. $\}$. For the H3K4me1 modification, 21.9% SNP loci were inferred as significant, i.e. "M" state or "P" state; 21.5% for H3K4me3, and 21.4% for H3K27ac. Proportion of strand specific reads at all of the significant loci were consistent with the inferred states. (Fig. 3.6(a)) In Binomial setting, significant contrast did not necessarily rule out modification events on the minor strand, i.e. the strand with fewer reads for the locus. However, allele-specific SNP loci generally exhibited relatively low count on the minor strand, indicating that histone modifications probably happened only on one of the strands. (Fig. 3.6(b))

We compared the inference result with the Binomial and Beta-Binomial test at FDR$=0.1$ (Fig. 3.7, Fig. 3.8, Fig. B.4, Fig. B.5, Fig. 3.7, and Fig. **??**). Con MAD-Bayes MBASIC performed far more sensitive than the Beta Binomial test: almost all the significant detections from Beta-Binomial test could be recovered. Moreover, inference by MAD-Bayes MBASIC were not only based on the dataset itself, but also on information borrowed from all of the datasets. For example, SNP loci inferred as "M" for H3K4me3 only by MAD-Bayes were very likely the ones with high maternal proportions for the other two histone marks (The left panel on second row in Fig. 3.7 and Fig. 3.8). MAD-Bayes MBASIC actually labeled several times the number of significant SNP loci as maternal-specific or paternal-specific (Table 3.3). The extra detections from MAD-Bayes

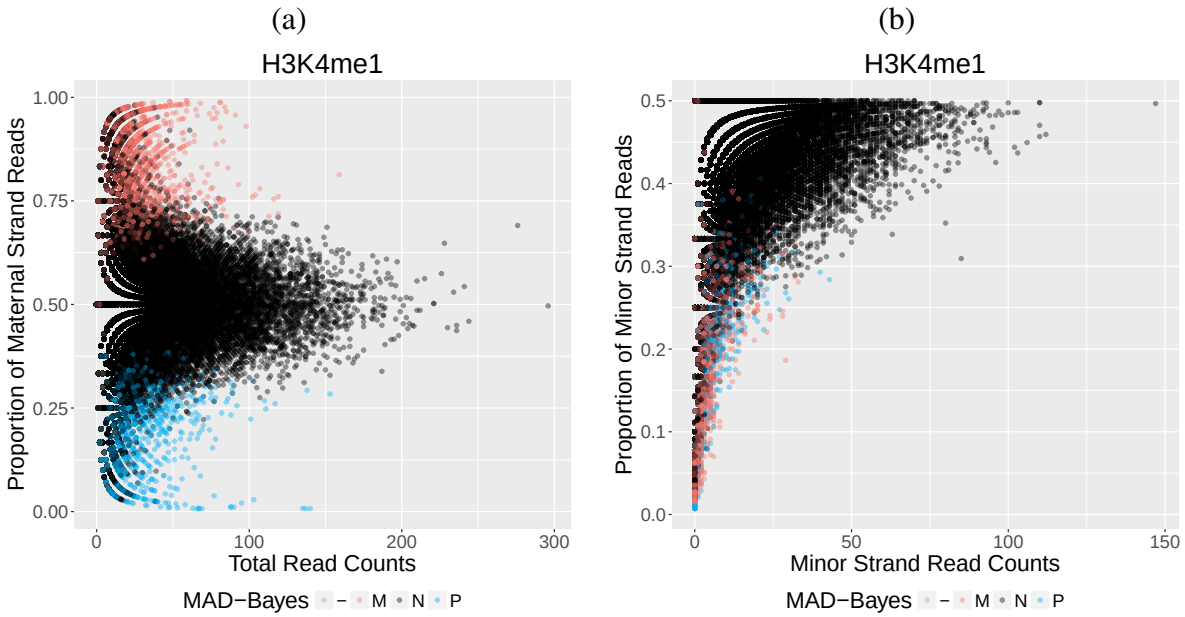


Figure 3.6: Strand proportions and read counts of different inferred states for H3K4me3, with 2 replicates together in one plot. (a) Maternal strand proportion v.s total read counts. (b) Minor strand proportion v.s minor strand counts.

were supported by the imbalance between maternal and paternal strand read count proportions and ranked relatively high in terms of p-values from Beta-Binomial test.

While K histone modifications could generate 4^K configurations: {"M","N","P", "-"}^{K} in theory, for the three active marks involved, it is reasonable to expect that allele specificity of different modifications would agree with each other. This expectation was supported by the data as most of the SNP loci had configurations where parental allele preference of different histone modifications did not contradict with each other. Only for 0.64% of the loci, an "M" state and a "P" state were assigned to the same locus for different modifications. (e.g. "PMM" combination at SNP locus Chr7:1,979,741, where H3K4me1 was inferred to be specific to paternal allele, while H3K4me3 and H3K27ac monotonously displayed preference for maternal allele.)

MAD-Bayes MBASIC tends to aggregate combinations together during clustering. It clustered the 23,836 loci into into 4 groups each dominated by a few combinations around 80% of each cluster. In Table 3.4, the first cluster (named as Cluster N) was dominated by the combinations of states related to neutral modifications on the two strands; the majority of SNP loci in the second cluster (Cluster P) were inferred as paternal-specific sites for the three histone marks; the third cluster (Cluster M1) contained SNP loci where modifications were in favor of maternal strand; the last cluster (Cluster M2) seemed quite interesting, with combinations of maternal specific H3K4me1 and neutral or non-peak for H3K4me3 and H3K27ac. Since H3K4me3 and H3K27ac were both predominantly promoter marks and H3K4me1 is an enhancer mark, the separation of Clusters M1 and M2 might be a result of functional difference between the two subgroups.

We next examined the allele-specific states of SNPs for each histone modification in the span of a peak. Peak widths of the three histone modifications range from several hundred to a few thousand. Some of the SNP loci were filtered by our criteria previously to maintain high quality at each locus; therefore 27.0%(3,778) of H3K4me1 peaks, 32.2%(2,847) of H3K4me3 peaks and 32.3%(3,453) of H3K27ac peaks had multiple SNP loci with qualified data counts. In general, allelic-specificities of SNP loci within the same peak were consistent, exceptions to which were due to allele-specific loci having a neutral neighbor, where difference between paternal and maternal strand read counts were not large enough to lead to significance, or the degree of allele-specificity

Table 3.3: Comparison of significant allele-specific histone modifications detected by MAD-Bayes MBASIC, Binomial and Beta-Binomial tests of FDR < 0.1. Each cell represents number of SNP loci with specific allelic patterns for (a) H3K4me1, (b) H3K4me3, and (c) H3K27ac.

(a) H3K4me1

	MAD-Bayes					MAD-Bayes			
Binomial	"M"	"N"	"P"	Total	Beta-Binomial	"M"	"N"	"P"	Total
"M"	**1,319**	260	0	1,579	"M"	**1,136**	0	0	1,136
"N"	2,289	**16,669**	732	19,690	"N"	2,472	**17,166**	903	20,541
"P"	0	237	**482**	873	"P"	0	0	**311**	311
Total	3,608	17,166	1,214	21,988	Total	3,608	17,166	1,214	21,988

(b) H3K4me3

	MAD-Bayes					MAD-Bayes			
Binomial	"M"	"N"	"P"	Total	Beta-Binomial	"M"	"N"	"P"	Total
"M"	**728**	420	0	1,148	"M"	**698**	11	0	709
"N"	989	**10,833**	886	12,708	"N"	1,019	**11,524**	1,054	13,597
"P"	0	282	**558**	840	"P"	0	0	**390**	390
Total	1,717	11,535	1,444	14696	Total	1,717	11,535	1,444	14696

(c) H3K27ac

	MAD-Bayes					MAD-Bayes			
Binomial	"M"	"N"	"P"	Total	Beta-Binomial	"M"	"N"	"P"	Total
"M"	**1,233**	873	0	2,106	"M"	**1,089**	3	0	1,092
"N"	1,235	**13,014**	663	14,912	"N"	1,379	**14,646**	964	16,989
"P"	0	761	**872**	1,633	"P"	0	0	**570**	570
Total	2,468	14,649	1,534	18,651	Total	2,468	14,649	1,534	18,651

might be gradually changing along genome (Fig. 3.9). We also observed a few interesting examples where loci within the same peak showed different ASHM. In some cases, the change of strand preference coincided a shift of the peak summit in the maternal and paternal coverage plot of the peaks. (Fig. B.8)

Evaluation with allele-specific expression

We next explored the association between allele-specific histone modification and allele specific expression under two settings: 1) for SNP loci in exons: ASE and ASHM at the same immediate SNP locus and 2) for SNP loci in regulatory regions: ASE at gene-level and ASHM at SNP loci in corresponding promoter or enhancer regions (to be defined in the following). We used the ASE data from AlleleDB ([13]), with the similar inferred state space of {"M":Expression specific to Maternal allele. "N": Expression Neutral to both alleles, "P": Expression specific to Paternal allele. } at FDR 0.1.

In the first setting, the distribution of ASE states of loci in four MAD-Bayes clusters agreed with the dominant combinations in MAD-Bayes clusters. In Fig. 3.10 (a), similar to the overall trend, SNP loci with ASHM states in Cluster N showed low representation of "M" and "P" in the ASE states. Cluster P exhibited an enrichment in the "P" state, agreeing with the several dominant combinations of paternal-specific histone modification states. In Cluster M1, enrichment in the "M" state was supported by the largely maternal-specific states. Interestingly, Cluster M2 included a large number of SNP loci inferred as maternal-specific for H3K4me1, (with "M–" or "MNN" patterns), and behaved similarly to the Cluster N in terms of ASE states distribution. Since H3K4me1 is a histone mark enriched in enhancers and other distal regulatory elements, it was reasonable to expect that the mere allele-specificity of H3K4me1 taking place at SNP loci in genebody actually take effect not on the expression of that gene.

To understand the association between ASHM and ASE at distal locations, we first focused on the ASHM of SNP loci in promoters, defined as the upstream 2,000 bps of the transcription starting site (TS). We selected 5,821 genes with at least one allele-specific expressed SNP, aggregated their

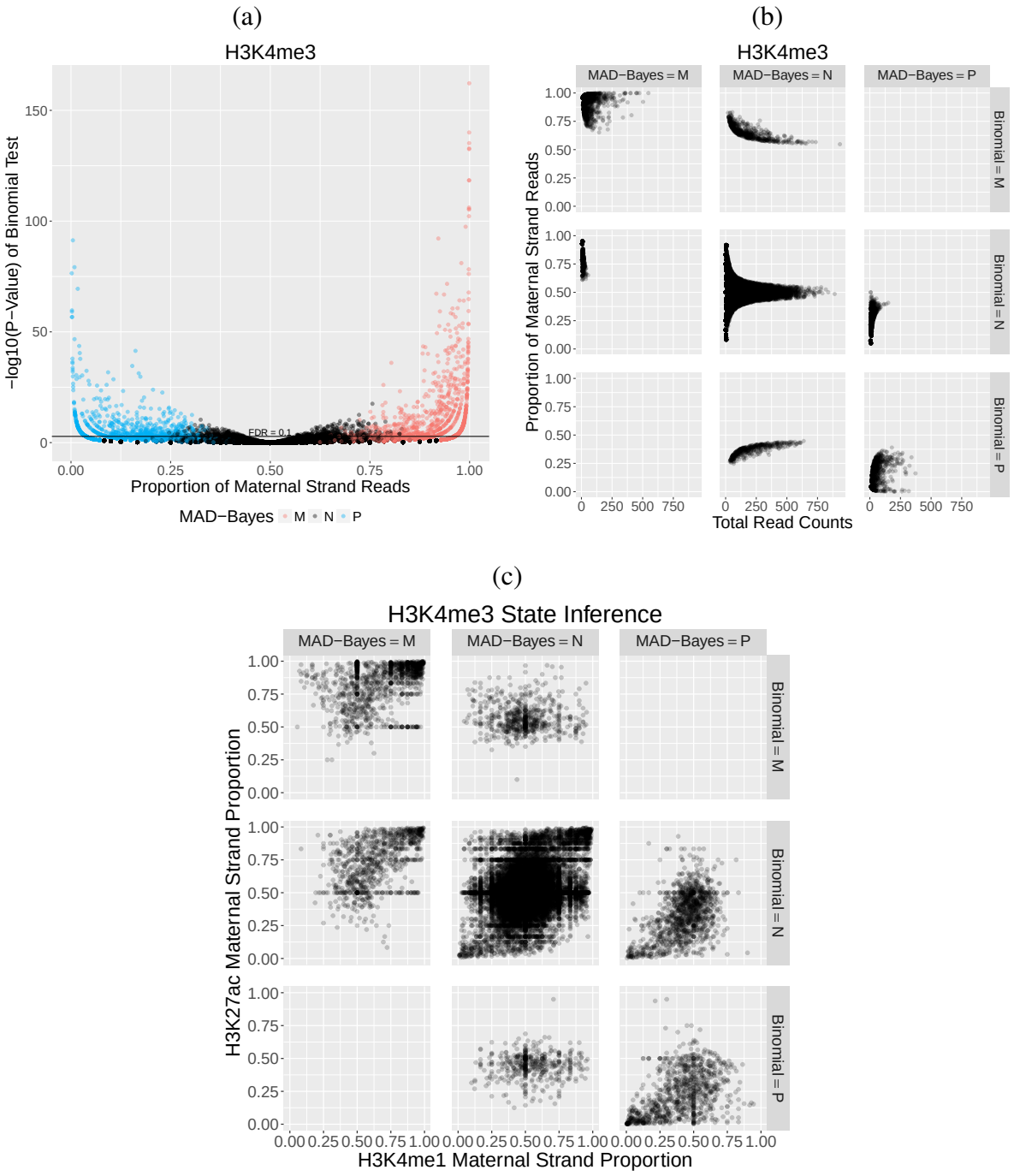


Figure 3.7: Comparison between the results of Binomial test and MAD-Bayes for H3K4me3. (a): Binomial test p-values versus maternal strand count proportions. (b): Maternal strand count proportions versus total counts. (c): Maternal strand counts of H3K4me1 and H3K27ac under different inferred states of H3K4me3.

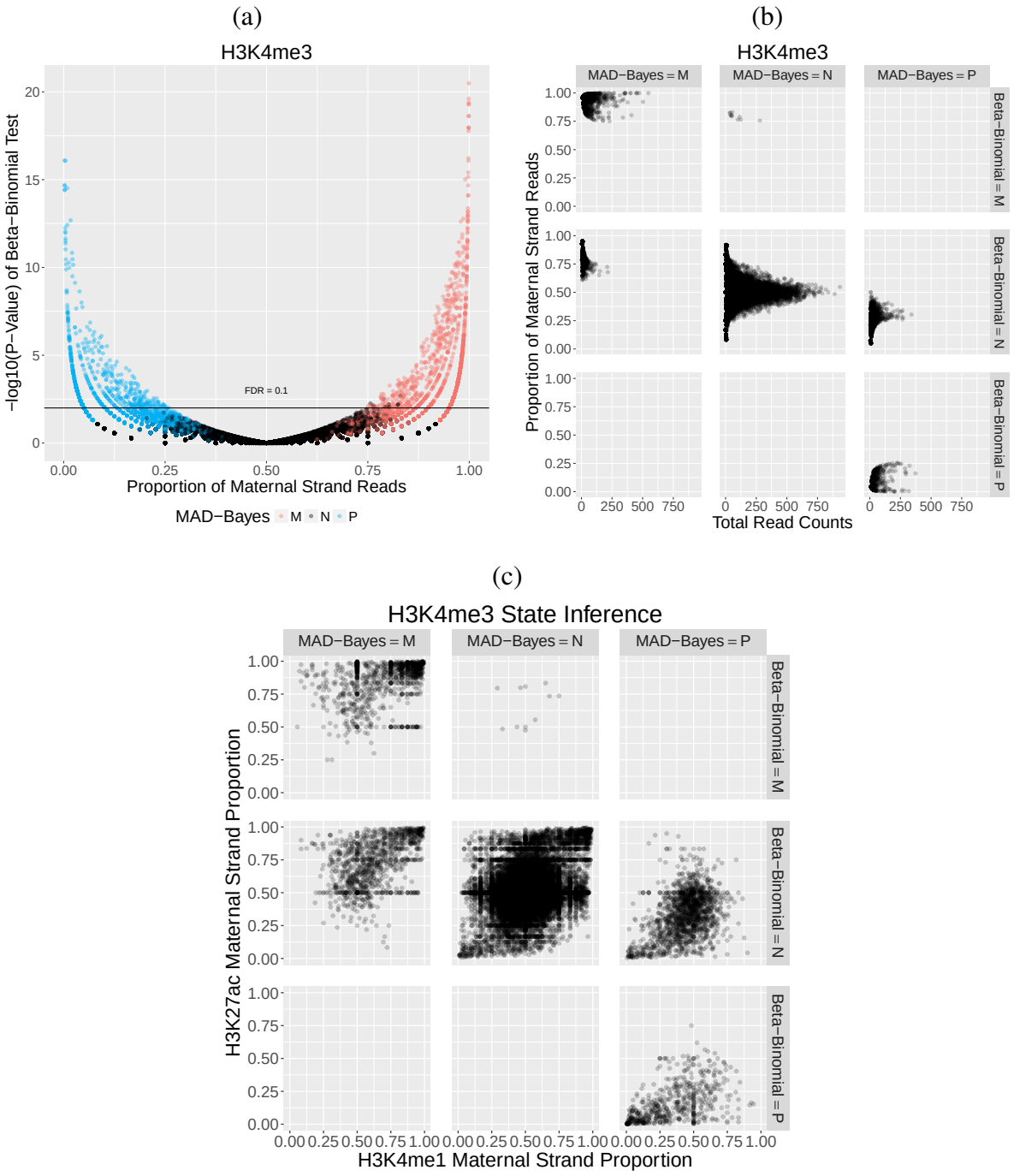


Figure 3.8: Comparison between the results of Beta-Binomial test and MAD-Bayes for H3K4me3. (a): Binomial test p-values versus maternal strand count proportions. (b): Maternal strand count proportions versus total counts. (c): Maternal strand counts of H3K4me1 and H3K27ac under different inferred states of H3K4me3.

Table 3.4: MAD-Bayes MBASIC clustered SNP loci by grouping similar combinations. A few dominant combinations covered the majority of each cluster.

Cluster	No. Loci	Dominant Combinations				Coverage
N	17,313	"NNN" (8,492)	"N- -" (2,922)	"N-N" (2,611)	"-NN"(1,029)	86.95%
P	1,803	"PPP" (494)	"NPP" (452)	"NPN" (334)	"P-P" (129)	78.15%
M1	2,467	"MMM" (833)	"M-M" (588)	"M- -" (328)	"NMM" (249)	80.99%
M2	2,253	"M- -" (839)	"-NN" (460)	"M-N" (347)	"MNN" (479)	94.14%

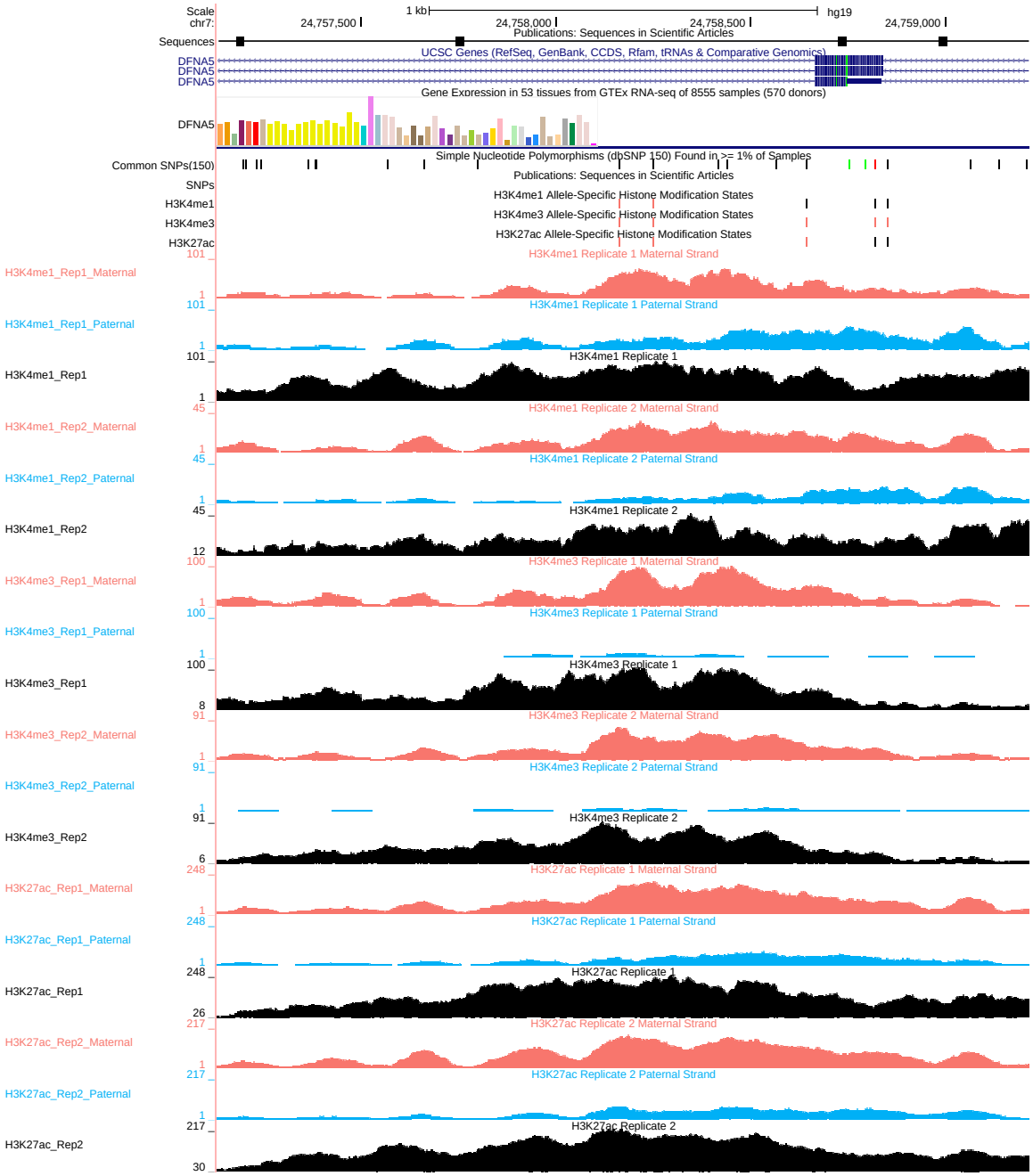


Figure 3.9: H3K4me3 displayed a consistent preference for mar maternal strand in the peak Chr7:24,757,161-24,759,248, with a string of allele specific states as: “M, M, M, M”. While for both H3K4me1 and H3K27ac, the difference between maternal and paternal strand counts of SNP loci at downstream were not sharp enough, and they we inferred as neutral, the peak in general were consistently in favor of maternal strand.

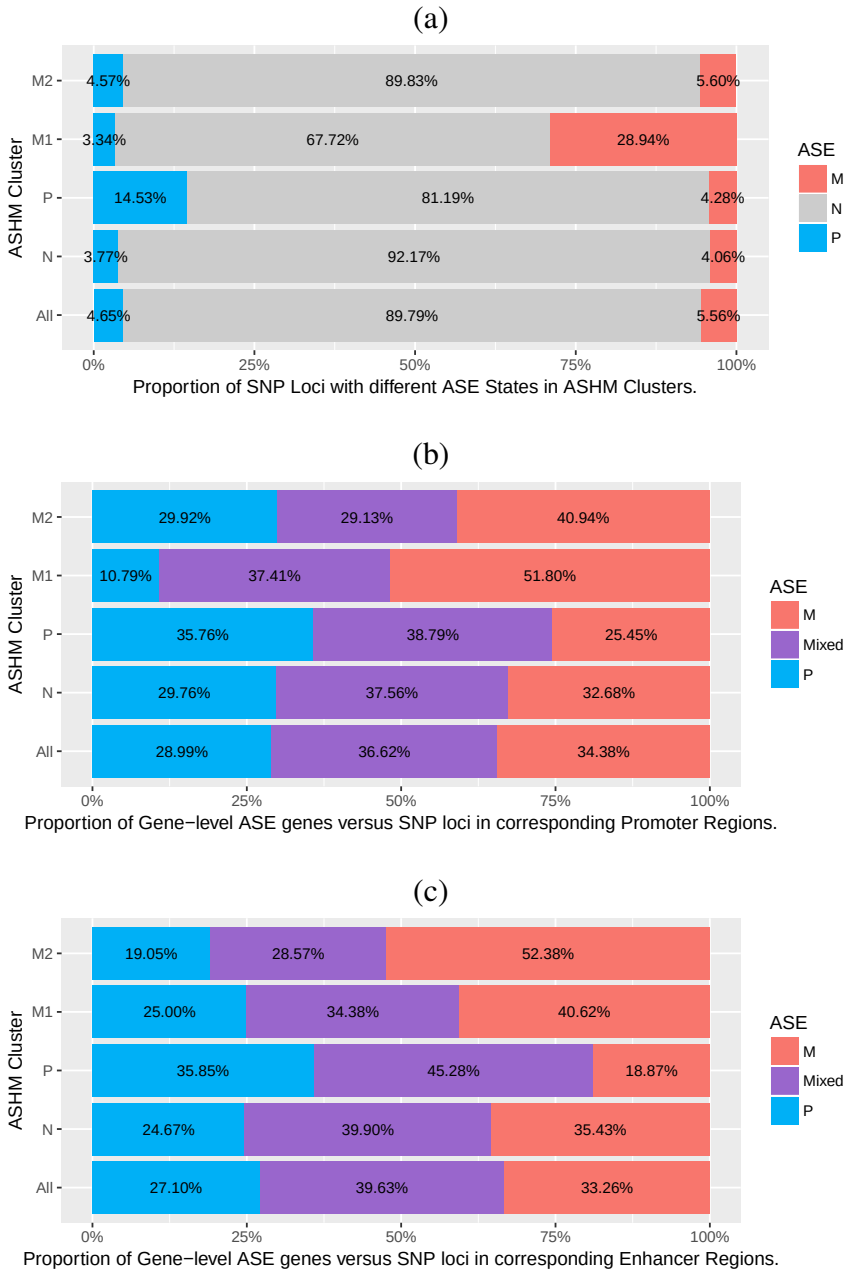


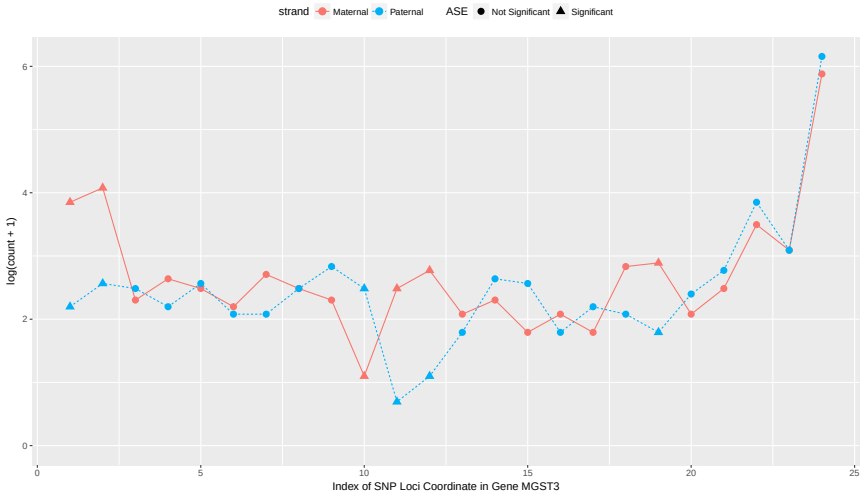
Figure 3.10: (a) Association between ASE and ASHM at the same SNP locus. (b) Association between aggregated gene-level ASE and ASHM of SNP loci in corresponding promoter regions. (c) Association between aggregated gene-level ASE and ASHM of SNP loci in corresponding enhancer regions.

gene-level ASE results as “M”, “P”, and “Mixed” (genes with both maternal-specific and paternal-specific expressed SNPs) to study each SNP-gene pair as long as SNP loci resided in the corresponding promoter. Overlapping the promoter regions with our candidate SNP loci, we built 2,097 SNP-gene pairs by connecting 2,006 SNP loci and 1,303 genes. We observed far more genes with both maternal-specific expression and paternal-specific SNP loci than genes where expression was consistently in favor of one strand. This observation can be expressed by the complication of ASE with alternative splicing, i.e., some of the transcripts were expressed maternally while expression of other transcripts were in favor of the paternal allele. For example, MGST3 was a “Mixed” gene, whose transcript ENST00000367889.3 was maternally expressed while paternal-specific expressed SNPs indicated some paternal-specific transcripts(Fig 3.11). (While it is more preferrably to infer transcript-level ASE to further interpret these results, the existing computational approaches toward transcript-level expression largely emphasize on the assignment of RNA-seq reads to different transcripts ([52, 37]), which creates another layer of uncertainty to the maternal and paternal strand counts; hence we sticked to the AlleleDB, which focuses on SNP-level ASE.)

In Fig. 3.10(b), association of gene-level ASE results and MAD-Bayes clusters of SNP in corresponding promoters supported the MAD-Bayes results, i.e., Cluster N as neutral, Cluster P as paternal-specific, and Cluster M1 as maternal-specific, where the preferred parental strand of ASHM in promoter SNP loci actually saw an enrichment in the genes demonstrating specificity for the same strand, and a clear depletion in genes expressed specific to the minor strand. Cluster M2 followed the neutral behavior of the overall trend of Cluster N, while including slightly more maternal-specific ASE genes than the paternal-specific ones. Association between ASE and ASHM of SNP loci in promoter region displayed the potential mediation role of histone modifications in regulatory regions.

Furthermore, we picked out 1,596 enhancer regions with the definition: (1) regions referred to as “enhancer” ([50]); and (2) regions which HiC ([4]) inferred as to encompass a significant long range interaction with the promoter regions we defined for GM12878 cell line. Similarly, we further restricted to genes whose enhancer regions contain SNP loci with ASHM information, and got down to 487 SNP-gene pairs with 412 SNP loci and 282 genes. While ASE trend of SNP loci

(a)



(b)

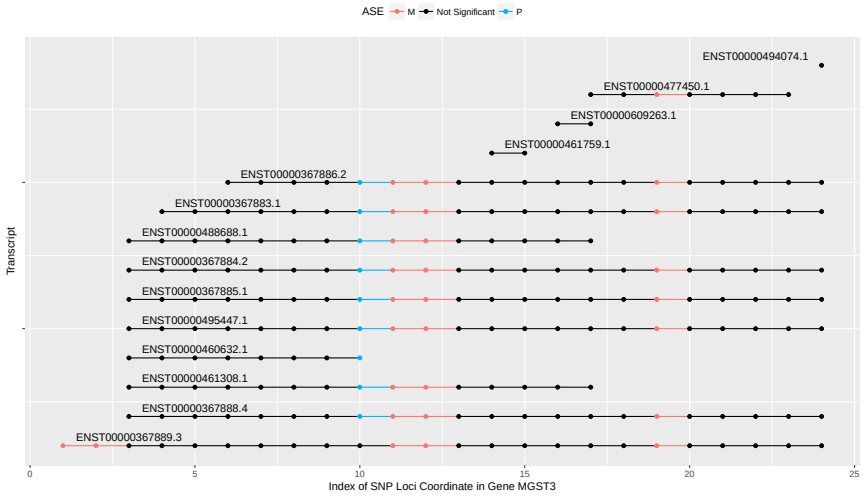


Figure 3.11: An example of a "Mixed" gene, MGST3, with both maternal-specifically and paternal-specifically expressed SNPs. (a) Log counts of the two strands and inferred ASE states of SNP loci indexed by coordinate. (b) Coverage of transcripts on SNP loci. Only ENST00000367889.3, with two maternal-specific SNP loci could be inferred as maternal-specific.

in Cluster N, Cluster P and Cluster M1 seemed similar to the ones of SNP-gene association from promoter region, expression of genes with SNP loci assigned to Cluster M2 was significantly in favor of the maternal strand (Fig. 3.10(c)). It implied that SNP loci in distal regulatory regions could also impact gene expression through histone modifications.

3.5 Conclusions and Discussion

In this chapter, we tried to bridge the gap between genomic variants and gene regulation through mediators such as histone modifications or transcription factor binding. By applying the extended MAD-Bayes MBASIC with variance-stabilizing transformation to the Binomial models of allele-specific inference of ChIP-seq data, we located SNP loci leading to allelic imbalance in ChIP-seq data as potential sites, whose regulatory path take histone modification or transcription factor binding events as an important chain.

MAD-Bayes MBASIC is a fast method for querying large sets (1000s) of ChIP-seq data with user-specified loci of interest. This represents a fast realization of the MBASIC framework on various models. We extended it to allele-specific analysis for transcription factor binding and histone modification problem settings via a variance-stabilizing transformation. Our evaluations indicated that such transformation preserved the time-efficiency and relatively high accuracy and low inference error of MAD-Bayes MBASIC.

We carried out ASB applications to transcription factor datasets from the GM12878 cell line. Comparison with Binomial and Beta-Binomial tests displayed the high power of MAD-Bayes MBASIC, and validation by atSNP, a method based on sequence-specificity, showed MAD-Bayes MBASIC was not making more mistakes than Binomial test. We also conducted the first ASHM application, revealed the consistency of SNP loci within a peak and the consistency between active histone marks, and a few interesting exceptions. By associating ASE results of genes and ASHM results in corresponding promoter and enhancer regions, we detected candidate genetic variants for further analysis, whose potential regulatory mechanism uses histone modifications as mediator in one of the chains.

Chapter 4

Conclusions

In this thesis, 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 ([29]) leaving the reconciliation of peaks over multiple replicates to the user.

We carried out ASB applications to transcription factor datasets from the GM12878 cell line. Comparison with Binomial and Beta-Binomial tests displayed the high power of MAD-Bayes MBASIC, and validation by atSNP, a method based on sequence-specificity, showed MAD-Bayes MBASIC was not making more mistakes than Binomial test. We also conducted the first ASHM application, revealed the consistency of SNP loci within a peak and the consistency between active histone marks, and a few interesting exceptions. By associating ASE results of genes and ASHM results in corresponding promoter and enhancer regions, we detected candidate genetic variants for further analysis, whose potential regulatory mechanism uses histone modifications as mediator in one of the chains.

In Chapter 2, I introduces the MAD-Bayes MBASIC for simultaneous binding state inference under strong assumptions. Employing a small variance asymptotics, the MAD-Bayes MBASIC improved significantly in computational time without impacting accuracy. The problem of selecting number of clusters is turned into the problem of selecting tuning parameters, which is addressed by a heuristic tuning method. Initialization is divided into two stage: state initialization and clustering initialization.

In Chapter 3, I extended the MAD-Bayes MBASIC to the allele-specifc analysis setting, via a variance stabilizing transformation. Computational experiments and applications to the allele-specific analysis of transcription factor binding and histone modification displays the power and accuracy of the this method compared to individual dataset level approach. It also provides the first systematic analysis of allele-specific histone modifications. Associating allele-specific histone modification results to allele-specific expression data, a small handful of SNP loci is picked out as the potential candidate for further studies in the causal regulatory mechanisms.

The R package *MBASIC* is available at http://github.com/KaileiChen/mbasic in latest version, and will be submitted to bioconductor soon.

Bibliography

[1] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[2] David J Aldous. Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII 1983*, pages 1–198, 1983.

[3] Francis J Anscombe. The transformation of poisson, binomial and negative-binomial data. *Biometrika*, 35(3/4):246–254, 1948.

[4] Ferhat Ay, Timothy L Bailey, and William Stafford Noble. Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome research*, 24(6):999–1011, 2014.

[5] Pierre Baldi and Anthony D Long. A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, 2001.

[6] A. Banerjee. Clustering with Bregman Divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[7] Y. Bao, V. Vinciotti, E. Wit, and P. AC't Hoen. Accounting for immunoprecipitation efficiencies in the statistical analysis of ChIP-seq data. *BMC bioinformatics*, 14(1):169, 2013.

[8] Y. Bao, V. Vinciotti, E. Wit, and P. 't Hoen. Joint modeling of ChIP-seq data via a Markov random field model. *Biostatistics*, 15(2):296–310, 2014.

[9] A. F. Bardet, Q. He, J. Zeitlinger, and A. Stark. A computational pipeline for comparative chip-seq analyses. *Nature protocols*, 7(1):45–61, 2012.

[10] D. Blackwell and J. B. MacQueen. Ferguson Distributions via Polya Urn Schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[11] T. Broderick, B. Kulis, , and MI. Jordan. MAD-Bayes: MAP-based asymptotic derivations from Bayes. Proceedings of the 30th International Conference on Machine Learning. In *Proceedings of the 30th International Conference on Machine Learning.*, 2013.

[12] Lawrence Brown, Tony Cai, Ren Zhang, Linda Zhao, and Harrison Zhou. The root–unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability theory and related fields*, 146(3-4):401–433, 2010.

[13] Jieming Chen, Joel Rozowsky, Timur R Galeev, Arif Harmanci, Robert Kitchen, Jason Bedford, Alexej Abyzov, Yong Kong, Lynne Regan, and Mark Gerstein. A uniform survey of allele-specific binding and expression over 1000-genomes-project individuals. *Nature communications*, 7, 2016.

[14] K. B. Chen, R. Hardison, and Y. Zhang. dCaP: detecting differential binding events in multiple conditions and proteins. *BMC Genomics*, 15(9):1–14, 2014.

[15] Ines de Santiago, Wei Liu, Ke Yuan, Martin OReilly, Chandra Sekhar Reddy Chilamakuri, Bruce AJ Ponder, Kerstin B Meyer, and Florian Markowetz. Baalchip: Bayesian analysis of allele-specific transcription factor binding in cancer genomes. *Genome biology*, 18(1):39, 2017.

[16] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc B Met*, 39:1–38, 1977.

[17] J. Ernst and M. Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–25, 2010.

[18] Kyle Kai-How Farh, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J Housley, Samantha Beik, Noam Shoresh, Holly Whitton, Russell JH Ryan, Alexander A Shishkin,

et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337, 2015.

[19] J. P. Ferguson, J. H. Cho, and H. Zhao. A new approach for the joint analysis of multiple ChIP-seq libraries with application to histone modification. *Statistical applications in genetics and molecular biology*, 11(3):Article 1, 2012.

[20] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, et al. Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100, 2012.

[21] K. J. Hewitt, D. H. Kim, P. Devadas, R. Prathibha, C. Zuo, R. Sanalkumar, K. D. Johnson, Y-A. Kang, J-S. Kim, C. N. Dewey, S. Keleş, and E. Bresnick. Hematopoietic signaling mechanism revealed from a stem/progenitor cell cistrome. *Molecular cell*, 59(1):62–74, 2015.

[22] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9:473–476, 2012.

[23] H. Ji, X. Li, Q-F. Wang, and Y. Ning. Differential principal component analysis of ChIP-seq. *Proceedings of the National Academy of Sciences of the United States of America*, 110(17):6789–6794, 2013.

[24] K. D. Johnson, A. Hsu, M-J R., M. E. Boyer, S. Keleş, J. Zhang, Y. Lee, S. M. Holland, and E. H. Bresnick. Cis-element mutation in a GATA-2-dependent immunodeficiency syndrome governs hematopoiesis and vascular integrity. *Journal of Clinical Investigation*, 10(122):36923704, 2012.

[25] Maya Kasowski, Sofia Kyriazopoulou-Panagiotopoulou, Fabian Grubert, Judith B Zaugg, Anshul Kundaje, Yuling Liu, Alan P Boyle, Qiangfeng Cliff Zhang, Fouad Zakharia,

Damek V Spacek, et al. Extensive variation in chromatin states across humans. *Science*, 342(6159):750–752, 2013.

[26] Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

[27] Felix Krueger and Simon R Andrews. Snpsplit: Allele-specific splitting of alignments between genomes with known snp genotypes. *F1000Research*, 5, 2016.

[28] P. F. Kuan, D. Chung, G. Pan, J. Thomson, R. Stewart, and S. Keleş. A statistical framework for the analysis of ChIP-Seq data. *Journal of the American Statistical Association*, 106:891–903, 2011.

[29] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, et al. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome research*, 22(9):1813–1831, 2012.

[30] K. Liang and S. Keleş. Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics*, 28(1):121–122, 2012.

[31] Michael Lynch, Bruce Walsh, et al. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998.

[32] S. Mahony, M. D. Edwards, E. O. Mazzoni, R. I. Sherwood, A. Kakumanu, C. A. Morrison, H. Wichterle, and D. K. Gifford. An integrated model of multiple-condition ChIP-Seq data reveals predeterminants of Cdx2 binding. *PLoS computational biology*, 10(3):e1003501, 2014.

[33] A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 42(D1):D142–D147, 2014.

[34] Graham McVicker, Bryce van de Geijn, Jacob F Degner, Carolyn E Cain, Nicholas E Banovich, Anil Raj, Noah Lewellen, Marsha Myrthil, Yoav Gilad, and Jonathan K Pritchard. Identification of genetic variants that affect histone modifications in human cells. *Science*, 342(6159):747–749, 2013.

[35] Athma A Pai, Jonathan K Pritchard, and Yoav Gilad. The genetic and mechanistic basis for variation in gene regulation. *PLoS genetics*, 11(1):e1004857, 2015.

[36] Tomi Pastinen. Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics*, 11(8):533, 2010.

[37] Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek, and Steven L Salzberg. Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown. *Nature protocols*, 11(9):1650, 2016.

[38] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, 2010.

[39] W. M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[40] Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[41] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[42] Joel Rozowsky, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, Robert Bjornson, Yong Kong, Naoki Kitabayashi, et al. Alleleseq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology*, 7(1):522, 2011.

[43] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):1–25, 2004.

[44] K-A. Sohn, J. W. K. Ho, D. Djordjevic, H-H. Jeong, P. J. Park, and J. H. Kim. hiHMM: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics*, pages btv117–, 2015.

[45] J. Song and K. C. Chen. Spectacle: fast chromatin state annotation using spectral learning. *Genome Biology*, 16(1):33, 2015.

[46] Q. Song and A. D. Smith. Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics*, 27:870–1, 2011.

[47] P-N. Tan, M. Steinbach, and V. Kumar. Chap 8 : Cluster Analysis: Basic Concepts and Algorithms. *Introduction to Data Mining*, page Chapter 8, 2005.

[48] C. Taslim, T. Huang, and S. Lin. DIME: R-package for identifying differential ChIP-seq based on an ensemble of mixture models. *Bioinformatics*, 27(11):1569–70, 2011.

[49] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 09 2012.

[50] The ENCODE Project Consortium. Encode portal. `https://www.encodeproject.org/`, 2014.

[51] Dechao Tian, Quanquan Gu, and Jian Ma. Identifying gene regulatory network rewiring using latent differential graphical models. *Nucleic Acids Research*, 44(17):e140–e140, 2016.

[52] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562, 2012.

[53] Bryce Van De Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods*, 12(11):1061, 2015.

[54] Jie Wang, Jiali Zhuang, Sowmya Iyer, XinYing Lin, Troy W. Whitfield, Melissa C. Greven, Brian G. Pierce, Xianjun Dong, Anshul Kundaje, Yong Cheng, Oliver J. Rando, Ewan Birney, Richard M. Myers, William S. Noble, Michael Snyder, and Zhiping Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, 22:1798–1812, 2012.

[55] Y. Wei, T. Tenzen, and H. Ji. Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics*, 16(1):31–46, 2015.

[56] Yingying Wei, Xia Li, Qian-fei Wang, and Hongkai Ji. iaseq: integrative analysis of allele-specificity of protein-dna interactions in multiple chip-seq datasets. *BMC genomics*, 13(1):681, 2012.

[57] Jian Yan, Martin Enge, Thomas Whitington, Kashyap Dave, Jianping Liu, Inderpreet Sur, Bernhard Schmierer, Arttu Jolma, Teemu Kivioja, Minna Taipale, et al. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell*, 154(4):801–813, 2013.

[58] X. Zeng, R.Sanalkumar, E. H. Bresnick, H. Li, Q. Chang, and S. Keleş. jMOSAiCS: Joint analysis of multiple ChIP-seq datasets. *Genome Biology*, 14:R38, 2013.

[59] C. Zuo and S. Keleş. A statistical framework for power calculations in ChIP-seq experiments. *Bioinformatics*, 30(6):853–860, 2014.

[60] Chandler Zuo, Kailei Chen, Kyle J Hewitt, Emery H Bresnick, and Sündüz Keleş. A hierarchical framework for state-space matrix inference and clustering. *The annals of applied statistics*, 10(3):1348, 2016.

[61] Chandler Zuo, Kailei Chen, and Sündüz Keleş. A mad-bayes algorithm for state-space inference and clustering with application to querying large collections of chip-seq data sets. *Journal of Computational Biology*, 24(6):472–485, 2017.

[62] Chandler Zuo, Sunyoung Shin, and Sündüz Keleş. atsnp: transcription factor binding affinity testing for regulatory snp detection. *Bioinformatics*, 31(20):3353–3355, 2015.

Appendix A: Supplementary Figures for Chapter 2

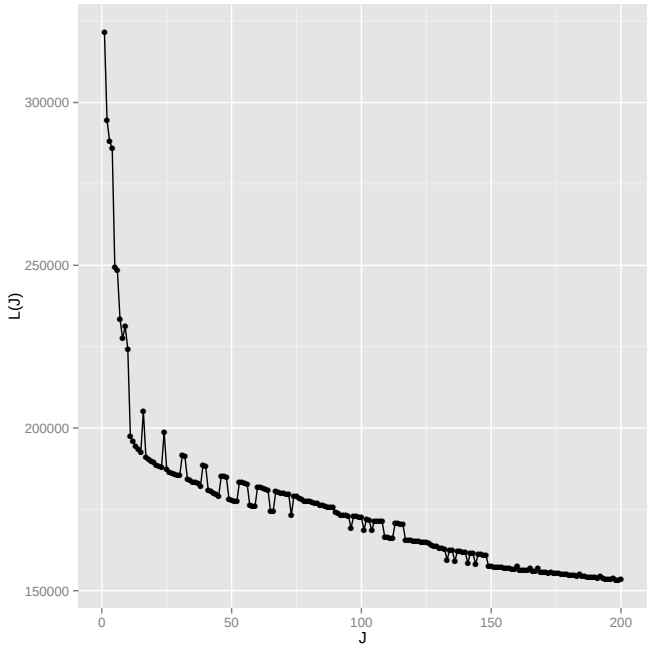


Figure A.1: A graphical interpretation of the conjugacy between λ_r and J. We use the K-means initialization to compute surrogate values for $L(J)$ for a large collection of $J \geq 1$. The λ_r value that can yield J clusters in the global solution must satisfy:
$\sup_{J'>J} \frac{L(J)-L(J')}{J-J'} \leq \lambda_r \leq \inf_{J'>J} \frac{L(J')-L(J)}{J'-J}$. 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 2.2** 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.

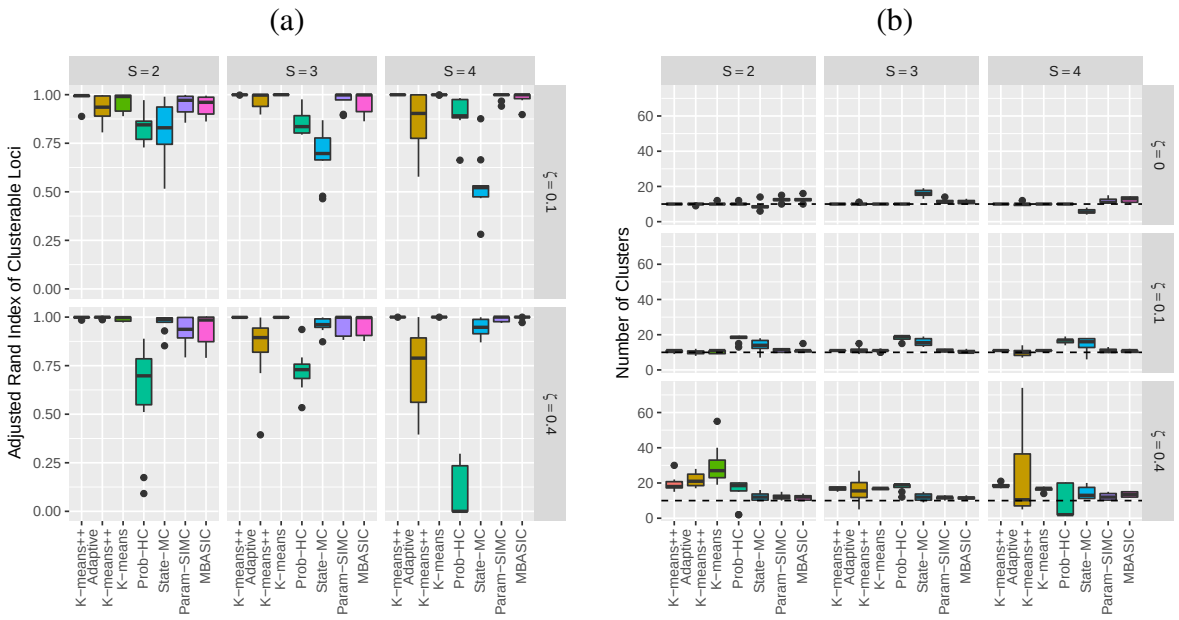


Figure A.2: S: Number of potential states. ($S = 2$ represents the usual binding state inference setting, while we may use more states to tell apart binding sites with different degree of signal strength.) ζ: Proportion of loci do not belong to any true cluster. (a) Clustering accuracy based on the adjusted Rand index after excluding unclusterable loci. (d) Number of clusters in fitted model.

Appendix B: Supplementary Figures for Chapter 3

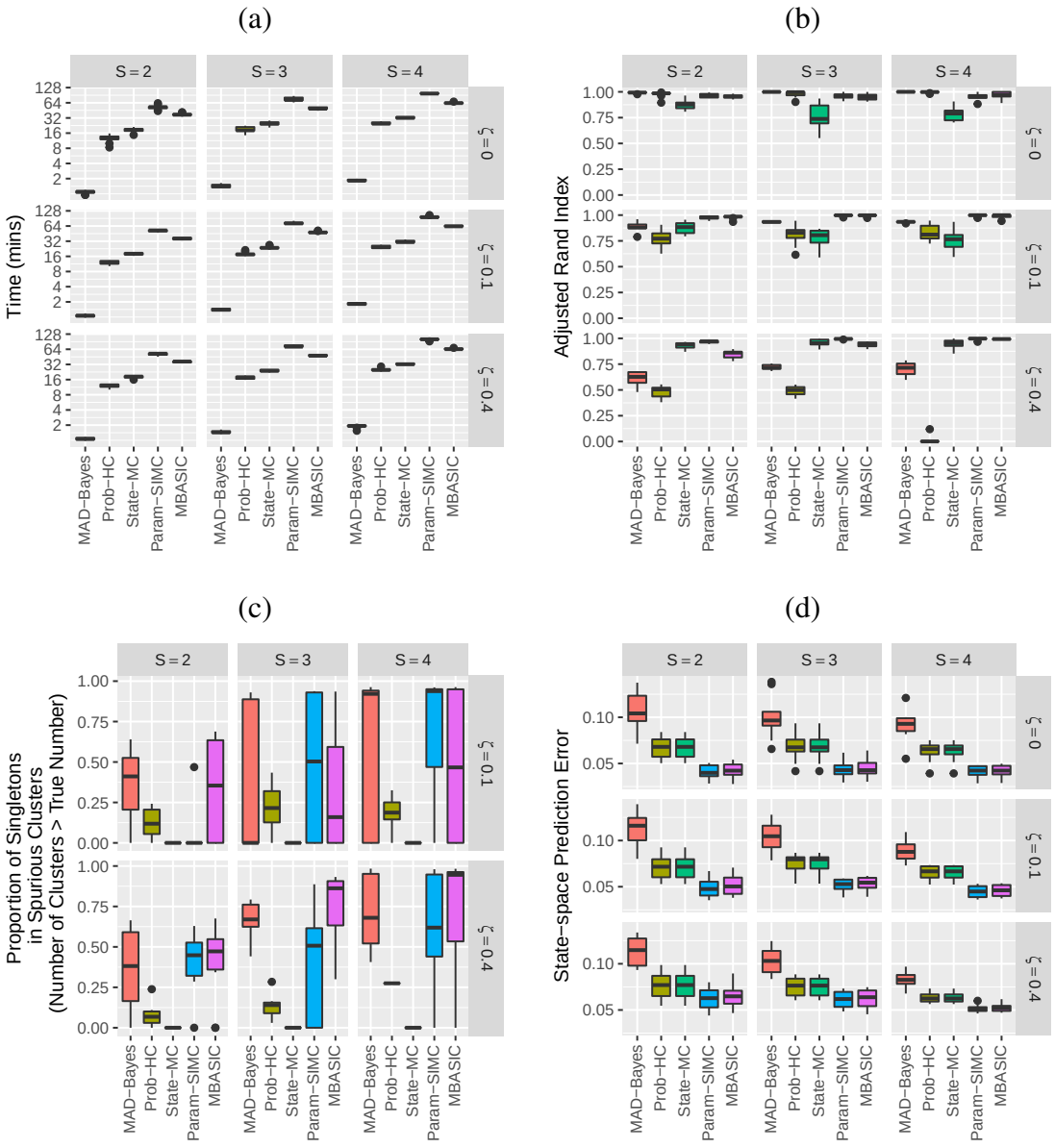


Figure B.1: Computational experiments with Zero-One-Inflated Poisson mixture distribution for inferring histone modification/TF binding sites. S: Number of potential states. ζ: Proportion of loci do not belong to any true cluster. (a) Run-time comparisons on a 64 bit machine with Intel Xeon 3.0GHz processor 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 unclusterable loci when 10% ($\zeta = 0.1$) and 40% ($\zeta = 0.4$) of the loci do not belong to a cluster.

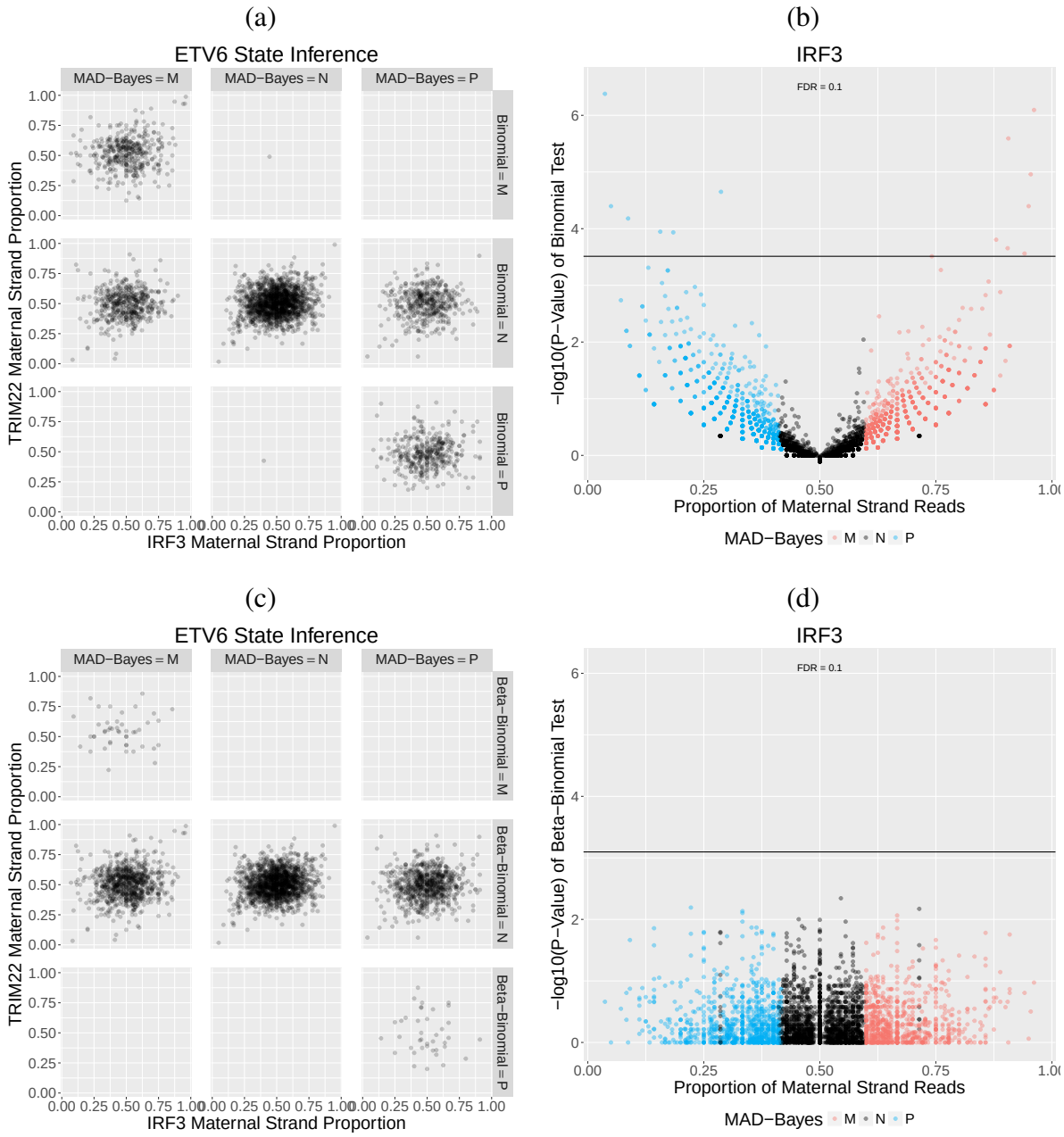


Figure B.2: (a) and (b): Comparison between the results of Binomial test and MAD-Bayes for IRF3. (c) and (d): Comparison between the results of Beta-Binomial test and MAD-Bayes for IRF3.

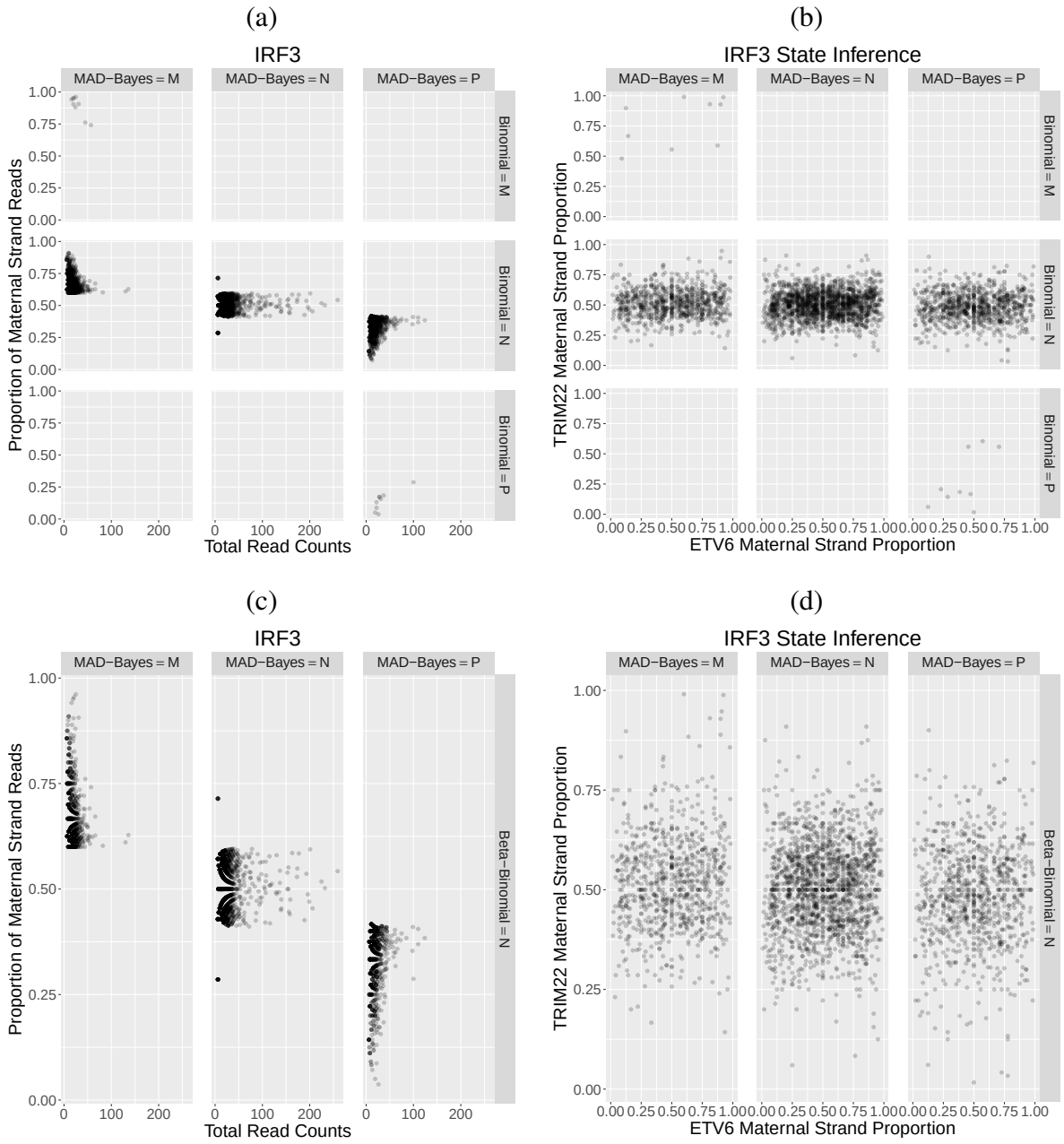


Figure B.3: (a) and (b): Comparison between the results of Binomial test and MAD-Bayes for TRIM22. (c) and (d): Comparison between the results of Beta-Binomial test and MAD-Bayes for TRIM22.

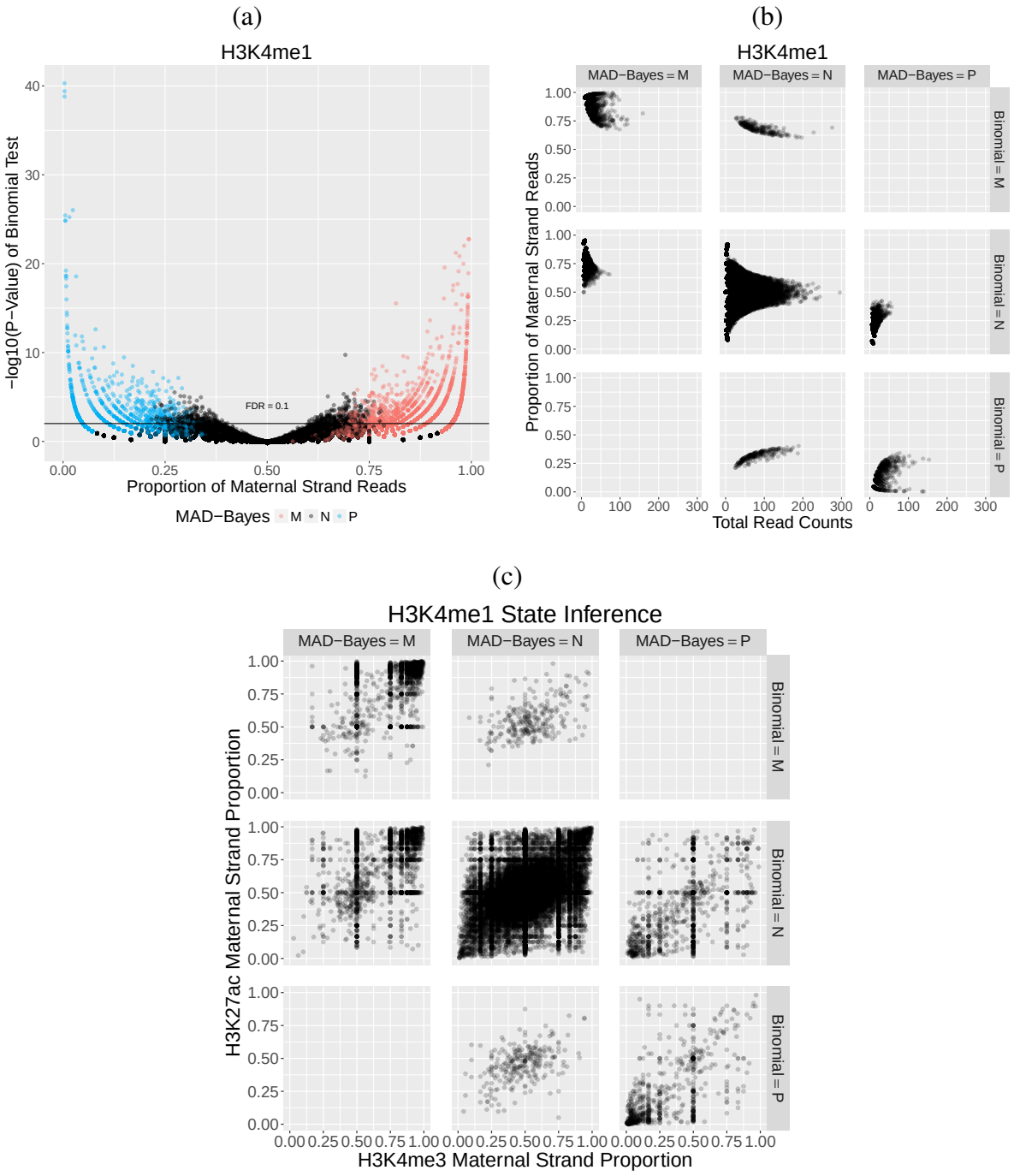


Figure B.4: Comparison between the results of Binomial test and MAD-Bayes for H3K4me1. (a): Binomial test p-values versus maternal strand count proportions. (b): Maternal strand count proportions versus total counts. (c): Maternal strand counts of H3K4me3 and H3K27ac under different inferred states of H3K4me1.

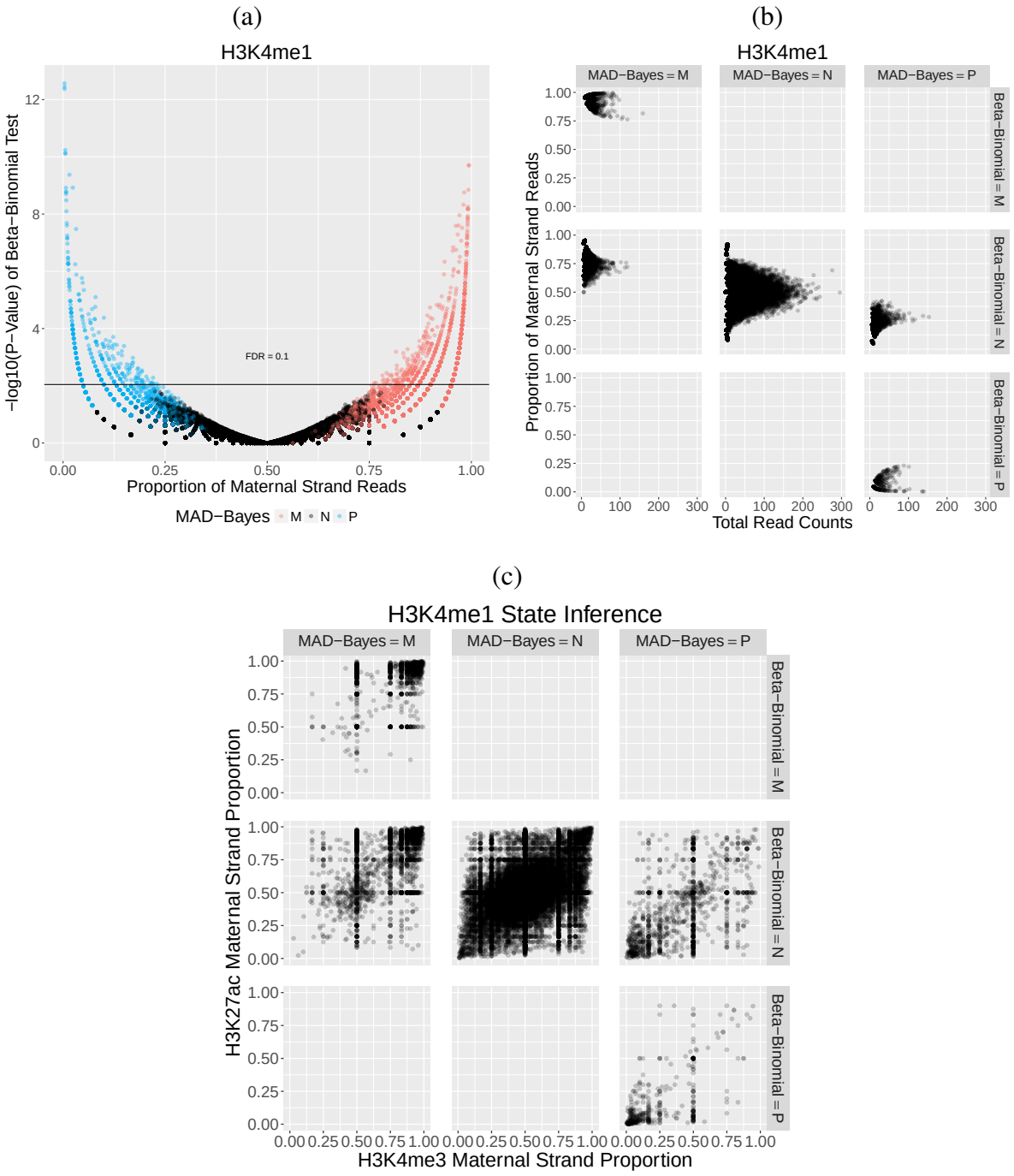


Figure B.5: Comparison between the results of Beta-Binomial test and MAD-Bayes for H3K4me1. (a): Binomial test p-values versus maternal strand count proportions. (b): Maternal strand count proportions versus total counts. (c): Maternal strand counts of H3K4me3 and H3K27ac under different inferred states of H3K4me1.

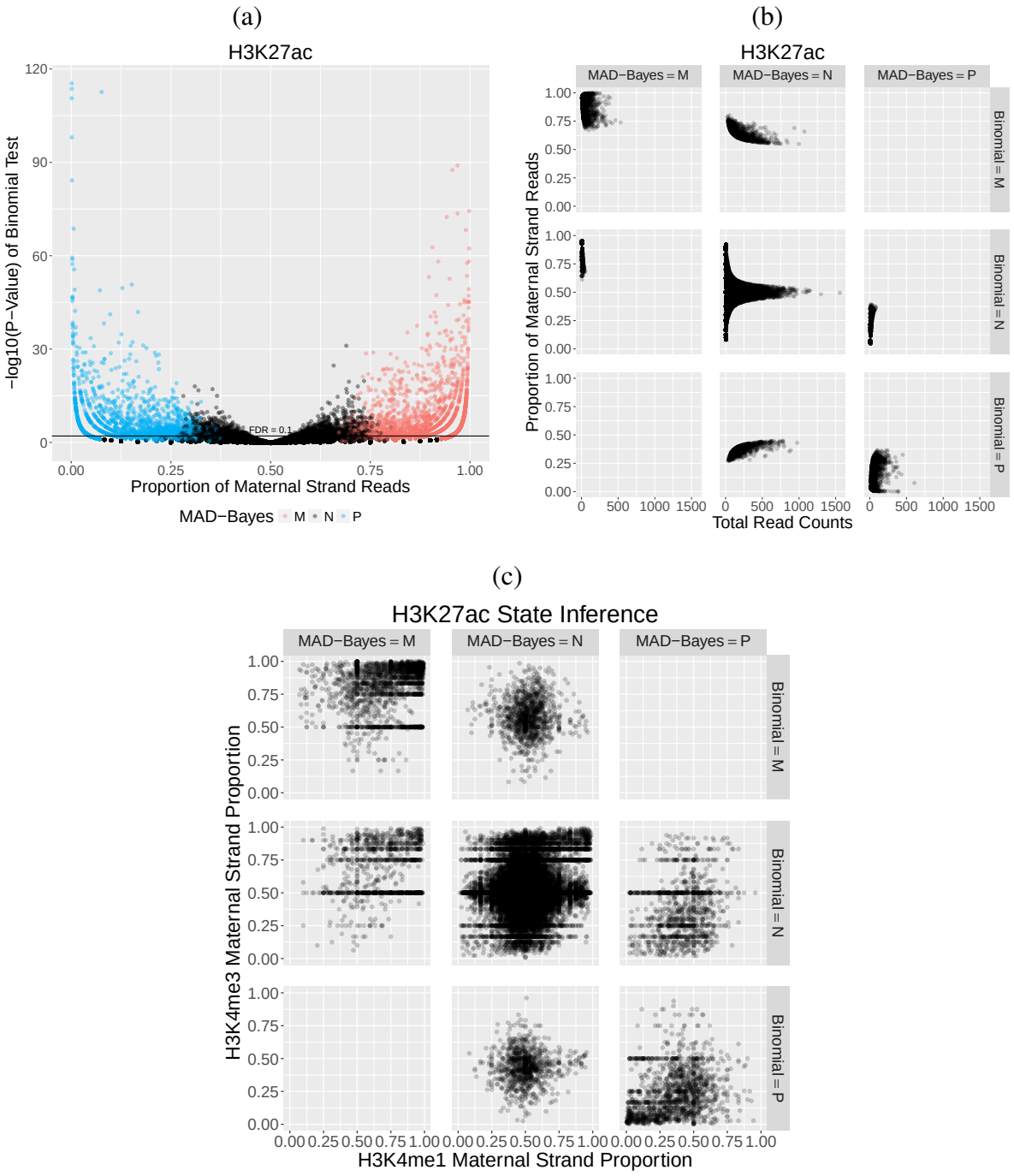


Figure B.6: Comparison between the results of Binomial test and MAD-Bayes for H3K27ac. (a): Binomial test p-values versus maternal strand count proportions. (b): Maternal strand count proportions versus total counts. (c): Maternal strand counts of H3K4me1 and H3K4me3 under different inferred states of H3K27ac.

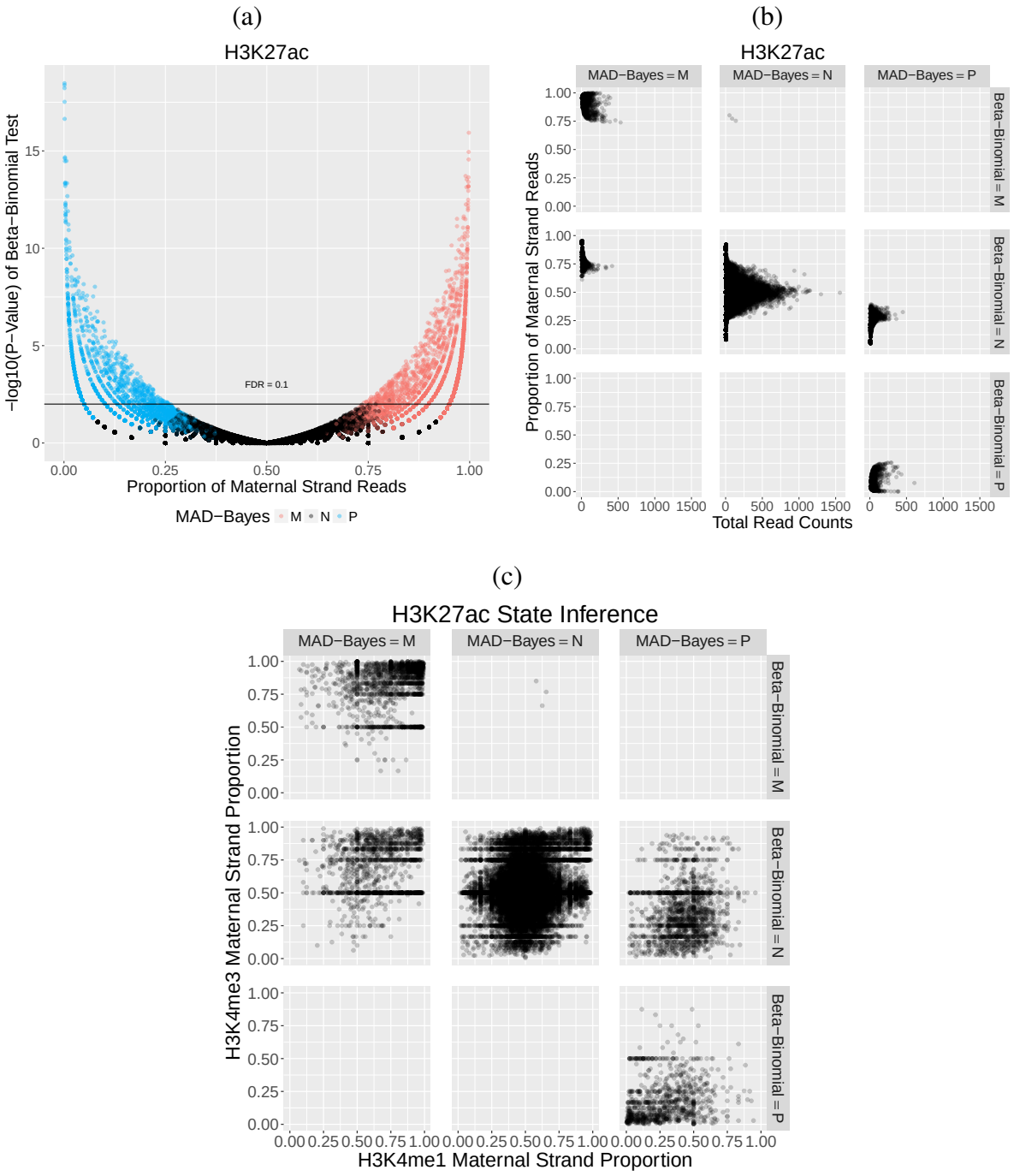


Figure B.7: Comparison between the results of Beta-Binomial test and MAD-Bayes for H3K27ac. (a): Binomial test p-values versus maternal strand count proportions. (b): Maternal strand count proportions versus total counts. (c): Maternal strand counts of H3K4me1 and H3K4me3 under different inferred states of H3K27ac.

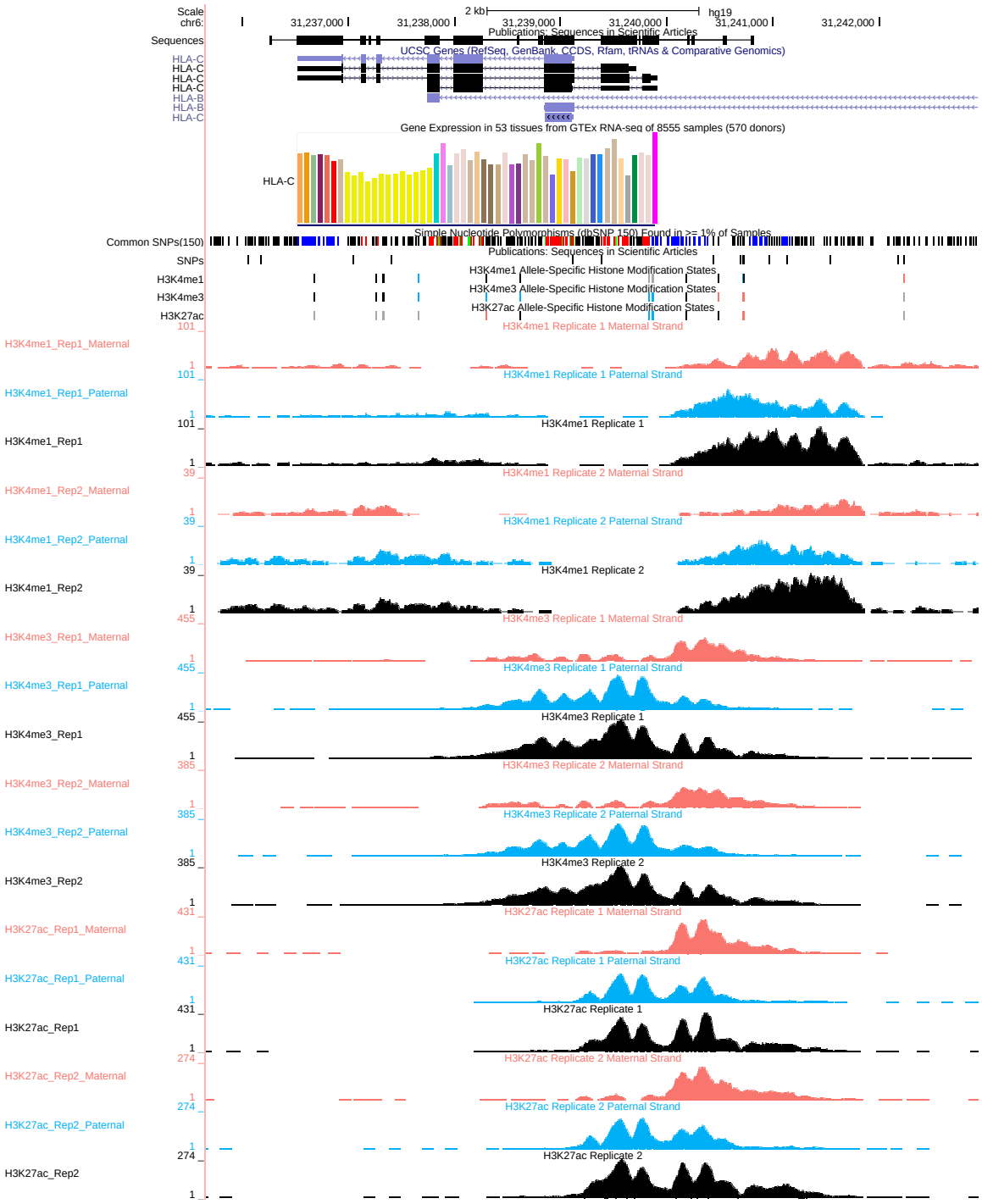


Figure B.8: H3K4me3 displayed preference for paternal strand in the upstream part of the peak chr6:31,235,668-31,242,931, but for maternal strand in the downstream part, with a string of allele specific states as: “N, N, N, N, P, P, P, P, P, P, N, M, M, M”.