jMOSAiCS: Joint Analysis of Multiple ChIP-seq Datasets

Xin Zeng[1], Rajendran Sanalkumar[3], Emery H. Bresnick[3], Hongda Li[4], Qiang Chang[4,5], and Sündüz Keleş*[1,2]

[1] Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin, USA. [2] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, USA. [3] Wisconsin Institutes for Medical Research, University of Wisconsin-Madison Carbone Cancer Center, Department of Cell and Regenerative Biology, University of Wisconsin School of Medicine and Public Health , Madison, Wisconsin, USA. [4] Genetics Training Program, Waisman Center, University of Wisconsin-Madison, Madison, Wisconsin, USA. [5] Department of Genetics and Neurology, University of Wisconsin-Madison, Madison, Wisconsin, USA.

Email: Sündüz Keleş - keles@stat.wisc.edu;

*Corresponding author

Abstract

The ChIP-seq technique enables genome-wide mapping of in vivo protein-DNA interactions and chromatin states. As ChIP-seq technology is becoming more economical, generation of multiple ChIP-seq samples to elucidate contributions of transcription factor binding and epigenome to phenotypic variation is becoming standard. Current analytical approaches for ChIP-seq analysis are largely geared towards single sample investigations, and therefore have limited applicability in comparative settings where the aim is to identify combinatorial patterns of enrichment across multiple datasets. We describe a novel probabilistic method, jMOSAiCS, for jointly analyzing multiple ChIP-seq datasets. We demonstrate the usefulness of this method with a wide range of data-driven computational experiments as well as with a case study of four histone modifications on GATA1 occupied segments during erythroid differentiation. Our analysis revealed a cluster of GATA1 occupied loci with a novel combinatorial pattern of histone modifications across the two cell lines involved in erythroid differentiation. We corroborated these data with gene expression data from erythroid differentiation and identified novel GATA1 target genes that exhibited changing patterns of histone modifications during erythroid differentiation. We validated a subset of the observed patterns for a number of these GATA1 occupied loci by independent quantitative real time ChIP analysis. Our results established that jMOSAiCS improves both the sensitivity and the specificity of detecting combinatorial enrichment patterns across multiple ChIP-seq datasets and is applicable with both transcription

factor and histone modification ChIP-seq data.

Background

The advent of high throughput next generation sequencing (NGS) technologies have revolutionized the fields of genetics and genomics by allowing rapid and inexpensive sequencing of billions of bases. Among the NGS applications, ChIP-seq (chromatin immunoprecipitation followed by NGS) is perhaps the most successful to date. Initial ChIP-seq studies largely focused on single sample investigations. However, as we begin to understand the role of epigenomics for biological variation, detailed comparisons of transcription binding (TF) and epigenomic marks between different tissues/individuals at single or multiple time points or developmental stages are becoming essential to understand the etiology and progression of many diseases. Therefore, comparative analysis of multiple ChIP-seq samples to identify combinatorial TF binding or epigenome profiles are rapidly emerging. Some examples include: (i) identifying differential binding of a TF or modification of a histone mark across multiple individuals, e.g., [1] studied variation in binding of NF-κB and RNA polymerase II(Pol II) across 10 individuals; [2] performed a genetic analysis of Ste12 binding in yeast by studying differential binding across 43 segregants of a cross between two yeast strains; (ii) genome-wide binding profiles of multiple TFs in a single tissue or cell line, e.g., comparative analysis of 22 *C. elegans* TFs [3]; (iii) time course or multiple developmental stage ChIP-seq experiments, e.g., Pol II binding at six developmental stages of *C. elegans* [3]; (iv) comparative analysis of binding profiles of one or more TFs with Pol II or modifications of histone marks, e.g., [4, 5].

Although there are already more than 30 algorithms/methods for ChIP-seq analysis (reviewed in [6]), all of them are limited to single sample analysis, and lack the ability to simultaneously compare multiple ChIP samples. Few number of available multi-sample ChIP-seq analysis tools are either specific to ChIP-seq design (e.g., [7] is specific to identifying chromatin states from ChIP-seq of histone modifications; [8] focuses on gene-centric analysis, exploratory [9] or difficult to generalize to more than two samples [10–12] due to computational reasons. This presents challenges for biological interpretation since combining results from individual analysis of multiple experiments can be a daunting task, especially for systematically enumerating combinatorial patterns of enrichment, controlling the overall false discovery rate (FDR), and prioritizing candidate regions for further experimental validation.

We introduce jMOSAiCS, **j**oint **Mo**del based One- and Two-**S**ample **A**nalysis and **I**nference for **C**hIP-**S**eq, as a probabilistic model for integrating multiple ChIP-seq datasets to identify combinatorial patterns of enrichment. The key components of jMOSAiCS are base models for the sequencing reads of each individual

ChIP-seq experiment and a model that governs the relationship of enrichment among different samples. We choose well-developed models from the ChIP literature for both of these key components. We evaluate jMOSAiCS with extensive data-driven computational experiments and compare it to both a separate analysis approach of multiple datasets and chromHMM [7]. We show that jMOSAiCS, which is applicable to both TF and histone ChIP-seq data, has better power and provides better false discovery rate control than the separate approach. We present an application of jMOSAiCS to multiple histone modifications during erythroid differentiation [5]. This analysis identified a cluster of GATA1 occupied loci exhibiting a pattern of enrichment that is different than that was identified by chromHMM analysis of the same datasets. We support our computational predictions by experimental validation of the predicted patterns of histone modifications for a number of selected loci. These results indicate that jMOSAiCS can reveal both global and local combinatorial enrichment patterns with high sensitivity.

Results

Model description

The most commonly used NGS platform for ChIP-seq is the Illumina platform [4, 13–15], which works by sequencing 25 to 100 bp from one or both ends of each DNA fragment in the sample of interest and generates millions of short reads. Standard pre-processing of reads involves mapping to a reference genome and summarizing total counts in each small non-overlapping interval (referred to as *bins*). Statistical analysis to detect enriched regions, i.e., peaks, in a single ChIP-seq sample is based on these counts and is carried out as a one- or two-sample analysis depending on the availability of a control sample. In contrast, inference from multiple samples involves classifying regions of genome into patterns of enrichment. For D samples, we can observe up to 2^D different enrichment patterns across genomic regions. For example, for $D = 2$, $\{00, 01, 10, 11\}$ denote the set of possible patterns: 00: not enriched in either of the samples; 10: enriched only in sample 1; 01: enriched only in sample 2; 11: enriched in both samples.

We consider I genomic regions of possibly different lengths across a reference genome. These initial set of I regions can be obtained by analyzing each dataset separately with one of the many available ChIP-seq analysis methods [6] and identifying regions of enrichment at a liberal FDR level. Let unobserved random variable $E_{id} \in \{0, 1\}$ denote enrichment for region i in dataset d. The overall enrichment pattern E_i is defined as the vector of $(E_{i1}, \cdots, E_{iD})$. Our joint model has three layers depicted in Figure 8. The first layer, named *E-layer*, concerns joint modeling of E_{id} for inferring combinatorial enrichment. This is enabled by defining a region-level random variable B_i as described below. The second layer, named *Y-layer*, concerns

observed read count data for region i across D samples: $Y_i = (Y_{i1}, \cdots, Y_{iD})$, where $Y_{id} = (Y_{id1}, \cdots, Y_{idL_i})$ and L_i denotes the number of bins in region i. In the case of a two-sample problem, Y_{idj} is vector-valued and denotes both the ChIP and control counts for jth bin of the ith region in dth sample. We assume that the counts from different samples are independent conditional on the enrichment pattern:

$$Y_{id} \perp Y_{id'} \mid E_i, \quad \forall d, d' = 1, \cdots, D, \quad \text{and hence} \quad Pr(Y_i) = \sum_{r=1}^{R} \left[\prod_{d=1}^{D} Pr(Y_{id} \mid E_i = r) \right] Pr(E_i = r),$$

where $r = 1, \cdots, R$ represents possible enrichment patterns. Note that $Pr(Y_{id} \mid E_i = r) = Pr(Y_{id} \mid E_{id} = r_d)$, $r_d = 0, 1$, and only concerns data for the L_i bins from the dth sample. $E_{id} = 0$ implies that all the bins in region i are from the background (unenriched) component in the dth sample. In contrast, if $E_{id} = 1$, one or more bins show enrichment. The third layer, named *Z-layer*, concerns Z_{idj} which we define as the bin-specific enrichment variable. If jth bin in ith region is enriched in dataset d, then $Z_{idj} = 1$ and 0 otherwise. We assume that $Z_{idj}, j = 1, \cdots, L_i \; \forall d, i$, are independent conditional on the region-specific enrichment indicator E_{id} and hence $Pr(Z_{id1}, \cdots, Z_{idL_i} \mid E_{id} = r_d) = \prod_{j=1}^{L_i} Pr(Z_{idj} \mid E_{id} = r_d)$.

The key to our joint modeling approach are the models we utilize for the *E-* and *Y-layers*. For the *E-layer*, we adopt the joint ChIP-chip model of JAMIE [16], which facilitates information sharing across experiments by capturing the correlation among datasets. In this model, the broad dependencies among the D samples are captured via unobserved variable B, where $B_i \in \{0, 1\}$ denotes whether region i is *potentially* enriched and E_{id} is defined to be 1 if region i is enriched in sample d. We assume that $E_{i1}, \cdots, E_{iD}$ are conditionally independent given B_i. Let $Pr(B_i = 1) = \tau_1$, $Pr(E_{id} = 1 \mid B_i = 1) = \eta_d$, and $Pr(E_{id} = 1 \mid B_i = 0) = 0$, i.e., the region cannot be enriched in any dataset if $B_i = 0$. Then, we have $Pr(E_{id} = r_d) = \tau_1 \eta_d^{r_d} (1 - \eta_d)^{1 - r_d} + (1 - \tau_1) \mathrm{I}(r_d = 0)$. The joint probability of $(E_{i1}, \cdots, E_{iD})$ is given by $Pr(E_{i1} = r_1, \cdots, E_{iD} = r_D) = \tau_1 \prod_{d=1}^{D} \eta_d^{r_d} (1 - \eta_d)^{1 - r_d} + (1 - \tau_1) \mathrm{I}(r_1 = 0, \cdots, r_D = 0)$.

For the *Y-layer*, we adopt the model-based approach of MOSAiCS [17] since MOSAiCS provides parametric models for read counts from both the enriched and unenriched regions in both the one- (without a control sample) and two-sample (with a control sample) problems. At the bin-level, $Y_{idj} \mid Z_{idj} = 0 \sim N_{idj}$, where $N_{idj} \sim NegBin\,(a, a/\mu_{idj})$ represents background read counts. Its mean μ_{idj} is parametrized as $\log \mu_{idj} = \beta_0 + \beta_1 X_{idj}^c$, where X_{idj} denotes the bin-level read counts in the control sample and c is a transformation parameter set data-adaptively. For one-sample analysis without a control sample or for two-sample analysis with a shallow sequenced control sample, MOSAiCS provides a parametrization of the bin-level counts that also depends on mappability and GC-content. For the enriched bins, $Y_{idj} \mid Z_{idj} = 1 \sim N_{idj} + S_{idj}$, where S_{idj} represents signal due to enrichment, i.e., protein binding or epigenomic marker modification. The

signal S_{idj} is modelled either as a single Negative Binomial distribution or a mixture of two Negative Binomial distributions. This choice is based on model fit and is determined through Bayesian Information Criterion (BIC) [18] by MOSAiCS. For model fitting, we utilize the fact that MOSAiCS provides fast and accurate estimates of dataset-specific background and signal distributions. Therefore, as part of model fitting, jMOSAiCS only needs to infer parameters associated with the B and E variables, namely τ_1 and η_d, $d = 1, \cdots, D$. In addition, jMOSAiCS provides posterior probabilities of the B and E variables that facilitate identification of region-specific enrichment patterns across the D datasets. We implemented jMOSAiCS as an `R` package and it is available from www.stat.wisc.edu/~keles (will be contributed to Bioconductor [19] and Galaxy [20] upon publication).

Data-driven computational experiments

We evaluated jMOSAiCS with data-driven computational experiments by simulating multiple ChIP-seq datasets based on model fits on actual datasets. We utilized ChIP-Seq experiments of STAT1 binding in interferon-γ-stimulated HeLa S3 cells by [21], H3K9me3 (repression mark) modification in peripheral blood mononuclear cells (PBMCs) from two unrelated individuals (Bresnick Lab, UW Madison), and methyl CpG binding protein MeCP2 in mouse cortex (Chang Lab, UW Madison). The model fits were obtained by MOSAiCS and the goodness-of-fit plots indicated satisfactory fits as discussed in [17]. We simulated multiple ChIP-seq datasets by using parameters that matched observed values in the STAT1, H3K9me3, and MeCP2 ChIP-seq experiments. The density plots of the read counts from the actual and sample simulated data are provided in Supplementary Figure S1 and indicate that the simulated data mimics the actual data well. In what follows, we first compared jMOSAiCs with a commonly practised separate analysis scheme where each ChIP-seq dataset is analyzed individually and the enrichment patterns are generated post-hoc analysis. Then, we compared jMOSAiCS to chromHMM [7] which is currently the state-of-the-art approach for discovering combinatorial patterns of chromatin states from multiple ChIP-seq data.

jMOSAiCS improves on separate analysis of multiple ChIP-seq datasets

Comparisons based-on data-driven STAT1 experiments: Analysis of multiple ChIP-seq datasets of two or more TFs under similar biological conditions

Data for this experiment uses the actual input experiment as the control sample and emulates ChIP-seq of multiple transcription factors in a single biological condition. Since we repeated each simulation experiment multiple times to assess variability, we restricted our data generation process to chromosome 12 of the human

genome to reduce computational time. We considered two settings with $D = 2$ and $D = 3$ datasets. The actual parameter values for each setting are summarized in Supplementary Table S1. For both settings, jMOSAiCS and the separate analysis approach, which identified enrichment for each individual dataset separately by MOSAiCS, are employed. Typical output from a ChIP-seq analysis is a ranked list of enriched regions. The length of the list can be based on a FDR cut-off, other types of Type-I error rate control, or the investigators may choose to consider certain number of high ranking regions. We evaluated the joint and the separate analysis approaches by taking this variation in reporting of the results into consideration. Specifically, we considered: (i) accuracy by plotting the proportion of correctly detected enriched regions obtained by the B variable and also correctly detected enrichments obtained by dataset-specific E variables as a function of top ranking enrichment regions; (ii) sensitivity by plotting the proportion of true set of enrichments that are detected as a function of nominal false discovery rate (reported are the total number of detected true enrichments identified at different FDR cut-offs divided by the total number true enrichments); (iii) false discovery rate control by plotting observed FDR as a function of target nominal FDR. Ranking of regions and FDR control for jMOSAiCS relied on the posterior inference with the B variable which captures whether or not any given region is enriched in any of the datasets and the E variables which infer whether or not the regions are enriched in specific datasets. We generated similar variables for the separate analysis in a post-hoc fashion after individual samples were analyzed with MOSAiCS.

Figure 9 summarizes these results for the $D = 2$ setting across 20 simulation runs (results for $D = 3$ are available in Supplementary Figure S2). This setting, on average, has $85,000$ enriched regions. Figure 9(a), which displays proportion of top ranking enriched regions that are true positives, indicates that jMOSAiCS and the separate analysis exhibit similar accuracy for the top 36% of the enriched regions; however, jMOSAiCS outperforms the separate approach significantly as we go down the list of top ranking regions. The differences in performances are significant both at the region-level (B-level, based on the B variable) for detecting whether or not there is any enrichment in a region in any of the D datasets and also at the individual dataset-level (E_1- and E_2-levels, based on the E variables). Beyond the 68% of the top enrichment regions (≥ 58000), the improvement in accuracy due to the joint analysis is about 10% at the individual dataset-level. In addition, jMOSAiCS exhibits much smaller variation in accuracy compared to the separate analysis as the number of top ranking regions considered increases. Since this setting had similar signal strengths for both datasets, dataset-specific accuracy improvements over the separate analysis captured by the E_1 and E_2 variables are similar.

Figure 9(b) evaluates the two approaches in terms of sensitivity and illustrates that jMOSAiCS has

better sensitivity than the separate approach at every nominal FDR level. Overall, jMOSAiCS identifies larger number of enriched regions and captures significantly higher proportion of the true set of enrichments compared to the separate approach at the same FDR level. When FDR is 0.01, the improvement in sensitivity is 9% at the B-level and more than 15% at the E-level. At the same FDR cutoff, jMOSAiCS identifies more true enrichments than the separate analysis. Next, we check how well the FDR is controlled by the two approaches in Figure 9(c), which depicts observed FDR across 20 simulations for different levels of nominal FDR. Overall, we observe that jMOSAiCS provides better FDR control than the separate approach and its FDR estimates at the E-level are more accurate. For the B-level, we observe some over-estimation of FDR by jMOSAiCS; however, this is still significant improvement over the separate analysis. Overall conclusions based on the H3K9me3 simulations which emulate data for a single epigenetic mark in two different conditions (two different individuals) agree with those of STAT1 results and the detailed results are provided in Supplementary Figure S3.

Comparisons based-on data-driven MeCP2 experiments: Joint analysis of replicate ChIP-seq experiments

ChIP-seq experiments are often carried out with at least two biological replicates to allow assessment of variability. Prior research suggests that non-specific biases such as GC content can vary significantly between biological replicates [17, 22]. As a result, it is not often clear whether or not data can be pooled at the biological replicate-level for the purpose of identifying enrichment. We studied a joint analysis strategy of multiple replicates with jMOSAiCS with a computational experiment based on MeCP2 binding in mouse. The data consisted of two biological replicates with 5 and 6 lanes of sequencing reads, respectively. The number of usable reads within a lane varied between 6.8 and 19.7 million reads. MOSAiCS provided adequate fits on each data set and the simulation parameters were set according to estimates from the MOSAiCS fits. Details on the parameter settings are available in Supplementary Table S1. Within this simulation, we varied the sequencing depth of one of the replicates (replicate 2) at 1, 3, and 6 lanes while keeping the other replicate at 5 lanes. One and three lane scenarios emulate the cases where one of the replicates has much lower sequencing depth than the other. This setting can arise in a variety of contexts, for example, when multiple samples are multiplexed together in one lane or when replicates are generated at different times. Figures 10, 11, and 12 summarize the results for these experiments. Figure 10(a) illustrates that, for lower depth scenarios of replicate 2, jMOSAiCS has significantly higher accuracy than the separate analysis at the B-level when inferring whether the regions are enriched in any of the replicate datasets. E-level comparisons of accuracy for replicate 2 (Figure 12(a)) reveal a consistent 15% difference in accuracy between jMOSAiCS and the separate approach. When both replicates have high sequencing depths, jMOSAiCS provides a

small but significant improvement over the separate analysis (jMOSAiCS (5-6) vs. Separate (5-6) across Figures 10(a), 11(a), and 12(a)). The differences in the sensitivities of the two approaches vary significantly with the number of lanes of replicate 2 (Figures 10(b), 11(b), and 12(b)). Overall, jMOSAiCS consistently identifies 10-15% more of the true enrichments when replicate 2 has lower depth. In Figure 11, as expected, the sensitivity of enrichment detection in replicate 1 is not affected by the number of lanes of replicate 2 in the separate analysis. However, jMOSAiCS also improves on this replicate as the number of lanes for the other replicate increases by sharing information across the two replicates through the B variable. The largest improvement due to jMOSAiCS is in the detection of enriched regions in the low depth replicate when it has only one lane of data (Figure 12(b)). In this setting, jMOSAiCS identifies 50% more of the true enrichment regions across all the nominal FDR levels. In Figures 10(c), 11(c), and 12(c), we observe that jMOSAiCS generally has more variable but accurate FDR estimation for both the B and E-levels. When replicate 1 has five lanes and replicate 2 only one lane, FDR controls by jMOSAiCS for the B- and E_2-levels are less accurate; however, the overall accuracy of jMOSAiCS is significantly better when fixed number of top ranking regions are considered (Figures 10(a) and 12(a)).

We also carried out a variation of this experimental setting by lowering the sequencing depths of both of the replicates to 1 and 3 lanes. The results are reported in Supplementary Figures S4, S5, and S6 and agree well with the overall conclusions reported here.

Comparison with chromHMM

chromHMM [7] is a hidden Markov model-based approach for partitioning a reference genome into multiple chromatin states based on multiple histone modification ChIP-seq datasets. The software accepts as input either aligned read files or enrichment/peak calls for each dataset. When provided with the aligned reads, it partitions the genome into 200 bps intervals and assigns each interval a 1 or 0 based on a local Poisson background distribution to depict enrichment. chromHMM aims to identify global patterns of enrichment and hence it approximates the space of 2^D enrichment patterns with a much smaller number as it is computationally prohibitive to consider the full state space with this model. As output, it reports the specific combination of epigenomic marks (enrichment patterns) associated with each chromatin state and the frequencies between 0 and 1 with which they occur. We compared jMOSAiCS and chromHMM in three settings using the data-driven experiments of STAT1 ChIP-seq data in HeLa cells. Although these initial parameters are derived from TF ChIP-seq data, they are able to generate ChIP-seq data with marginal density similar to those of histone data. The specific simulation settings are as follows:

SE1 Same as the STAT1 simulation described in the earlier section.

SE2 Lowered η_2 from 0.9 to 0.5 to increase the number of regions with 10 pattern.

SE3 Strengthened the ChIP signal by substituting b_1 and b_2 with $2 \times b_1$ and $2 \times b_2$.

One of the major differences between chromHMM and jMOSAiCS is that chromHMM models binary enrichment indicators as the observable data whereas jMOSAiCS models the actual read counts (Y-layer). In addition, jMOSAiCS can capture all possible enrichment patterns even for large number of datasets (D) because the joint distribution of the enrichment variables is governed by the univariate B variable. To investigate the effect of the binarization in chromHMM, we considered three versions of chromHMM: (i) original chromHMM; (ii) chromHMM coupled with true binarization; (iii) chromHMM where bin-level binarization is based on peak calling with MOSAiCS at nominal FDR levels of 0.05 and 0.2. Detailed results for setting SE2 are provided in Figure 13. Figure 13(a) summarizes enrichment pattern identification results for the 11 and 10 patterns based on the genome annotations obtained by jMOSAiCS and variations of chromHMM. The results for the 01 pattern are not displayed because there are very few regions with this pattern and they are mostly misannotated by chromHMM. Overall, these results illustrate that jMOSAiCS outperforms the 4-state chromHMM in this setting. When coupled with true binary data, chromHMM annotated all chromatin states accurately. Using peaks called by MOSAiCS increased the accuracy compared to original 4-state chromHMM but identified fewer correct regions in the 10 state. Figure 13(b) provides detailed comparison of jMOSAiCS with the 2-state chromHMM where chromHMM approximates the full state space of dimension 4 by only 2 states. A similar comparison between jMOSAiCS and the 4-state chromHMM is provided in Supplementary Figure S7. We observe that approximating the state space of dimension of 4 by 2 dimensions leads to significant loss in accuracy for chromHMM. At the individual dataset-level, the difference in accuracy between the 2-state chromHMM and jMOSAiCS can be as large as 20% (comparing jMOSAiCS-E2 with chromHMM-E2 in Figure 13(b)). The results for simulation settings SE1 and SE3 are similar and provided as Supplementary Figures S8 and S9.

Application to mouse ENCODE data of multiple histone modifications during erythroid differentiation

We applied jMOSAiCS to ChIP-seq data with antibodies specific to the histone modifications H3K4me3, H3K4me1, H3K27me3, and H3K9me3 in the G1E and G1E-ER4+E2 cells [5]. These data were generated as part of the mouse ENCODE project and analyzed by chromHMM to segment the mouse erythroid genome based on chromatin modifications in [5]. The original analysis by [5] focused on segmentation of GATA1

occupied segments since G1E cells are a GATA-(null) cell line derived from targeted disruption of GATA1 in embryonic stem cells whereas G1E-ER4 cells are G1E cells engineered to express a conditionally active estrogen receptor (ER) ligand binding domain fusion to GATA1 (ER-GATA1). When estradiol is added to the culture medium (G1E-ER4+E2), the ER-GATA1 fusion protein gets activated and binds to GATA1 specific sites. chromHMM analysis approximated $2^4 = 16$ dimensional state space with only 6 states. Our jMOSAiCS application explored the full state space and, in addition to the 6 states identified by chromHMM, identified 5 more states to which significant number of GATA1 occupied segments were assigned. Figure 14(a) enumerates the state space for jMOSAiCS and Figure 14(b) lists the number of GATA1 occupied segments for each state in the G1E and G1E-ER4+E2 cells. Overall, we observe that chromHMM captures broad dominating patterns and jMOSAiCS improves resolution for identifying local structures. In Figure 14(c), we provide normalized read data for the 316 GATA1 occupied peaks (with width less than 1400 bps out of a total of 371) identified to switch from state 1101 in G1E to state 1111 in G1E-ER4+E2. We note that chromHMM output does not include the 1101 or the 1111 pattern and distributes these loci over the 6 patterns it utilizes. However, as evidenced from the heatmaps, these GATA1 occupied segments lack the repressive mark H3K27me3 in G1E cells and exhibit the mark upon activation of GATA1 in G1E-ER4+E2.

We annotated these GATA1 occupied segments with respect to gene locations and identified that a large subset of them (48%) map to immediate $5'$ or $3'$-end, or within introns of known genes. We studied expression profiling data from GATA1-null erthyroid precursor cells that stably express a conditionally active allele of GATA1 fused to the estrogen receptor ligand binding domain (G1E-ER-GATA-1). Differential expression analysis of uninduced and beta-estradiol-induced G1E-ER-GATA-1 cells [23] identified Elf1, Atp6v1e1, Cmas, Ech1, Extl3, Rab4a, Casc3, and Lrrf1p2 as significantly induced upon GATA1 activation with beta-estradiol treatment for 24 hours (FDR adjusted p-value ≤ 0.05). Although H3K27me3 is conventionally viewed as inhibitory to transcription, [24] recently identified an enrichment profile of H3K27me3 in the promoter of genes associated with active transcription. The genes we identified constitute further examples of this class. Several of these significantly expressed genes have established functions in stem cell biology and hematopoiesis. For example, Elf1 is an Ets transcription factor involved in the control of hematopoiesis through participating in the transcriptional activation of the Stem Cell Leukemia (SCL)/T-cell Acute Lymphocytic Leukmia-1 (TAL1) gene [25, 26]. We performed quantitative ChIP analysis of these four loci and validated the H3K4me1, H3K4me3, H3K27me3, and H3K9me3 marks at these loci in beta-estradiol-induced G1E-ER-GATA-1 cells (Supplementary Table S2). We provide detailed read coverage plots of these regions in Supplementary Figures S10-S13 along with their chromHMM annotations to further support their

jMOSAiCS annotation.

Discussion

Integrative analysis of multiple ChIP-seq datasets for enumerating enrichment patterns is an emerging need. We have introduced jMOSAiCS to enable efficient one- or two-sample integrative analysis of multiple ChIP-seq datasets. jMOSAiCS capitalizes on the dataset-specific accurate model fits by MOSAiCS and efficient encoding of the joint distribution of the enrichment across multiple datasets by the JAMIE approach of [16]. Diagnostics is an important component of probabilistic model-based approaches. jMOSAiCS inherits the goodness-of-fit plots provided by MOSAiCS for model checking and diagnostics. In contrast to some of the few available joint analysis methods for multiple ChIP-seq data (e.g., [10]), jMOSAiCS can efficiently handle multiple datasets and is accurate at both obtaining global and local structures. A comparison of jMOSAiCS with chromHMM reveals that jMOSAiCS is better at identifying local structures since it can capture any specific enrichment pattern and does not rely on approximating the number of states with a smaller number of patterns. This observation is further supported by identification of a considerable number of GATA1 occupied segments in a different state than that was identified by chromHMM.

Our analyses illustrate that jMOSAiCS is powerful in analyzing biological replicates simultaneously when it is not appropriate to pool them due to non-specific sequencing biases such as the GC-content. When one or more of the replicates is shallowly sequenced compared to others, jMOSAiCS boosts the power for these replicates. Another particularly attractive use for jMOSAiCS is when the TF of interest interacts with reference genome through another DNA binding protein. For example, virus-host interactions are typically facilitated by virus proteins interacting with the host DNA via host proteins. Joint analysis of ChIP-seq data for the host and virus proteins has the potential to boost power for detecting regions enriched for the virus protein (e.g., [27]).

Meta-analysis of multiple samples is another integrative approach to multiple ChIP-seq samples. However, the focus of such meta approaches (e.g., MM-ChIP [28], ChIPMeta [29]) is the analysis of ChIP (-chip or -seq) data of the same protein under similar biological conditions but by different platforms or laboratories for the purpose of boosting power of peak detection. The focus in jMOSAiCS is combinatorial pattern detection across multiple datasets (same TF in different biological conditions or different TFs or epigenomic marks in the same biological conditions). Therefore, our computational experiments focused on comparing jMOSAiCS with chromHMM which is suitable for the latter task. jMOSAiCS can handle multiple ChIP-seq datasets with varying experimental parameters such library size and read length because marginal distributions of

read counts in each dataset are modelled in a dataset-specific manner.

jMOSAiCS currently implements a naive Bayes model for the joint distribution of the dataset-specific enrichment indicators. This model captures broad dependencies among the samples via an unobserved variable. A potential improvement is to consider how enrichment of a region in a sample depends on its enrichment in other samples. A general way to induce such a structure is by Bayesian Networks, where a directed acyclic graph represents the dependencies. Trees, which generalize first order Markov chains, and mixtures of trees for which efficient structure learning algorithms exist [30] are two appealing, flexible candidates that can encode for increasingly complex dependencies. Furthermore, they can be tailored for specific characteristics of analyzed samples, e.g., a Markov structure for time course ChIP-seq experiments.

Conclusion

jMOSAiCS facilitates joint analysis of multiple ChIP-seq datasets for both identifying enrichment patterns of a single TF across multiple conditions and characterizing enrichment patterns of multiple epigenomic marks in one or more conditions. Given model fits from the peak/enrichment caller MOSAiCS, a typical jMOSAiCS run takes about 30 minutes to identify combinatorial patterns of four datasets across the whole mouse genome with a single CPU on a 64 bit machine with Intel Xeon 3.0GHz processor.

Materials and Methods

Model fitting and parameter estimation in jMOSAiCS

Let f_{0d} and f_{1d} denote read count distributions for unenriched and enriched bins in dataset d. We will denote estimates of these by MOSAiCS with $\hat{f}_{0d}$ and $\hat{f}_{1d}$. When a region is not enriched in dataset d, data for all the bins within that region are generated from f_{0d}. Hence,

$$p_{0id} \equiv Pr(Y_{id}|E_{id}=0) = \prod_{l=1}^{L_i} f_{0d}(Y_{idl}).$$

If region i is enriched in dataset d, then read counts for one or more consecutive bins within region i are generated from f_{1d}. This enforces local spatial coherence and is motivated by the wide range of enriched region widths observed in ChIP-seq data of histone modifications. Note that this kind of spatial dependence is also capture by the chromHMM model. Let V_{id} denote the the number of enriched bins and $S_{id} \in \{1, \cdots, L_i\}$ the starting position of the set of enriched bins in region i. Then, we have

$$p_{1id} \equiv Pr(Y_{id}|E_{id}=1)$$

$$
\begin{aligned}
&= \sum_{v=1}^{L_i} Pr(Y_{id}|E_{id}=1,B_i=1,V_{id}=v)Pr(V_i=v|E_{id}=1,B_i=1) \\
&= \sum_{v=1}^{L_i}\left(\sum_{s=1}^{L_i-v+1} Pr(Y_{id}|E_{id}=1,B_i=1,V_{id}=v,S_{id}=s)Pr(S_{id}=s|E_{id}=1,B_i=1,V_{id}=v)\frac{1}{L_i}\right) \\
&= \sum_{v=1}^{L_i}\left(\sum_{s=1}^{L_i-v+1} Pr(Y_{id}|S_{id}=s,V_{id}=v,E_{id}=1,B_i=1)\frac{1}{L_i-v+1}\frac{1}{L_i}\right) \\
&= \sum_{v=1}^{L_i}\sum_{s=1}^{L_i-v+1}\left(\frac{1}{L_i}\frac{1}{L_i-v+1}\prod_{l=1}^{s-1} f_{0d}(Y_{idl})\prod_{l=s+v}^{L_i} f_{0d}(Y_{idl})\prod_{l=s}^{s+v-1} f_{1d}(Y_{idl})\right),
\end{aligned}
$$

where we assume that the run of enriched bins can start anywhere within the region with equal probability of $1/L_i$ and the length of the run has a uniform discrete distribution, i.e., $Pr(S_{id}=s|E_{id}=1,B_i=1,V_{id}=v)=1/(L_i-v+1)$, $s=1,\cdots,L_i-v+1$. The likelihood of full data is a product over I regions:

$$
\begin{aligned}
Pr(Y,E,B) &= \prod_{i=1}^{I} Pr(Y_i,E_i,B_i) \\
&= \prod_{i=1}^{I} Pr(Y_i|E_i,B_i)P(E_i|B_i)P(B_i) \\
&= \prod_{i=1}^{I}\left(\left[(1-\tau_1)\prod_{d=1}^{D}(1-E_{id})p_{0id}\right]^{1-B_i}\left[\tau_1\prod_{d=1}^{D}((1-\eta_d)p_{0id})^{1-E_{id}}(\eta_d p_{1id})^{E_{id}}\right]^{B_i}\right). \quad (1)
\end{aligned}
$$

We estimate f_{0d} and f_{1d} for each individual dataset separately using the MOSAiCS algorithm. Therefore, the quantities p_{0id} and p_{1id}, $i=1,\cdots,I$, $d=1,\cdots,D$ are fixed given $\hat{f}_{0d}$ and $\hat{f}_{1d}$. Because B, E, S and V are unobserved variables, we derive an Expectation-Maximization [31] algorithm to obtain maximum likelihood estimators of τ_1 and $\eta=(\eta_1,\cdots,\eta_d)$ based on the likelihood in (1). The full data log likelihood can be written as:

$$
\begin{aligned}
L(\tau_1,\eta) &= \sum_{i=1}^{I}\left[(1-B_i)\log(1-\tau_1)+B_i\log\tau_1\right] \\
&+ \sum_{i=1}^{I}\sum_{d=1}^{D}\left[B_i(1-E_{id})\log(1-\eta_d)+B_iE_{id}\log(\eta_d)+C\right],
\end{aligned}
$$

where C is a constant that does not contain the parameters to be estimated and can be computed given $\hat{f}_{0d}$ and $\hat{f}_{1d}$. Taking expectation of the full data likelihood conditional on observed read counts Y, we obtain the following E- and M-steps, where τ_1^t, η^t denote parameter estimates from the t-th iteration:

E-step:

$$
\begin{aligned}
a_i^{(t+1)} &\equiv \mathrm{E}(B_i|Y,\tau_1^t,\eta^t) \\
&= \frac{Pr(Y_i|B_i=1,\tau_1^t,\eta^t)\tau_1^t}{Pr(Y_i|B_i=1,\tau_1^t,\eta^t)\tau_1^t+Pr(Y_i|B_i=0,\tau_1^t,\eta^t)(1-\tau_1^t)}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\tau_1^t \prod_{d=1}^{D}[\eta_d^t p_{1id} + (1-\eta_d^t)p_{0id}]}{\tau_1^t \prod_{d=1}^{D}[\eta_d^t p_{1id} + (1-\eta_d^t)p_{0id}] + (1-\tau_1^t)\prod_{d=1}^{D} p_{0id}}, \\
b_{id}^{(t+1)} &\equiv \mathrm{E}(B_i E_{id} | Y, \tau_1^t, \eta^t) \\
&= \frac{\eta_d^t p_{1id} a_i^{(t+1)}}{\eta_d^t p_{1id} + (1-\eta_d^t)p_{0id}}.
\end{aligned}
$$

M-step:

$$
\begin{aligned}
\tau_1^{(t+1)} &= \frac{\sum_{i=1}^{I} a_i^{(t+1)}}{I}, \\
\eta_d^{(t+1)} &= \frac{\sum_{i=1}^{I} b_{id}^{(t+1)}}{\sum_{i=1}^{I} a_i^{(t+1)}}, \quad d = 1, \cdots, D.
\end{aligned}
$$

This EM algorithm converged within 100 iterations in both the computational experiments and the analysis of ChIP-seq data of histone modifications used in the case study. We used the posterior probabilities $Pr(B_i|Y_i, \hat{\tau}_1, \hat{\eta})$ and $Pr(E_{id}|Y_i, \hat{\tau}_1, \hat{\eta})$ for false discovery rate control with a direct posterior probability approach [32] in the computational experiments.

Computational experiments

All the computational experiments were based on the following procedure. The reference genome (human for STAT1 and H3K9me3 or mouse for MeCP2) was divided into bins (50 bp for STAT1, 250 bp for H3K9me3, and 200 bp for MeCP2) based on average fragment size in the actual experiment. Every consecutive $n \in \{3, 5\}$ bins were organized into non-overlapping regions to facilitate B-level data generation. For each region i, $i = \cdots, I$, the B_i variable was set to 1 with probability τ_1. If $B_i = 0$, then all the E_{id} and Z_{idj} variables were set to 0 for that region, indicating no enrichment for all the bins in the region across all the datasets. For regions with $B_i = 1$, E variable was simulated at the dataset-level, e.g., E_{id} was set to 1 with probability η_d. The bin-level Z variables were generated based on E_{id}. For $E_{id} = 1$, the region i should have at least one enriched bin in dataset d. To ensure this, we selected the bin that the enrichment starts within a region at random and allowed the number of consecutive bins with enrichment to vary within each region. For non-enriched bins, the Z_{idj} was set to 0 and the corresponding *Y-layer* data (read counts) were generated from the background distribution. For enriched bins, Z_{idj} was set to 1 or 2 with probabilities p_1 and $1 - p_1$, and denoted the components of the mixture distribution for the signal. Specifically, $Z_{idj} = 1$ implied that $Y_{idj} \sim N_{idj} + NegBin(b_1, c_1/(1 + c_1))$, whereas $Z_{idj} = 2$ referred to $Y_{idj} \sim N_{idj} + NegBin(b_2, c_2/(1 + c_2))$. We generated multiple ChIP-seq datasets by varying the signal component parameters b_1, b_2, c_1, c_2, and p_1 of this procedure according to the parameters estimated from the actual ChIP-seq studies (Supplementary Table S1).

Separate analysis of multiple ChIP-seq datasets and annotation of genomes into combinatorial patterns in the computational experiments

In the separate analysis, we analyzed each dataset by MOSAiCS [17]. This allowed us to quantify the gain due to the joint modeling approach rather than differences in modelling of the read count data by different ChIP-seq analysis methods. MOSAiCS reports bin-level posterior probabilities of enrichment (posterior probabilities at the Z-layer). For the sensitivity and empirical FDR calculations, enriched bins were identified at the various levels of nominal FDR using a direct posterior probability approach [32]. Then, dataset-specific E variables were set to 1 if there was at least one enriched bin in a region. Similarly, region-specific B variables were set to 1 if at least one of the E variables for a given region was set to 1. The accuracy calculations required ranking of regions based on the B and E variables. For this purpose, we followed a meta-analytic approach and used the maximum of bin-level posterior probabilities of enrichment within each region for inference at the E-level and the maximum within each region across D datasets for inference at the B-level. Then, these posterior probabilities were used for ranking the regions in the accuracy plots. We also considered FDR control over these meta-analytically defined B and E variables as an alternative to the above approach for identifying set of enriched regions in the separate analysis; however this modification yielded similar results and did not change the overall conclusions. Ranking for the joint analysis in the accuracy plots utilized posterior inferences for the B and E variables based on the jMOSAiCS model. Accuracy as a function of top number of detected enriched regions required ranking of regions by chromHMM. For each region, we summed over chromHMM estimated pattern probability times the pattern-specific emission probability of each bin within the region and generated pattern-specific posterior probabilities for ranking.

Comparison of chromHMM and jMOSAiCS required annotation of genome into TF binding/chromatin states based on the jMOSAiCS fit. We calculated the joint posterior probability of the E variables $Pr(E_{i1} = r_1, \cdots, E_{iD} = r_D \mid Y_i, \tau_1, \eta)$ for each combination of $r_1, \cdots, r_D$, where $r_i = 0, 1$. The enrichment pattern (or state) of each region is assigned as the one with the maximum joint posterior probability.

jMOSAiCS analysis of multiple histone modification ChIP-seq datasets from [5]

We partitioned the mouse genome into 200 bp intervals and applied jMOSAiCS to data from the G1E and G1E-ER4+E2 cells separately. Enriched regions were identified by controlling the FDR at 0.01 through the E-variable. In the downstream analysis, we focused on 11485 GATA1 occupied segments defined by [5] and enumerated H3K4me3, H3K4me1, H3K27me3, and H3K9me3 modification patterns of these regions across the two cell types. The size of the GATA1 occupied segments ranged from 400 bp to 36000 bp with a median

width of 800 bp.

Quantitative ChIP assay

Quantitative ChIP analysis was conducted with two independent biological replicates of beta-estradiol-induced G1E-ER-GATA-1 cells using control and specific antibodies as described in [33]. The relative levels of the specific histone marks are indicated in the Supplementary Table S2. The PCR primers used to analyze the four loci are provided in Supplementary Table S3.

Author's contributions

S.K. conceived and designed the method, computational experiments, and the data analysis and wrote the paper. X.Z. designed and implemented the method, computational experiments, and the data analysis, and wrote the paper. R.S. and E.H.B. performed experimental validation of the selected targets. H.L. and Q.C. contributed data. All authors read and approved the final manuscript.

Acknowledgements

H3K9me3 ChIP-seq was performed by Henriette O'Geen in the lab of Peggy Farnham (University of Southern California), using PBMC provided by Emery Bresnick (University of Wisconsin, Madison) and Swee Lay Thein (King's College London). We thank Jason Ernst, Ph.D., (Department of Biological Chemistry, UCLA) for discussions on the chromHMM software and Weisheng Wu (Hardison Lab, Penn State University) for information on the ChIP-seq datasets. This research is supported by a National Institutes of Health Grant (HG006716) to SK.

References

1. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, Hong MY, Karczewski KJ, Huber W, Weissman SM, Gerstein MB, Korbel JO, Snyder M: **Variation in transcription factor binding among humans**. *Science* 2010, **328**(5975):232–235.

2. Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M: **Genetic analysis of variation in transcription factor binding in yeast.** *Nature* 2010, **464**(7292):1187–1191.

3. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, Agarwal A, Alexander RP, Barber G, Brdlik CM, Brennan J, Brouillet JJ, Carr A, Cheung MS, Clawson H, Contrino S, Dannenberg LO, Dernburg AF, Desai A, Dick L, Dose AC, Du J, Egelhofer T, Ercan S, Euskirchen G, Ewing B, Feingold EA, Gassmann R, Good PJ, Green P, Gullier F, Gutwein M, Guyer MS, Habegger L, Han T, Henikoff JG, Henz SR, Hinrichs A, Holster H, Hyman T, Iniguez AL, Janette J, Jensen M, Kato M, Kent WJ, Kephart E, Khivansara V, Khurana E, Kim JK, Kolasinska-Zwierz P, Lai EC, Latorre I, Leahey A, Lewis S, Lloyd P, Lochovsky L, Lowdon RF, Lubling Y, Lyne R, MacCoss M, Mackowiak SD, Mangone M, McKay S, Mecenas D, Merrihew G, Miller DM, Muroyama A, Murray JI, Ooi SL, Pham H, Phippen T, Preston EA, Rajewsky N, Ratsch G, Rosenbaum H, Rozowsky J, Rutherford K, Ruzanov P, Sarov M, Sasidharan R, Sboner A, Scheid P, Segal E, Shin H, Shou C, Slack FJ, Slightam C, Smith R, Spencer WC, Stinson EO, Taing S, Takasaki T, Vafeados D, Voronina K, Wang G, Washington NL, Whittle CM, Wu B, Yan KK, Zeller G, Zha Z, Zhong M, Zhou X, modENCODE Consortium, Ahringer J, Strome S, Gunsalus KC, Micklem G, Liu XS, Reinke V, Kim SK, Hillier LW, Henikoff S, Piano F, Snyder M, Stein L, Lieb JD, Waterston RH: **Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project**. *Science* 2010, **330**(6012):1775–1787.

4. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome**. *Cell* 2007, **129**:823–837.

5. Wu W, Cheng Y, Keller CA, Ernst J, Kumar SA, Mishra T, Morrissey C, Dorman CM, Chen KB, Drautz D, Giardine B, Shibata Y, Song L, Pimkin M, Crawford GE, Furey TS, Kellis M, Miller W, Taylor J, Schuster SC, Zhang Y, Chiaromonte F, Blobel GA, Weiss MJ, Hardison RC: **Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration**. *Genome Research* 2011, **21**(10):1659–1671.

6. Wilbanks EG, Facciotti MT: **Evaluation of Algorithm Performance in ChIP-seq Peak Detection**. *PLoS ONE* 2010, **5**(7):e11471, [http://dx.doi.org/10.1371%2Fjournal.pone.0011471].

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

8. Ferguson JP, Cho JH, Zhao H: **A new approach for the joint analysis of multiple chip-seq libraries with application to histone modification.** *Statistical applications in genetics and molecular biology* 2012, **11**(3).

9. Ye T, Krebs AR, Choukrallah MA, Keime C, Plewniak F, Davidson I, Tora L: **seqMINER: an integrated ChIP-seq data interpretation platform**. *Nucleic acids research* 2011, **39**(6):e35–e35.

10. Johannes F, Wardenaar R, Colome-Tatche M, Mousson F, de Graaf P, Mokry M, Guryev V, Timmers HTM, Cuppen E, Jansen RC: **Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq**. *Bioinformatics* 2010, **26**(8):1000–1006.

11. Song Q, Smith AD: **Identifying dispersed epigenomic domains from ChIP-Seq data**. *Bioinformatics* 2011, **27**(6):870–1.

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

13. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells**. *Nature* 2007, **448**:653–560.

14. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in Vivo protein-DNA interactions**. *Science* 2007, **316**:1749–1502.

15. Seo YK, Chong HK, Infante AM, In SS, Xie X, Osborne TF: **Genome-wide analysis of SREBP-1 binding in mouse liver chromatin reveals a preference for promoter proximal binding to a new motif**. *PNAS* 2009, **106**(33):13765–9.

16. Wu H, Ji H: **JAMIE: joint analysis of multiple ChIP-chip experiments**. *Bioinformatics* 2010, **26**(15):1864–1870.

17. Kuan PF, Chung D, Pan G, Thomson J, Stewart R, Keleş S: **A statistical framework for the analysis of ChIP-Seq data**. *Journal of the American Statistical Association* 2011, **106**:891–903. [Software available on Galaxy http://toolshed.g2.bx.psu.edu/ and also on Bioconductor http://bioconductor.org/packages/2.8/bioc/html/mosaics.html].

18. Schwarz G: **Estimating the Dimension of a Model**. *The Annals of Statistics* 1978, **6**(2):461–464.

19. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini A, Sawitzki G, Smith C, Smyth G, Tierney L, Yang J, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biology* 2004, **5**(10):R80+, [http://dx.doi.org/10.1186/gb-2004-5-10-r80].

20. **Galaxy Tool Shed**. [http://toolshed.g2.bx.psu.edu/].

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

22. Benjamini Y, Speed TP: **Summarizing and correcting the GC content bias in high-throughput sequencing**. *Nucleic Acids Research* 2012, **40**(10):e72.

23. Fujiwara* T, andS Keleş* HO, Blahnik K, Linneman AK, Kang YA, Choi K, Farnham PJ, Bresnick EH: **Discovering hematopoietic mechanisms through genomewide analysis of GATA factor chromatin occupancy**. *Moelcular Cell* 2009, **36**(4):667–681. [*: co-first authors].

24. Young MD, Willson TA, Wakefield MJ, Trounson E, Hilton DJ, Blewitt ME, Oshlack A, Majewski IJ: **ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity**. *Nucleic Acids Research* 2011, [http://nar.oxfordjournals.org/content/early/2011/06/07/nar.gkr416.abstract].

25. Chan WY, Follows GA, Lacaud G, Pimanda JE, Landry JRR, Kinston S, Knezevic K, Piltz S, Donaldson IJ, Gambardella L, Sablitzky F, Green AR, Kouskoff V, Göttgens B: **The paralogous hematopoietic regulators Lyl1 and Scl are coregulated by Ets and GATA factors, but Lyl1 cannot rescue the early Scl-/- phenotype**. *Blood* 2007, **109**(5):1908–1916.

26. Göttgens B, Broccardo C, Sanchez MJ, Deveaux S, Murphy G, Göthert J, Kotsopoulou E, Kinston S, Delaney L, Piltz S, Barton L, Knezevic K, Erber W, Begley C, Frampton J, Green A: **The scl +18/19 stem cell enhancer is not required for hematopoiesis: identification of a 5' bifunctional hematopoietic-endothelial enhancer bound by Fli-1 and Elf-1**. *Molecular and cellular biology* 2004, **24**:1870–1883.

27. Zhao B, Zou J, Wang H, Johannsen E, Peng CW, Quackenbush J, Mar JC, Morton CCC, Freedman ML, Blacklow SC, Aster JC, Bernstein BE, Kieff E: **Epstein-Barr virus exploits intrinsic B-lymphocyte transcription programs to achieve immortal cell growth.** *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(36):14902–14907.

28. Chen Y, Meyer CA, Liu T, Li W, Liu JS, Liu XS: **MM-ChIP enables integrative analysis of cross-platform and between-laboratory ChIP-chip or ChIP-seq data**. *Genome Biology* 2011, **12**(2):R11.

29. Choi H, Nesvizhskii AI, Ghosh D, Qin ZS: **Hierarchical hidden Markov model with application to joint analysis of ChIP-chip and ChIP-seq data**. *Bioinformatics* 2009, **25**(14):1715–1721.

30. Friedman N, Geiger D, Goldszmidt M: **Bayesian network classifiers**. *Machine Learning* 1997, **29**:131–163.

31. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm**. *JRSSB* 1977, **39**:1–38.

32. Newton MA, Noueiry A, Sarkar D, Ahlquist P: **Detecting differential gene expression with a semiparametric hierarchical mixture model**. *Biostatistics* 2004, **5**(2):155–176.

33. Im H, Grass JA, Johnson KD, Boyer ME, Wu J, Bresnick EH: **Measurement of protein-DNA interactions in vivo by chromatin immunoprecipitation**. *Methods in Molecular Biology* 2004, **284**:129–146.

Figures

Figure 1: *Pictorial depiction of the jMOSAiCS model for a region across two ChIP-seq datasets.* Region i consists of three bins. The B variable governs whether or not the region is enriched in any of the two samples. E variables denote sample-specific enrichments and are conditionally independent given the B variable. Z variables depict enrichment at the bin-level and are conditionally independent given the sample-specific E variables. When $E_{id} = 1$, one or more consecutive Z variables are set to 1 to capture enrichment. Observed read count Y can be scalar or vector-valued depending on the availability of a control input sample. Data fits at the Y-layer are obtained by MOSAiCS [17] on individual samples and evaluated by the goodness-of-fit (GOF) plots.

Figure 2: *Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from the STAT1 ChIP-seq experiment.* 'jMOSAiCS-B', 'jMOSAiCS-E1', and 'jMOSAiCS-E2' represent results derived from posterior probability inferences of the B, E_1, and E_2 variables. 'Separate-B', 'Separate-E1', and 'Separate-E2' represent results derived from separate analysis of each dataset.

Figure 3: *Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from the MeCP2 ChIP-seq experiment.* Comparisons of region-level (B) results of jMOSAiCS and separate analysis. 'jMOSAiCS (x-y)' and 'Separate (x-y)' refer to jMOSAiCS and separate analysis of x lanes of replicate 1 with y lanes of replicate 2.

Figure 4: *Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from the MeCP2 ChIP-seq experiment.* Comparison of dataset-specific region-level enrichment detection (E_1) results of jMOSAiCS and separate analysis on replicate 1. 'jMOSAiCS (x-y)' and 'Separate (x-y)' refer to jMOSAiCS and separate analysis of x lanes of replicate 1 with y lanes of replicate 2.

Figure 5: *Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from MeCP2 ChIP-seq data.* Comparison of dataset-specific region-level enrichment detection (E_2) results of jMOSAiCS and separate analysis on replicate 2 for which the number of data lanes varies. 'jMOSAiCS (x-y)' and 'Separate (x-y)' refer to jMOSAiCS and separate analysis of x lanes of replicate 1 with y lanes of replicate 2.

Figure 6: *Comparisons between jMOSAiCS and chromHMM based on data simulated from ChIP-seq experiment of STAT1 in HeLa3 cells (Setting SE2).* (a) Identification of combinatorial patterns: '11': enriched in both samples; '10': enriched only in sample 1. 'True': number of enriched regions; 'chromHMM': results by original 4-state chromHMM; 'chromHMM-true': 4-state chromHMM coupled with true binary data for the bins; 'chromHMM-0.05': 4-sate chromHMM coupled with MOSAiCS binarization of the bins at an FDR of 0.05; 'chromHMM-0.2': 4-state chromHMM coupled with MOSAiCS binarization of the bins at an FDR of 0.2. TP and FP denote true and false positives, respectively. (b) Accuracy of enrichment detection at the region (B) and dataset-specific region (E_1 and E_2) levels by jMOSAiCS and 2-state chromHMM.

Figure 7: *Analysis of mouse ENCODE histone ChIP-seq datasets.* (a) List of combinatorial patterns identified by jMOSAiCS. Patterns 1-6 are also identified by chromHMM. (b) Changes in chromatin states between G1E and G1E-ER4+E2 cells for DNA segments occupied by GATA1 in the latter cells. (c) Heatmap of normalized raw data for a group of 316 GATA1 occupied segments identified to switch from '1101' in G1E cells to '1111' in G1E-ER4+E2 cells by jMOSAiCS. Enriched regions (excluding segments longer than 1400 bp in size) identified across different marks are aligned and depicted in between the dashed lines.

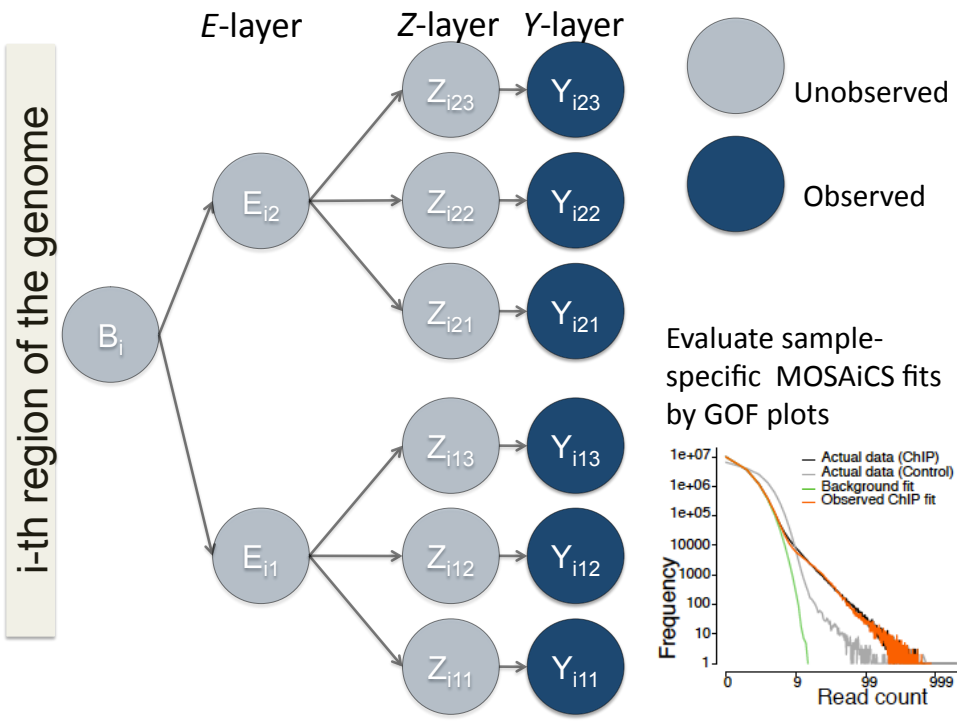


Figure 8: *Pictorial depiction of the jMOSAiCS model for a region across two ChIP-seq datasets.* Region i consists of three bins. The B variable governs whether or not the region is enriched in any of the two samples. E variables denote sample-specific enrichments and are conditionally independent given the B variable. Z variables depict enrichment at the bin-level and are conditionally independent given the sample-specific E variables. When $E_{id} = 1$, one or more consecutive Z variables are set to 1 to capture enrichment. Observed read count Y can be scalar or vector-valued depending on the availability of a control input sample. Data fits at the Y-layer are obtained by MOSAiCS [17] on individual samples and evaluated by the goodness-of-fit (GOF) plots.

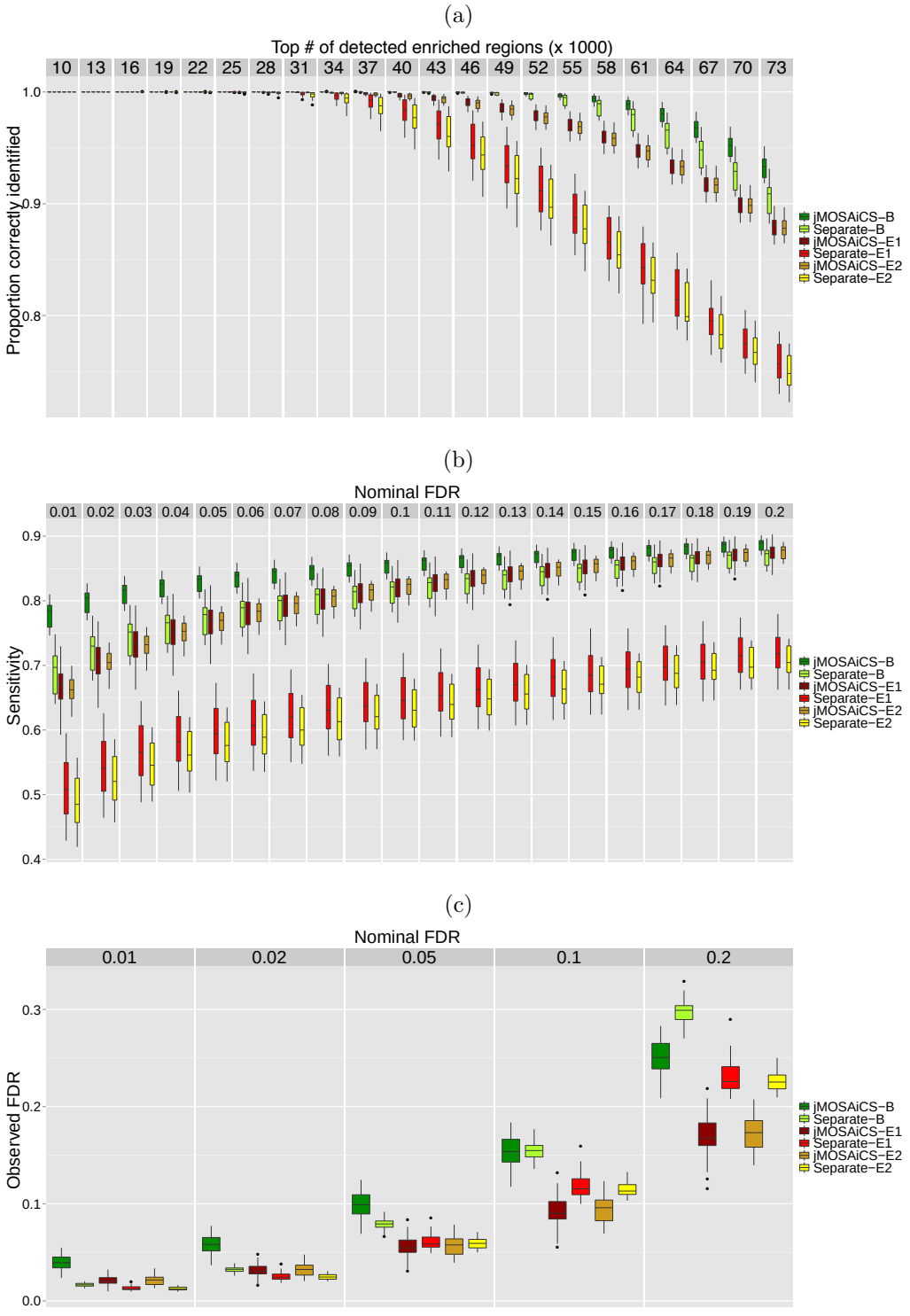


Figure 9: *Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from the STAT1 ChIP-seq experiment.* 'jMOSAiCS-B', 'jMOSAiCS-E1', and 'jMOSAiCS-E2' represent results derived from posterior probability inferences of the B, E_1, and E_2 variables. 'Separate-B', 'Separate-E1', and 'Separate-E2' represent results derived from separate analysis of each dataset.

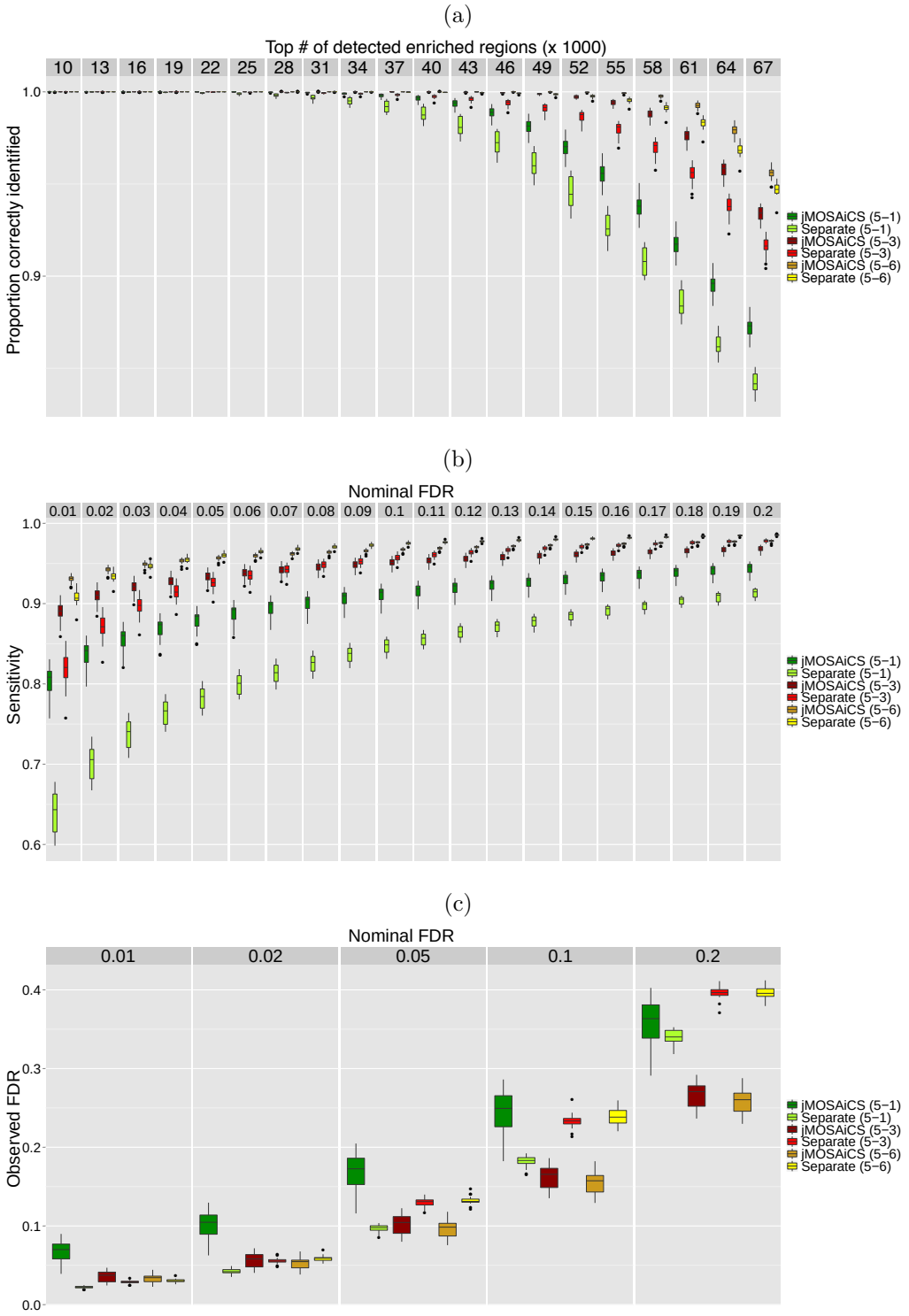


Figure 10: *Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from the MeCP2 ChIP-seq experiment.* Comparisons of region-level (B) results of jMOSAiCS and separate analysis. 'jMOSAiCS (x-y)' and 'Separate (x-y)' refer to jMOSAiCS and separate analysis of x lanes of replicate 1 with y lanes of replicate 2.

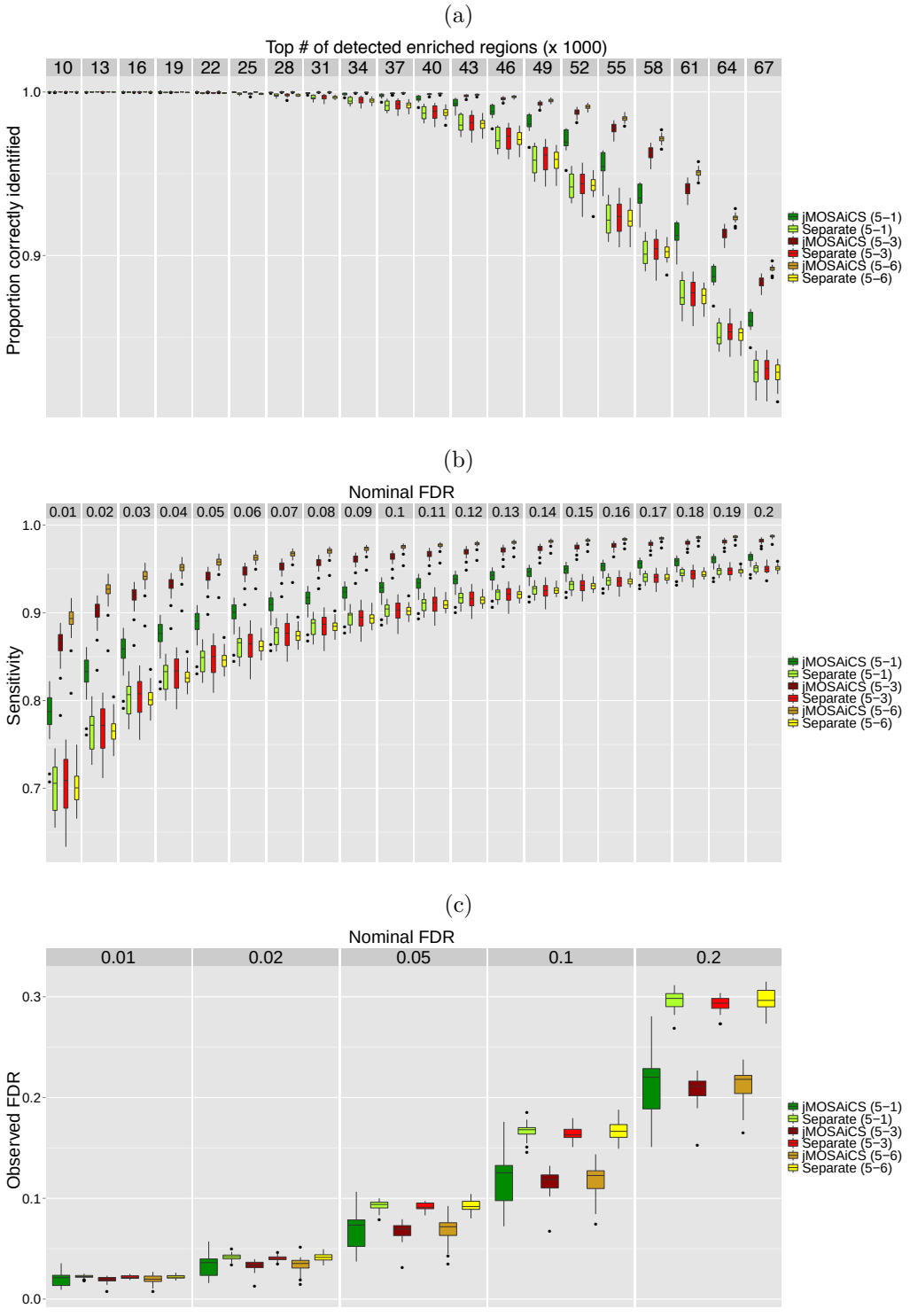


Figure 11: *Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from the MeCP2 ChIP-seq experiment.* Comparison of dataset-specific region-level enrichment detection (E_1) results of jMOSAiCS and separate analysis on replicate 1. 'jMOSAiCS (x-y)' and 'Separate (x-y)' refer to jMOSAiCS and separate analysis of x lanes of replicate 1 with y lanes of replicate 2.

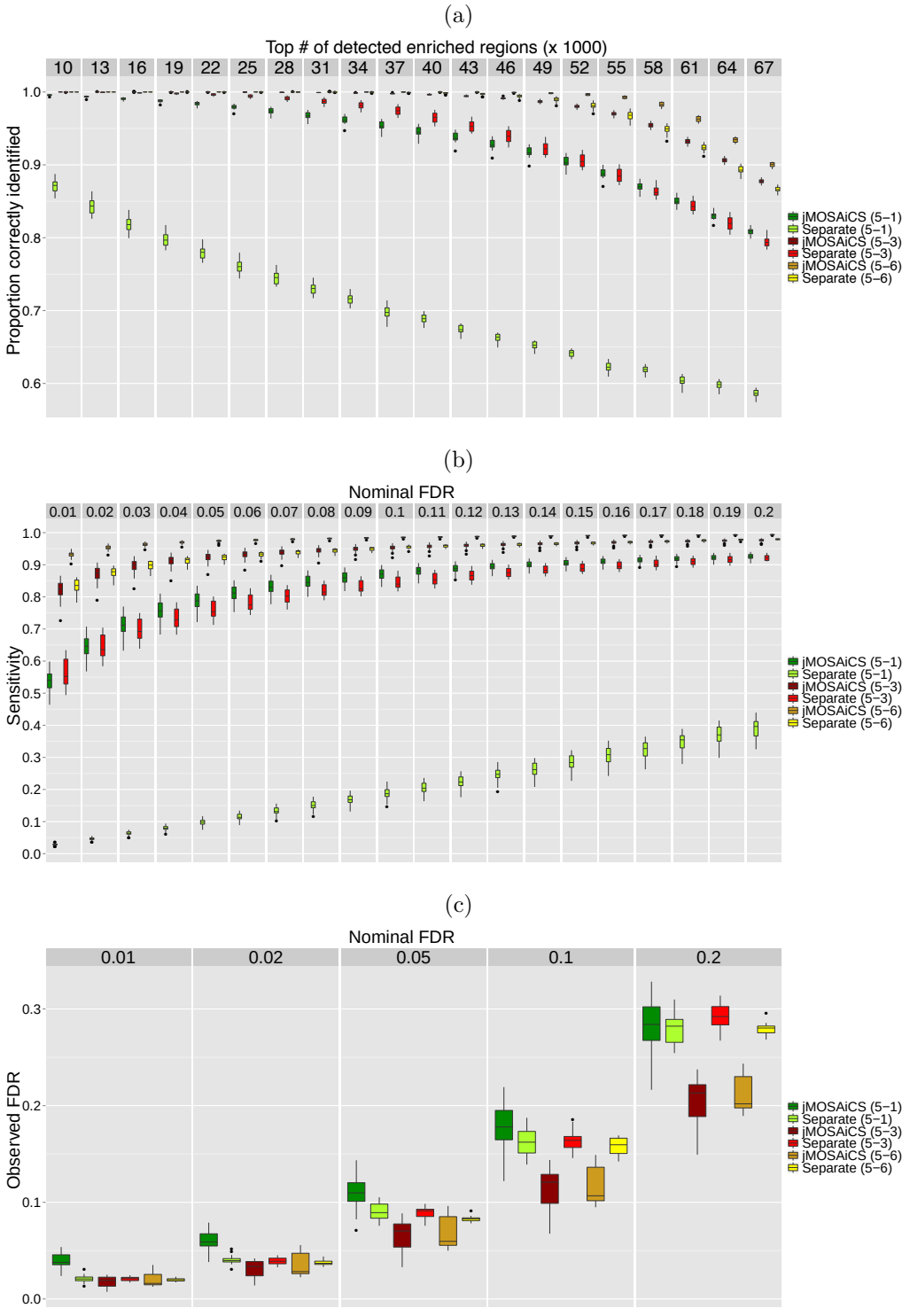
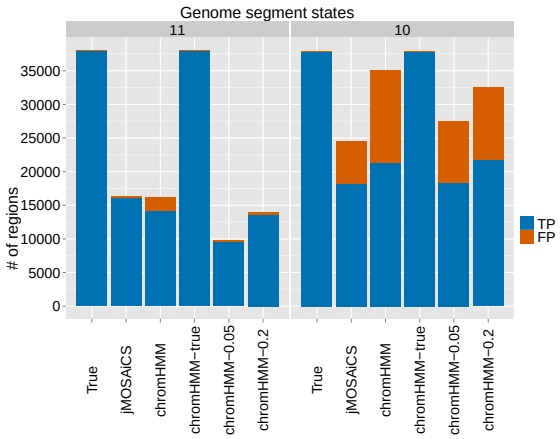


Figure 12: *Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from MeCP2 ChIP-seq data.* Comparison of dataset-specific region-level enrichment detection (E_2) results of jMOSAiCS and separate analysis on replicate 2 for which the number of data lanes varies. 'jMOSAiCS (x-y)' and 'Separate (x-y)' refer to jMOSAiCS and separate analysis of x lanes of replicate 1 with y lanes of replicate 2.

(a)



(b)

Figure 13: *Comparisons between jMOSAiCS and chromHMM based on data simulated from ChIP-seq experiment of STAT1 in HeLa3 cells (Setting SE2).* (a) Identification of combinatorial patterns: '11': enriched in both samples; '10': enriched only in sample 1. 'True': number of enriched regions; 'chromHMM': results by original 4-state chromHMM; 'chromHMM-true': 4-state chromHMM coupled with true binary data for the bins; 'chromHMM-0.05': 4-sate chromHMM coupled with MOSAiCS binarization of the bins at an FDR of 0.05; 'chromHMM-0.2': 4-state chromHMM coupled with MOSAiCS binarization of the bins at an FDR of 0.2. TP and FP denote true and false positives, respectively. (b) Accuracy of enrichment detection at the region (B) and dataset-specific region (E_1 and E_2) levels by jMOSAiCS and 2-state chromHMM.

(a)

State		H3K4me1	H3K4me3	H3k27me3	H3K9me3
1	Active	1	1	0	0
2	Active	1	0	0	0
3	Inactive	1	0	1	0
4	Inactive	0	0	1	0
5	Inactive	0	0	0	1
6	Inactive	0	0	0	0
7	Inactive	0	0	1	1
8	Active	0	1	0	0
9	Bivalent	0	1	0	1
10	Inactive	0	1	1	0
11	Bivalent	0	1	1	1
12	Bivalent	1	0	0	1
13	Bivalent	1	1	0	1
14	Active	1	1	1	0
15	Bivalent	1	1	1	1

(b)

Inactive or Bivalent to Active | No Change | Active to Inactive or Bivalent

Number of GATA1 OSs	Chromatin state in G1E-ER4+E2														
Chromatin state in G1E	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	7416	256	10	0	1	3	0	152	1	6	3	20		325	91
2	186	77	3	0	0	2	0	0	0	0	0	11	14	14	3
3	16	1	1	0	0	0	1	0	0	0	0	1	2	7	3
4	0	0	1	1	0	1	0	0	0	1	0	0	0	0	0
5	8	1	0	0	3	0	0	0	0	0	1	2	5	0	0
6	96	41	6	0	7	11	1	3	0	1	1	5	5	13	3
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	51	6	2	0	0	4	1	53	2	9	3	2	7	29	7
9	11	0	0	0	0	0	1	0	27	2	23	0	12	6	19
10	4	0	0	0	0	0	0	0	0	2	0	0	0	15	4
11	1	0	0	0	0	0	0	0	0	0	1	0	0	0	5
12	2	3	0	0	1	0	0	0	0	0	0	2	9	1	1
13	169	6	0	0	0	0	0	0	31	2	73	19	296	29	371
14	288	6	1	0	0	0	0	1	0	5	1	1	21	370	107
15	22	0	0	0	0	0	0	0	0	0	4	0	21	26	120

(c)

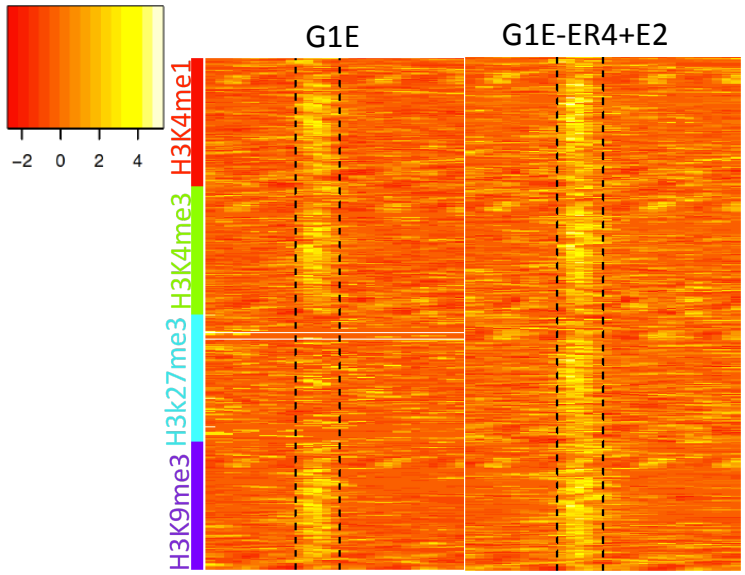


Figure 14: *Analysis of mouse ENCODE histone ChIP-seq datasets.* (a) List of combinatorial patterns identified by jMOSAiCS. Patterns 1-6 are also identified by chromHMM. (b) Changes in chromatin states between G1E and G1E-ER4+E2 cells for DNA segments occupied by GATA1 in the latter cells. (c) Heatmap of normalized raw data for a group of 316 GATA1 occupied segments identified to switch from '1101' in G1E cells to '1111' in G1E-ER4+E2 cells by jMOSAiCS. Enriched regions (excluding segments longer than 1400 bp in size) identified across different marks are aligned and depicted in between the dashed lines.

Additional Files

Additional file 1: Supplementary Materials

This file contains further details on and additional results from the computational experiments presented as supplementary text and figures.

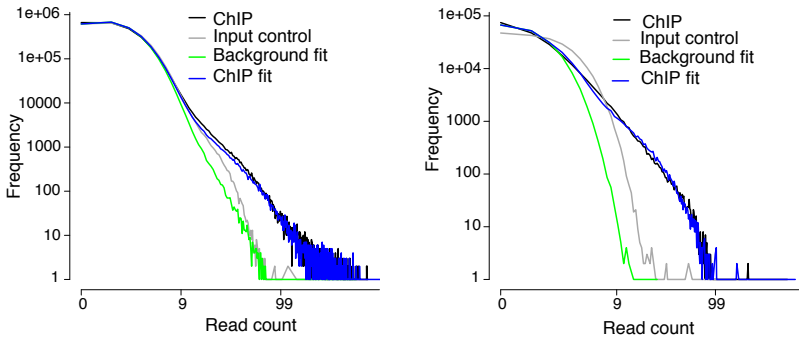
Additional file 2: An initial implementation of the R package for jMOSAiCS

Additional file 3: Vignette for the R package `jmosaics`

Supplementary Materials

Density plots for experimental and simulated data

(a) STAT1 - experimental data (b) H3K9me3 - experimental data



(c) STAT1 - simulated data (d) H3K9me3 - simulated data

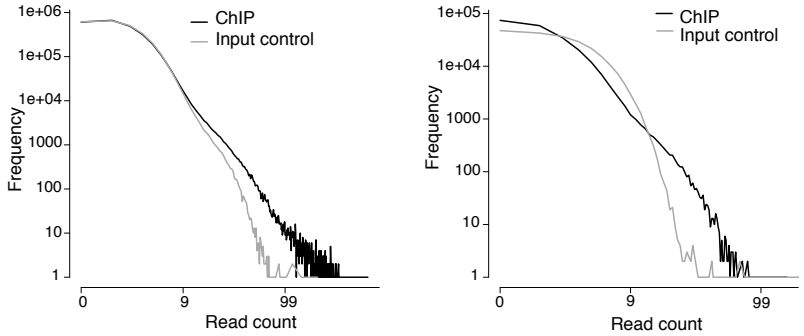
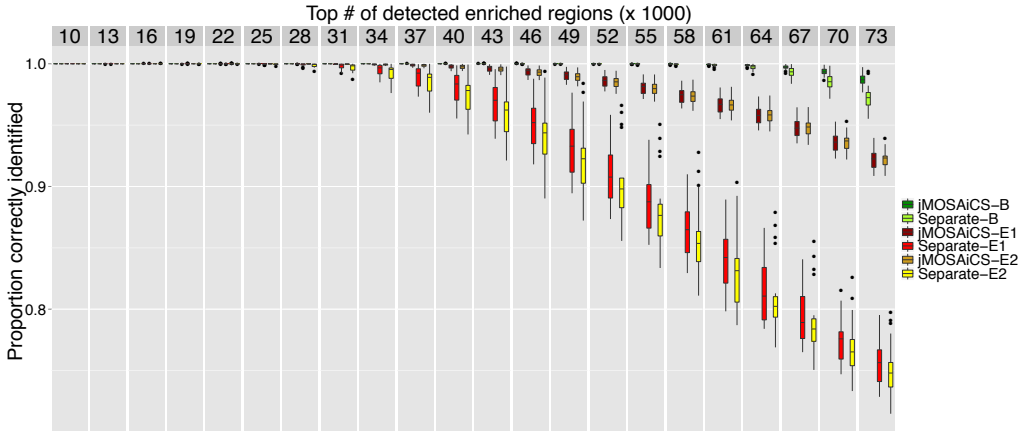


Figure S1. Density plots for experimental and sample simulated data.

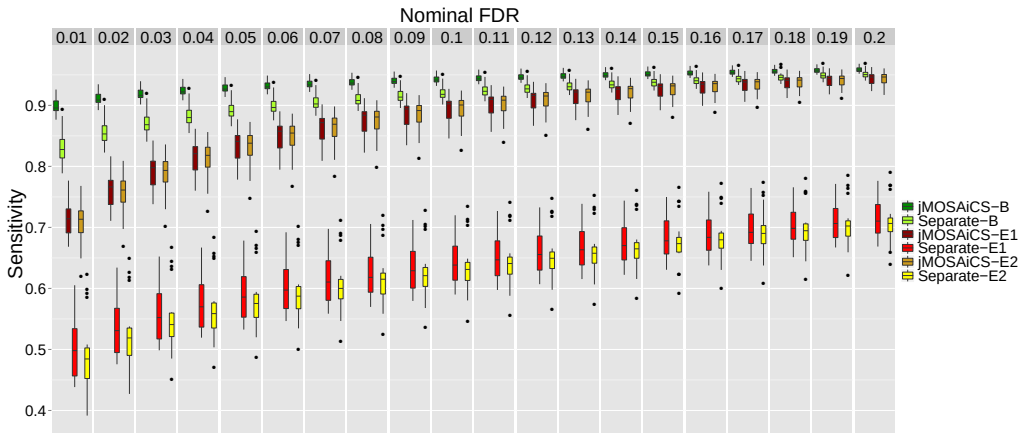
Supplementary figure for the STAT1 simulation with $D = 3$.

Supplementary Figure S2 reports the STAT1 simulation results for the setting with three ChIP-seq samples.

(a)



(b)



(c)

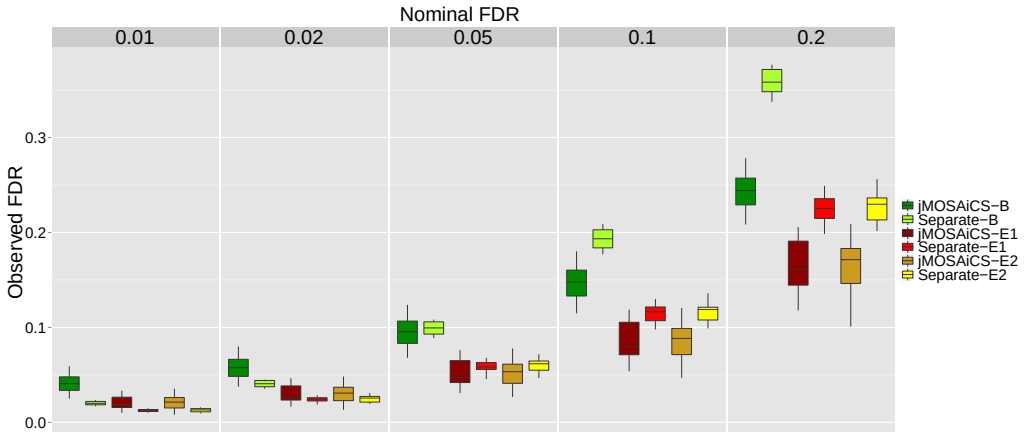
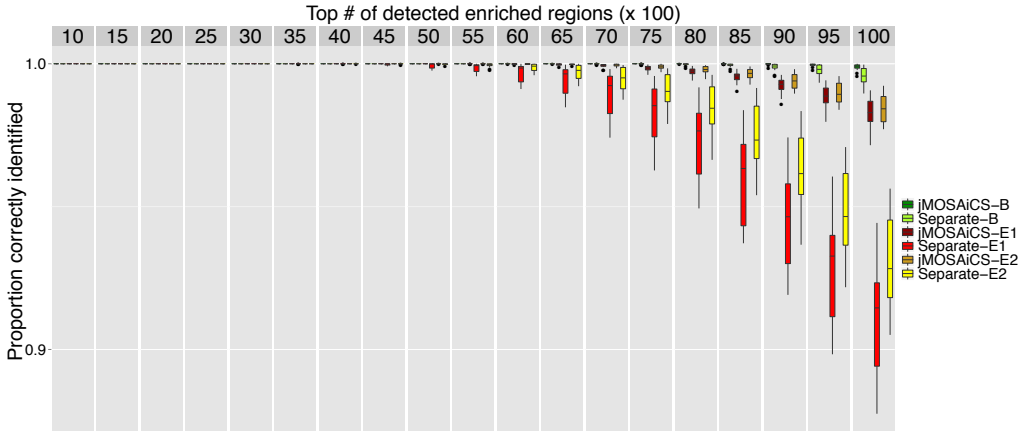


Figure S2: *Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from the STAT1 ChIP-seq experiment ($D = 3$).* 'jMOSAiCS-B', 'jMOSAiCS-E1', and 'jMOSAiCS-E2' represent results derived from posterior probability inferences of the B, E_1, and E_2 variables. 'Separate-B', 'Separate-E1', and 'Separate-E2' represent results derived from separate analysis of each dataset. Results for the third dataset based on E_3 are not depicted since they are similar to those of E_1 and E_2.

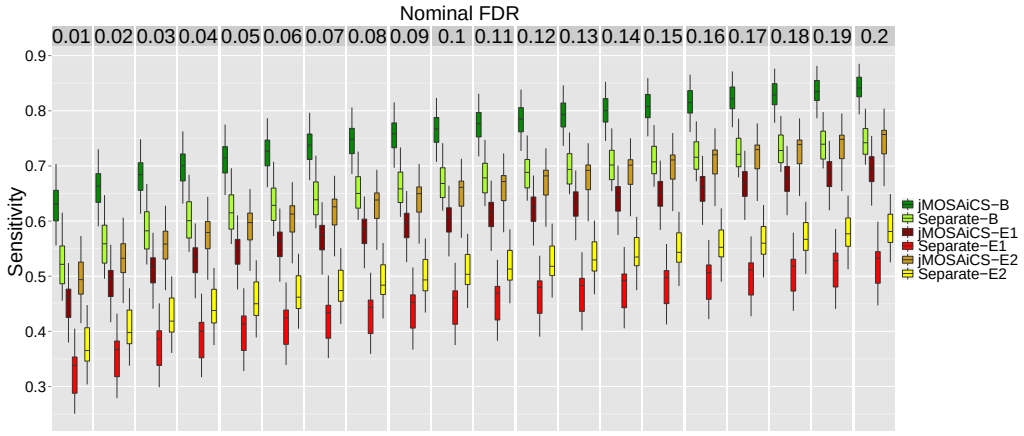
Supplementary figure for the H3K9me3 ChIP-seq simulation.

Supplementary Figure S3 reports accuracy, sensitivity, and FDR control results from simulation experiments with parameters matching to ChIP-seq data of H3K9me3 modification in peripheral blood mononuclear cells (PBMCs) from two individuals. One of the datasets, hence the data simulated from its parameters, has much lower signal strength. As a result, both the accuracy and sensitivity results captured by the E_1 variable for this sample is much lower compared to the other sample. Furthermore, the FDR control with the separate analysis is very conservative owing to the low signal strength.

(a)



(b)



(c)

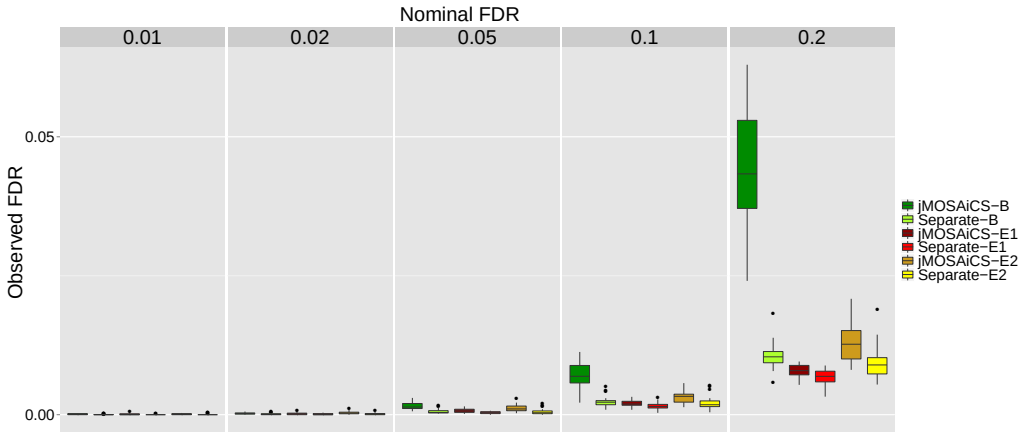


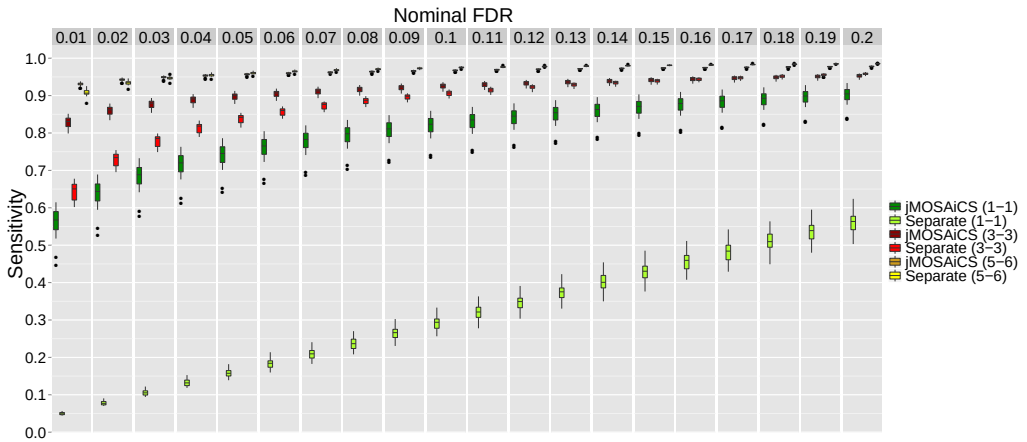
Figure S3: *Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from H3K9me3 ChIP-seq experiments in PBMCs of two individuals.* 'jMOSAiCS-B', 'jMOSAiCS-E1', and 'jMOSAiCS-E2' represent results derived from posterior probability inferences of the B, E_1, and E_2 variables. jMOSAiCS-E2 corresponds to the analysis of the lower signal dataset. 'Separate-B', 'Separate-E1' and 'Separate-E2' represent results derived from separate analysis of each dataset.

Supplementary figures for the MeCP2 ChIP-seq simulation.

Supplementary Figures S4, S5, and S6 report results for the setting that lowers sequencing depths of both of the MeCP2 replicates. When both replicates have one lane of data, the signal to noise ratios of the samples are very low and, as a result, jMOSAiCS has poor FDR control at the region-level as depicted by the box plots labelled with jMOSAiCS (1-1) in Supplementary Figures S4(c), S5(c), and S6(c)). However, at the dataset-specific region-level, jMOSAiCS improves both accuracy and sensitivity while providing better FDR control than the separate analysis.

(a)

(b)



(c)

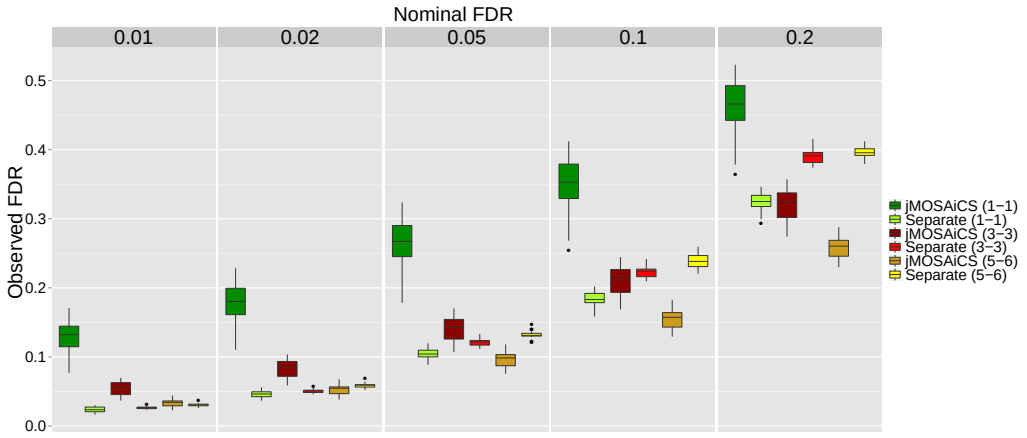


Figure S4: *Computational experiments comparing jMOSAiCS with a separate analysis approach on data simulated from MeCP2 ChIP-seq experiment.* Comparisons of region-level (B) results of jMOSAiCS and separate analysis. 'jMOSAiCS (x-y)' and 'Separate (x-y)' refer to jMOSAiCS and separate analysis of x lanes of replicate 1 with y lanes of replicate 2.

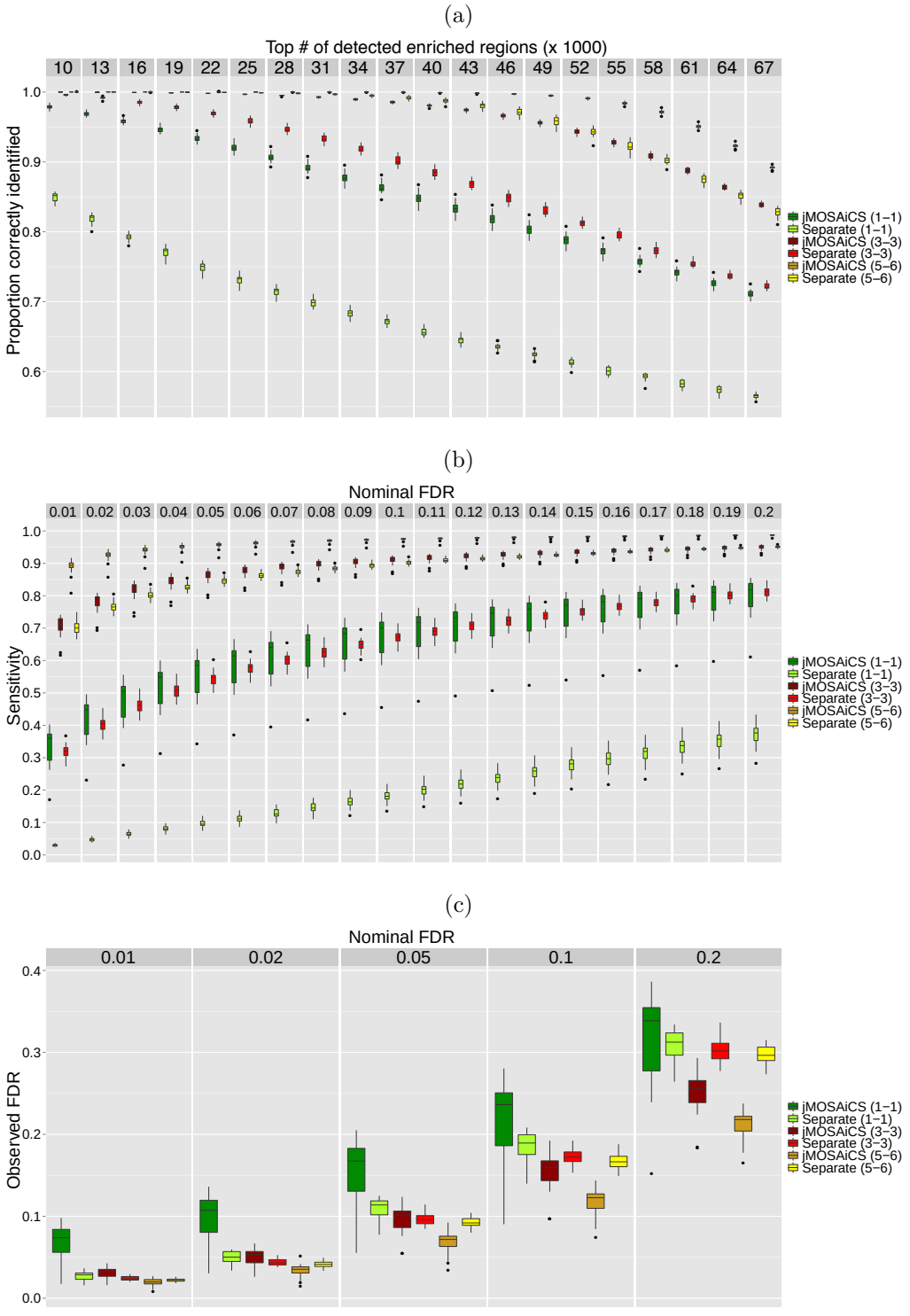
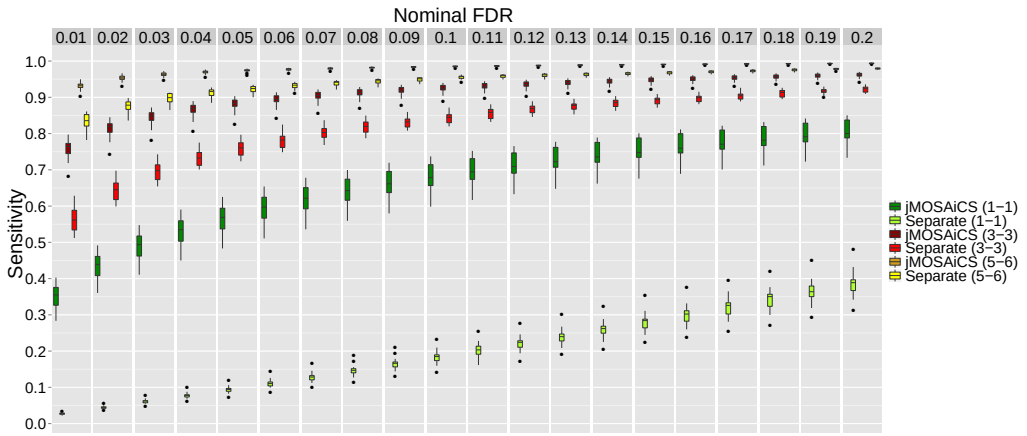


Figure S5: *Computational experiments comparing jMOSAiCS with a separate analysis approach on data simulated from the MeCP2 ChIP-seq experiment.* Comparison of dataset-specific region-level enrichment detection (E_1) results of jMOSAiCS and separate analysis on replicate 1. 'jMOSAiCS (x-y)' and 'Separate (x-y)' refer to jMOSAiCS and separate analysis of x lanes of replicate 1 with y lanes of replicate 2.

(a)

(b)



(c)

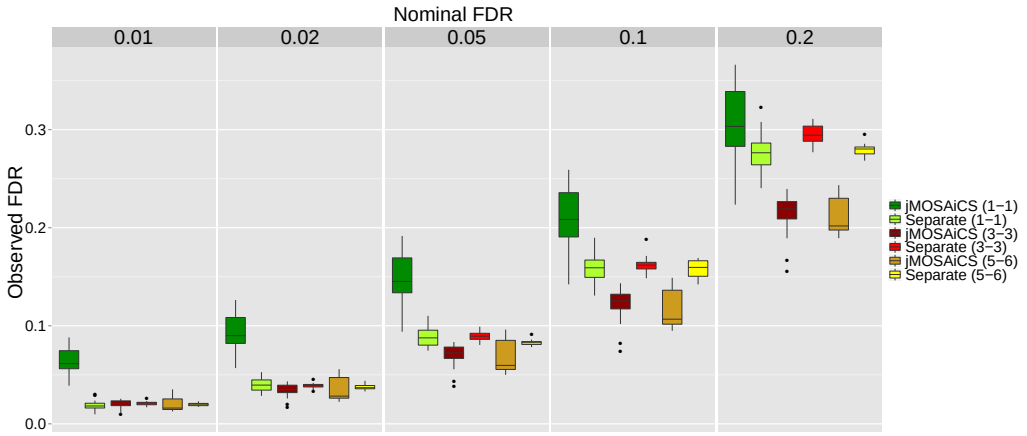


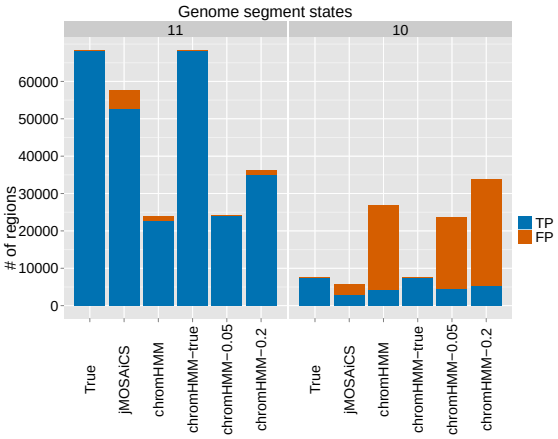
Figure S6: *Computational experiments comparing jMOSAiCS with a separate analysis approach on data simulated from the MeCP2 ChIP-seq experiment.* Comparison of dataset-specific region-level enrichment detection (E_2) results of jMOSAiCS and separate analysis on replicate 2. 'Joint (x-y)' and 'Separate (x-y)' refer to jMOSAiCS and separate analysis of x lanes of replicate 1 with y lanes of replicate 2.

Supplementary figures for comparisons with chromHMM

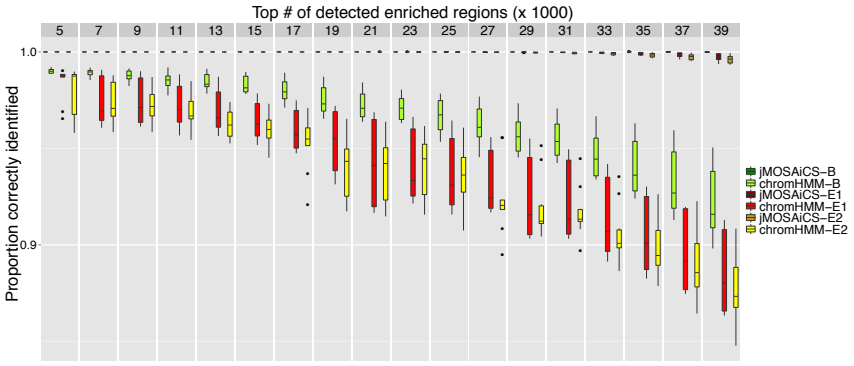
Supplementary Figure S7 compares jMOSAiCS with the 4-state chromHMM in terms of accuracy in setting SE2 and Supplementary Figures S8 and S9 present results for the 2-state and 4-state chromHMM in the first (SE1) and third simulation (SE3) settings. For the first simulation setting, the numbers of regions in states 10 and 01 are much smaller compared to state 11. As a result, the enrichment detection accuracies of the 2-state chromHMM and 4-state chromHMM do not differ significantly (Figures S8(b) vs. S8(c)). This agrees well with the many successful uses of chromHMM for identifying global patterns of epigenetic marks by approximating the size of the state space. jMOSAiCS exhibits best accuracy since it outperforms chromHMM among regions in state 11. In the third setting (SE3), when the signal strength is increased, the numbers of correctly identified regions in the 10 state increase for both the 4-state chromHMM and jMOSAiCS compared to the second simulation setting SE2 (Figure S7(a)). In addition, enrichment detection accuracy of 4-state chromHMM is much higher and less variable than that of the 2-state chromHMM (Figures S9(b) and S9(c)).

Figure S7: *Comparisons between jMOSAiCS and chromHMM based on data simulated from ChIP-seq experiment of STAT1 in HeLa3 cells.* Accuracy of enrichment detection at the region (B) and dataset-specific region (E_1 and E_2) levels by jMOSAiCS and 4-state chromHMM.

(a)



(b)



(c)

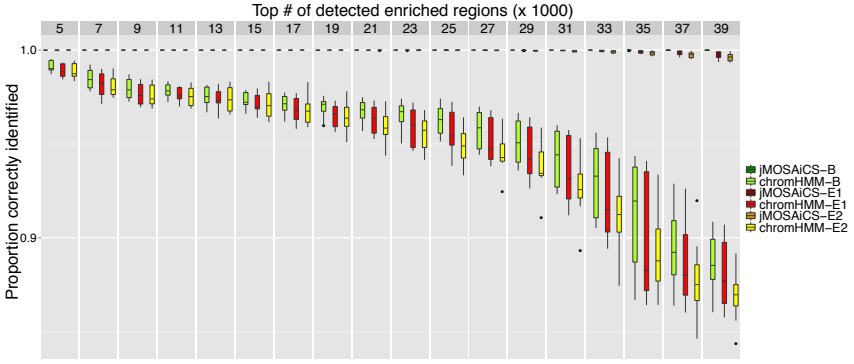
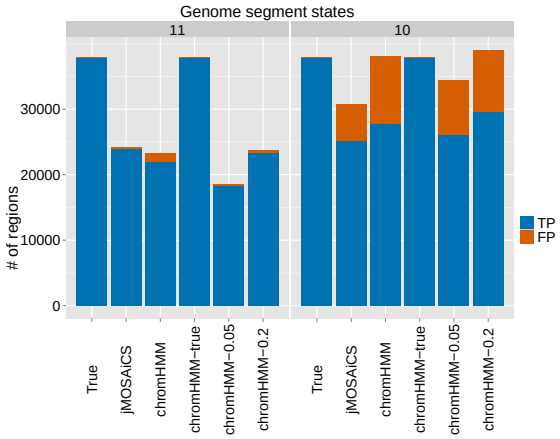
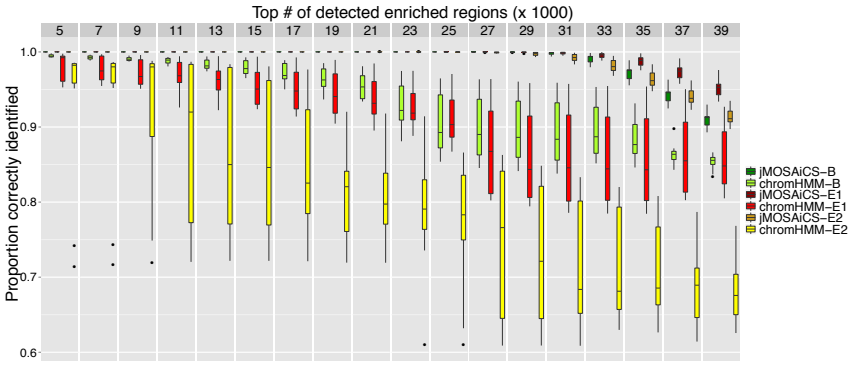


Figure S8: *Comparisons between jMOSAiCS and chromHMM based on data simulated from ChIP-seq experiment of STAT1 in HeLa3 cells (Setting: SE1).* (a) Identification of combinatorial patterns: '11': enriched in both samples; '10': enriched only in sample 1. 'True': number of enriched regions; 'Joint': results by jMOSAiCS; 'chromHMM': results by original chromHMM; 'chromHMM-true': chromHMM coupled with true binarization of the bins; 'chromHMM-0.05': chromHMM coupled with MOSAiCS labelling of the bins at an FDR of 0.05;'chromHMM-0.2': chromHMM coupled with MOSAiCS labelling of the bins at an FDR of 0:05. TP and FP denote true and false positives, respectively. (b)-(c) Accuracy of enrichment detection at the region (B) and dataset-specific region (E_1 and E_2) levels by jMOSAiCS, 2-state chromHMM (b), and 4-state chromHMM (c).

(a)



(b)



(c)

Figure S9: *Comparisons between jMOSAiCS and chromHMM based on data simulated from ChIP-seq experiment of STAT1 in HeLa3 cells (Setting: SE3).* (a) Identification of combinatorial patterns: '11': enriched in both samples; '10': enriched only in sample 1. 'True': number of enriched regions; 'Joint': results by jMOSAiCS; 'chromHMM': results by original chromHMM; 'chromHMM-true': chromHMM coupled with true binarization of the bins; 'chromHMM-0.05': chromHMM coupled with MOSAiCS labelling of the bins at a FDR of 0.05;'chromHMM-0.2': chromHMM coupled with MOSAiCS labelling of the bins at an FDR of 0:05. TP and FP denote true and false positives, respectively. (b)-(c) Accuracy of enrichment detection at the region (B) and dataset-specific region (E_1 and E_2) levels by jMOSAiCS, 2-state chromHMM (b), and 4-state chromHMM (c).

Parameter settings for the data-driven computational experiments

	b_1	c_1	b_2	c_2	τ_1	p_1^*	η_1	η_2
STAT1 (D=2, D=3)	0.3388	0.0373	1.3769	0.0077	0.1619	0.9826	0.9	0.9
H3K9me3 (D = 2, sample 1)	0.5000	0.0600	0.4700	0.0060	0.1900	0.995	0.9	0.9
H3K9me3 (D = 2, sample 2)	0.5800	0.0600	10.01	0.0200	0.1900	0.999	0.9	0.9
MeCP2 replicate 1 (1 lane)	0.0028	0.0013	0.6621	0.6588	0.2306	0.0303	0.9	0.9
MeCP2 replicate 1 (3 lanes)	1.3620	0.3127	0.0031	0.0004	0.2306	0.9905	0.9	0.9
MeCP2 replicate 1 (5 lanes)	0.0010	0.0003	1.8577	0.1760	0.2306	0.0226	0.9	0.9
MeCP2 replicate 2 (1 lane)	0.0023	0.0013	0.8255	0.6237	0.2306	0.0291	0.9	0.9
MeCP2 replicate 2 (3 lanes)	0.0009	0.0004	1.7181	0.2289	0.2306	0.0268	0.9	0.9
MeCP2 replicate 2 (6 lanes)	0.0009	0.0002	1.9469	0.1456	0.2306	0.0238	0.9	0.9

Table S1: *Parameters used for the computational experiments.* * mixing proportion for the components of the two component Negative Binomial signal distribution. The I genomic regions consisted of 5, 3, and 3 bins for the STAT1, H3K9me3, and MeCP2 simulations, respectively. Multiple samples of $D = 2$ and $D = 3$ settings in the STAT1 experiments are generated by adding small random perturbations to the parameters presented in the first row.

Quantitative ChIP results for the G1E-ER4+E2 cells

Mark	Atp6v1e	Elf1	Extl3	Cmas
H3K4me1	0.228	1.444	0.198	0.378
H3K4me3	0.075	0.254	0.082	0.293
K27me3	0.107	0.122	0.182	0.103
K9me3	0.103	0.142	0.103	0.086
PI	0.004	0.003	0.008	0.003

Table S2: Relative levels of the specific histone marks from quantitative ChIP analysis averaged over two independent biological replicates of beta-estradiol-induced G1E-ER-GATA-1 cells. Pre-immune denotes negative control.

Extl3 F	TCTCATTACAGGTGGTTGTGAGC
Extl3 R	GTGTTGGCTGGTGAGATGGCT
Elf1 F	GCCACCATGCCCGGC
Elf1 R	TTCACCTTTTCAGCTTTGAGG
Atp6v1e F	GAACTGAATGGACAAACCAGGG
Atp6v1e R	TCTTCTGCCCATACCTCACACCT
Cmas F	GGGAGGTGTGCATATAGAACA
Cmas R	CCTCCCAGCTCATCG

Table S3: Primer sets used in the quantitative ChIP assays.

Coverage plots for the loci with validated patterns

Coverage plots display the total number of reads mapping to each nucleotide separated by strand (forward: black, backward: red). Each read only contributes to the nucleotide that its $5'$ end maps to. Coordinates in Figures S10 to S13 are based on mouse genome version mm8. jMOSAiCS identified "1101" for G1E and "1111" for G1E-ER4+E2 as patterns of these GATA1 occupied loci. Control input values of these regions are much lower than the displayed ChIP values even after adjusting for the sequencing depths (data not shown). The dashed lines denote the boundaries of the GATA1 occupied loci.

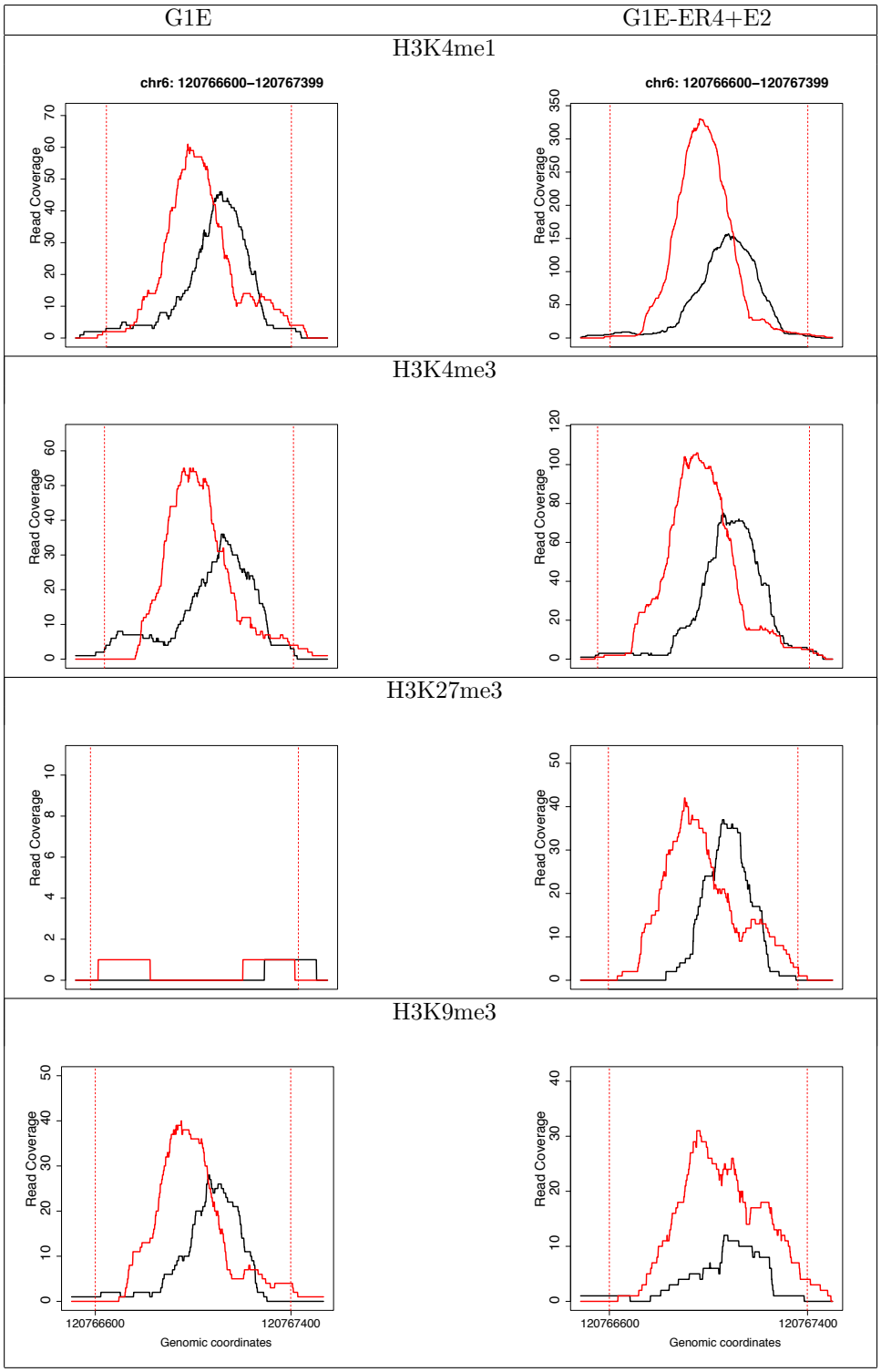


Figure S10: *Coverage plot for the Atp6v1e1 locus.* ChromHMM patterns: G1E: State 2 (”1000”); G1E-ER4+E2: State 2 (”1000”).

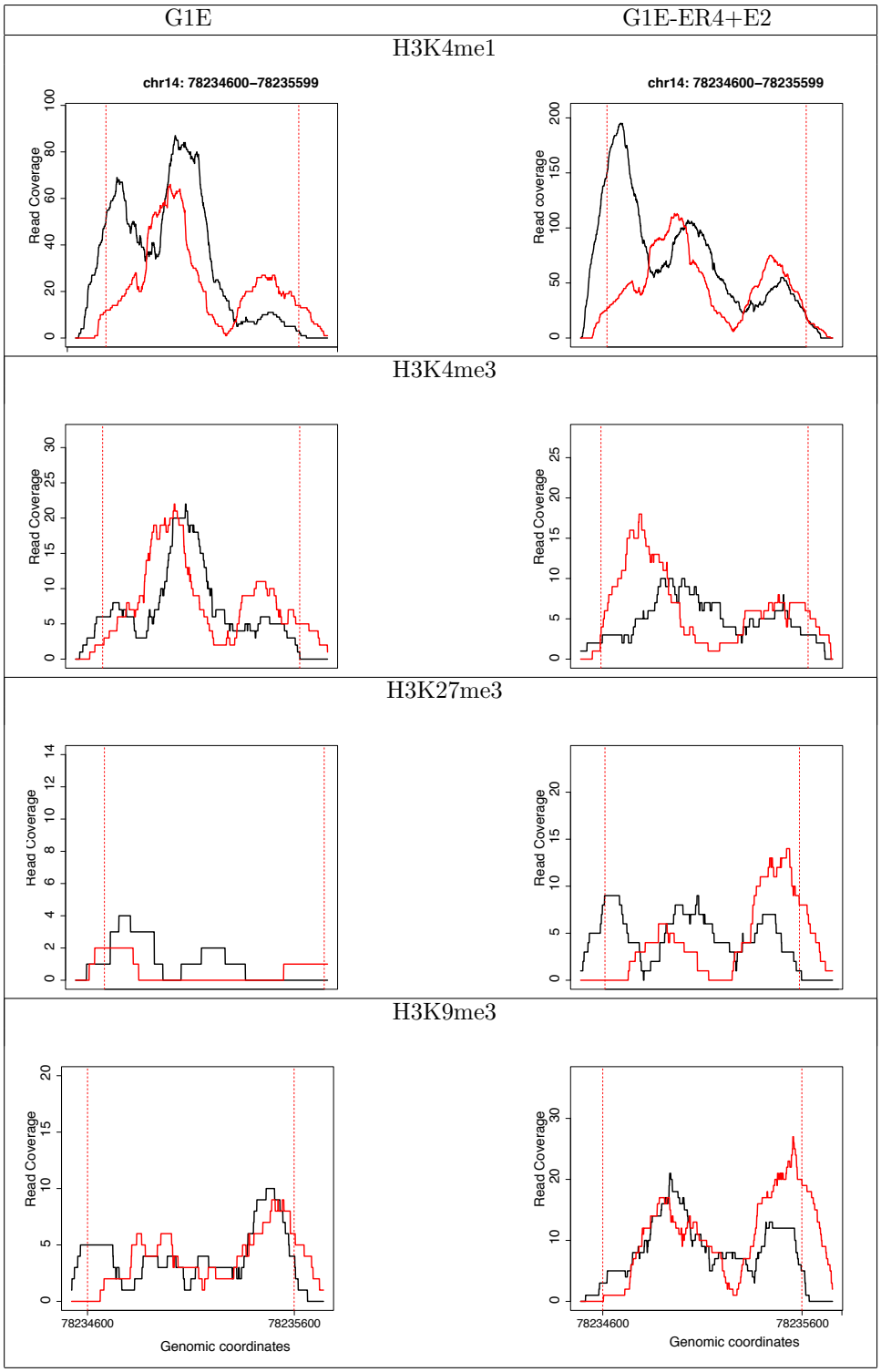


Figure S11: *Coverage plot for the Elf1 locus.* ChromHMM patterns: G1E: State 6 ("0000"); G1E-ER4+E2: State 6 ("0000").

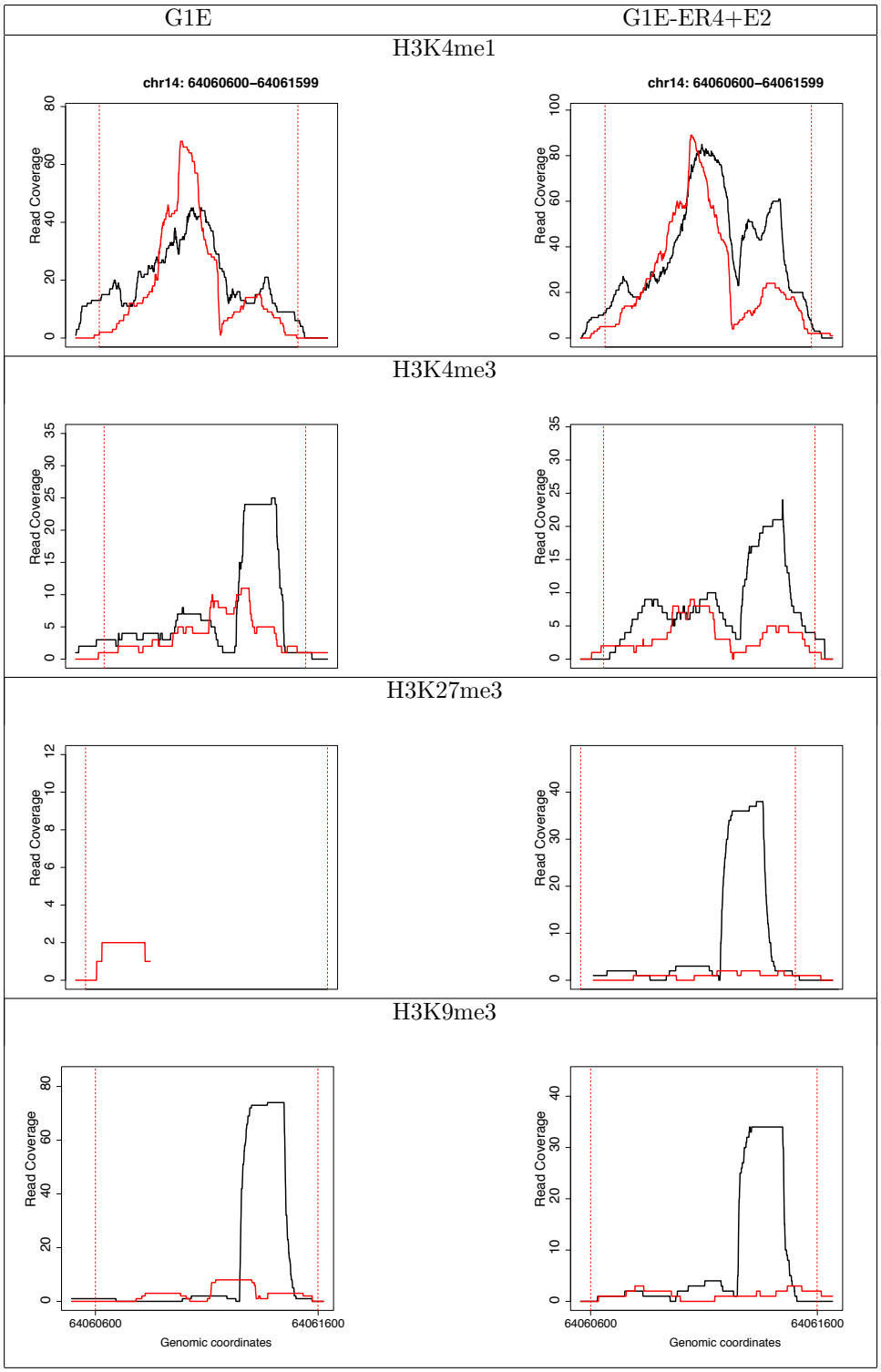


Figure S12: *Coverage plot for the Extl3 locus.* ChromHMM patterns: G1E: State 6 ("0000"); G1E-ER4+E2: State 4 ("0010").

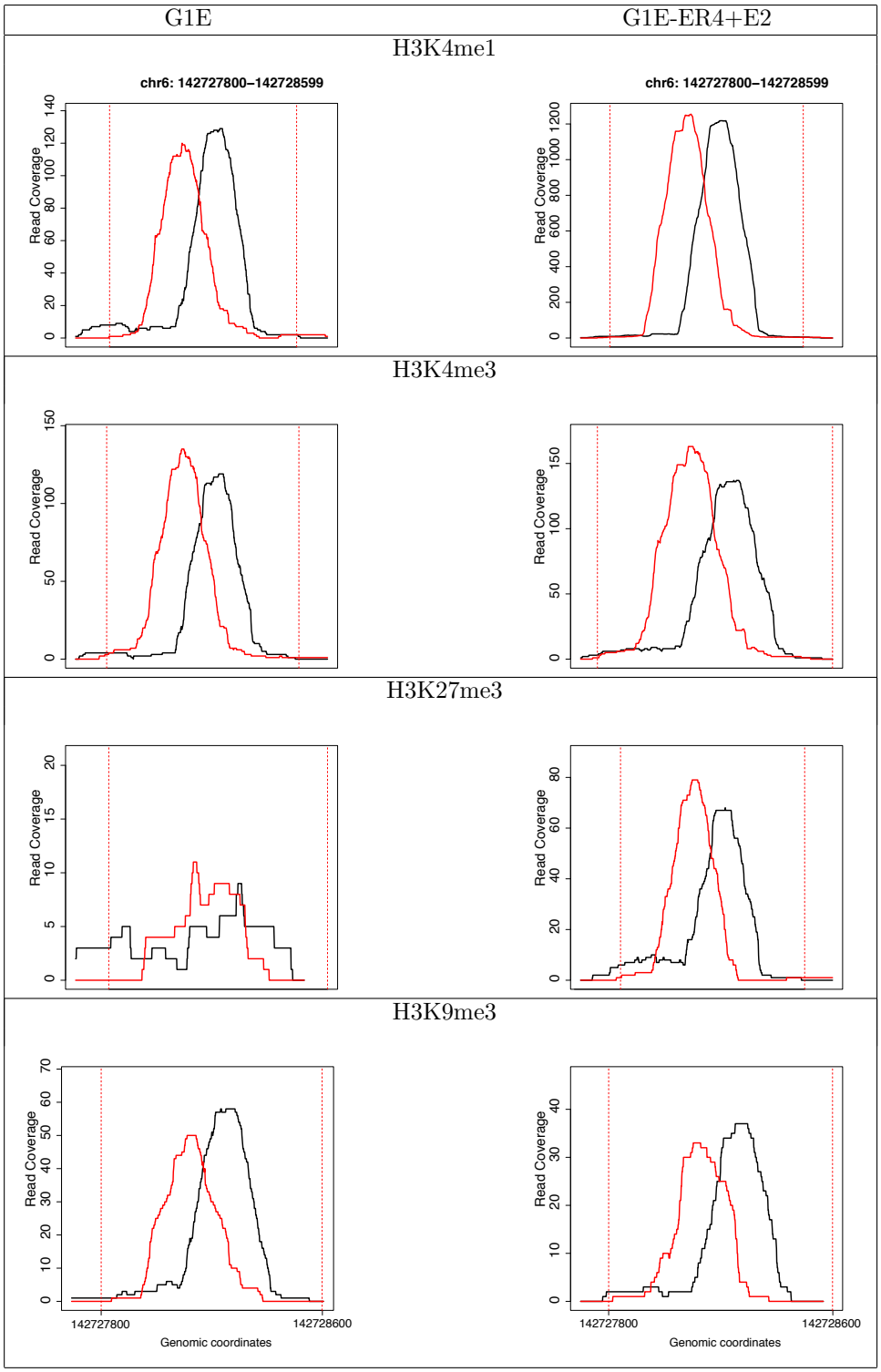


Figure S13: *Coverage plot for the Cmas locus.* ChromHMM patterns: G1E: State 2 ("1000"); G1E-ER4+E2: State 1 ("1100").