Genome Biology

METHOD Open Access

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*]

Abstract

The ChIP-seq technique enables genome-wide mapping of *in vivo* protein-DNA interactions and chromatin states. Current analytical approaches for ChIP-seq analysis are largely geared towards single-sample investigations, and have limited applicability in comparative settings that aim 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 its usefulness with a wide range of data-driven computational experiments and with a case study of histone modifications on GATA1-occupied segments during erythroid differentiation. jMOSAiCS is open source software and can be downloaded from Bioconductor [1].

Background

The advent of high-throughput next generation sequencing (NGS) technologies has 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 in biological variation, detailed comparisons of transcription factor (TF) binding and epigenomic marks between different tissues and 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, for example, [2] studied variation in binding of NF-κB and RNA polymerase II (Pol II) across ten individuals; [3] 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, for example, comparative analysis of 22 *Caenorhabditis elegans* TFs [4]; (iii) time course or multiple developmental stage ChIP-seq experiments, for example, Pol II binding at six developmental stages of *C. elegans* [4]; and (iv) comparative analysis of binding profiles of one or more TFs with Pol II or modifications of histone marks, for example, [5,6].

Although there are already more than 30 algorithms and methods for ChIP-seq analysis (reviewed in [7]), all of them are limited to single-sample analysis and lack the ability to simultaneously compare multiple ChIP samples. The small number of available multi-sample ChIP-seq analysis tools are either specific to ChIP-seq design (for example, [8] is specifically for identifying chromatin states from ChIP-seq of histone modifications; [9] focuses on gene-centric analysis), exploratory [10] or difficult to generalize to more than two samples [11-13] 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. A more recent genome segmentation algorithm, Segway [14], designed for multiple ChIP-seq datasets, utilizes a dynamic Bayesian network method and offers flexibility by enabling analysis at 1 bp resolution. However, the current Segway implementation requires specialized cluster management systems and is not readily available for running on standard desktops or computing clusters without a cluster management system.

* Correspondence: keles@stat.wisc.edu
[1]Department of Statistics, University of Wisconsin-Madison, 1220 Medical Sciences Center, 1300 University Avenue, Madison, WI 53706, USA
Full list of author information is available at the end of the article

We introduce jMOSAiCS (joint model-based one- and two-sample analysis and inference for ChIP-seq), which is 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 chose 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 [8]. 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 [6]. This analysis identified a cluster of GATA1-occupied loci exhibiting a pattern of enrichment that is different than that 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 [5,15-17], 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, that is, 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} is the set of possible patterns: 00 means not enriched in either of the samples; 10 enriched only in sample 1; 01 enriched only in sample 2 and 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 [7] 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 $(E_{i1}, ..., E_{iD})$. Our joint model has three layers, which are depicted in Figure 1. The first layer, called the 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, called the Y layer, concerns observed read count data for region i across D samples: $Y_i = (Y_{i1}, ..., Y_{iD})$, where $Y_{iD} = (Y_{id1}, \ldots, Y_{idL_i})$ 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 the jth bin of the ith region in the dth sample. We assume that the counts from different samples are independent conditional on the enrichment pattern:

$$Y_{id} \perp Y_{id'} | E_i, \forall d, d' = 1, \ldots, D,$$

and hence

$$\Pr(Y_i) = \sum_{r=1}^{R} \left[\prod_{d=1}^{D} \Pr(Y_{id} | E_i = r) \right] \Pr(E_i = r),$$

where $r = 1, ..., R$ 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, called the Z layer, concerns Z_{idj}, which we define as the bin-specific enrichment variable. If the jth bin in the ith region is enriched in dataset d, then $Z_{idj} = 1$ and is 0 otherwise. We assume that $Z_{idj}, j = 1, \ldots, L_i, \forall d, i$ are independent and conditional on the region-specific enrichment indicator E_{id} and hence:

$$\Pr(Z_{id1}, \ldots, Z_{idL_i} | E_{id} = r_d) = \prod_{j=1}^{L_i} \Pr(Z_{idj} | E_{id} = r_d)$$

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 [18], 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 $E_{i1}, ..., E_{iD}$ are conditionally independent given B_i. Let $\Pr(B_i = 1) = \tau_1$,

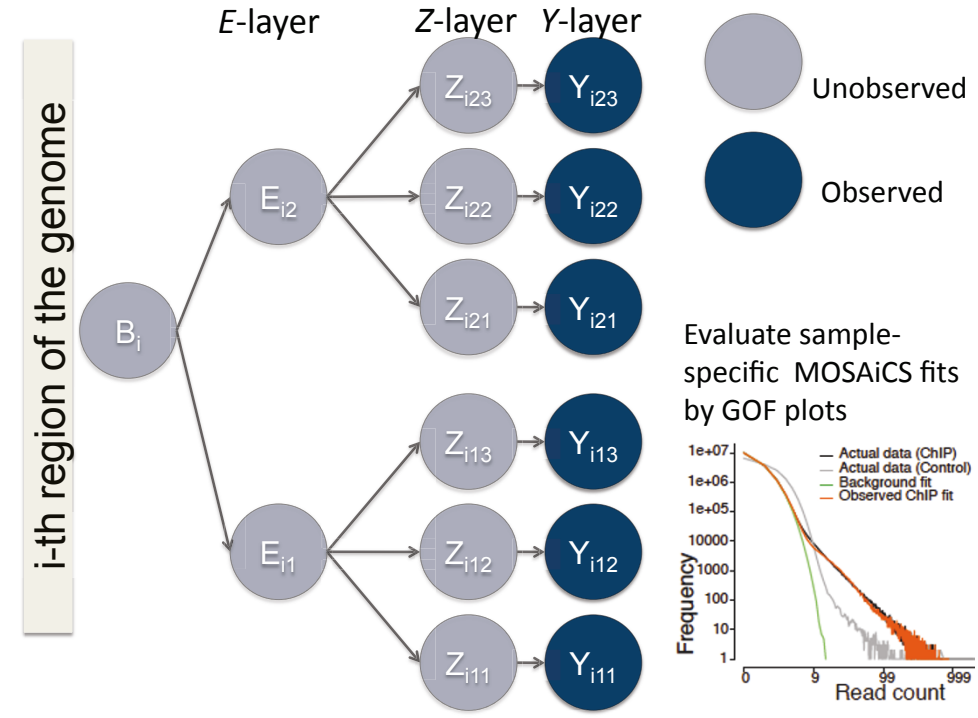

Figure 1 Pictorial depiction of the jMOSAiCS model for a region across two ChIP-seq datasets. Region *i* consists of three bins. The *B* variable governs whether or not the region is enriched in any of the two samples. The *E* variables denote sample-specific enrichments and are conditionally independent given the *B* variable. The *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. The 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 [19] on individual samples and evaluated by the goodness-of-fit (GOF) plots. ChIP: chromatin immunoprecipitation; jMOSAiCS: joint model-based one- and two-sample analysis and inference for ChIP-seq; MOSAiCS: model-based analysis and inference for ChIP-seq data

$\Pr(E_{id} = 1 \mid B_i = 1) = \eta_d$, and $\Pr(E_{id} = 1 \mid B_i = 0) = 0$, that is, 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 $(E_{i1}, ..., E_{iD})$ is given by:

$$\Pr(E_{i1} = r_1, \ldots, E_{iD} = r_D) = \tau_1 \prod_{d=1}^{D} \eta_d^{r_d}(1-\eta_d)^{1-r_d} + (1-\tau_1)\mathrm{I}(r_1 = 0, \ldots, r_D = 0).$$

For the *Y* layer, we adopt the model-based approach of MOSAiCS [19] 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 $(a, a/\mu_{idj})$ are the background read counts. Its mean μ_{idj} is parameterized as $\log \mu_{idj} = \beta_0 + \beta_1 X_{idj}^c$, where X_{idj} are 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 parameterization of the bin-level counts that also depends on mappability and guanine-cytosine (GC) content. For the enriched bins, $Y_{idj} \mid Z_{idj} = 1 \sim N_{idj} + S_{idj}$, where S_{idj} is the signal due to enrichment, that is, protein binding or epigenomic marker modification. The signal S_{idj} is modeled 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) [20] by MOSAiCS. For model fitting, we utilize the fact that MOSAiCS provides fast and accurate estimates of the dataset-specific background and signal distributions. Therefore, as a 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, ..., 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 Bioconductor [1]. Additional Files 1 and 2 provide a freeze of the R package and its vignette. These are included in the manuscript for archival

purposes only. We recommend that users download the most recent version of the software from Bioconductor [1].

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 [19]. We simulated multiple ChIP-seq datasets 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 Additional file 3, Figure S1, and indicate that the simulated data mimics the actual data well. In what follows, we first compared jMOSAiCS with a commonly practiced separate analysis scheme where each ChIP-seq dataset is analyzed individually and the enrichment patterns are generated by *post hoc* analysis. Then, we compared jMOSAiCS to chromHMM [8], which is currently the state-of-the-art approach for discovering combinatorial patterns of chromatin states from multiple ChIP-seq data.

jMOSAiCS improves on a 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 Additional file 3, 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 cutoff, other types of type-I error rate control or the investigators may choose to consider a 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 the true set of enrichments that are detected as a function of the nominal false discovery rate (the total number of detected true enrichments identified at different FDR cutoffs divided by the total number of true enrichments are reported); (iii) false discovery rate control by plotting the 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 2 summarizes these results for the $D = 2$ setting across 20 simulation runs (results for $D = 3$ are available in Additional file 3, Figure S2). This setting, on average, has 85,000 enriched regions, that is, regions with $B = 1$. Figure 2(a), which displays the 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) in 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 (≥58,000), 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 2(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 a larger number of enriched regions and captures a 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

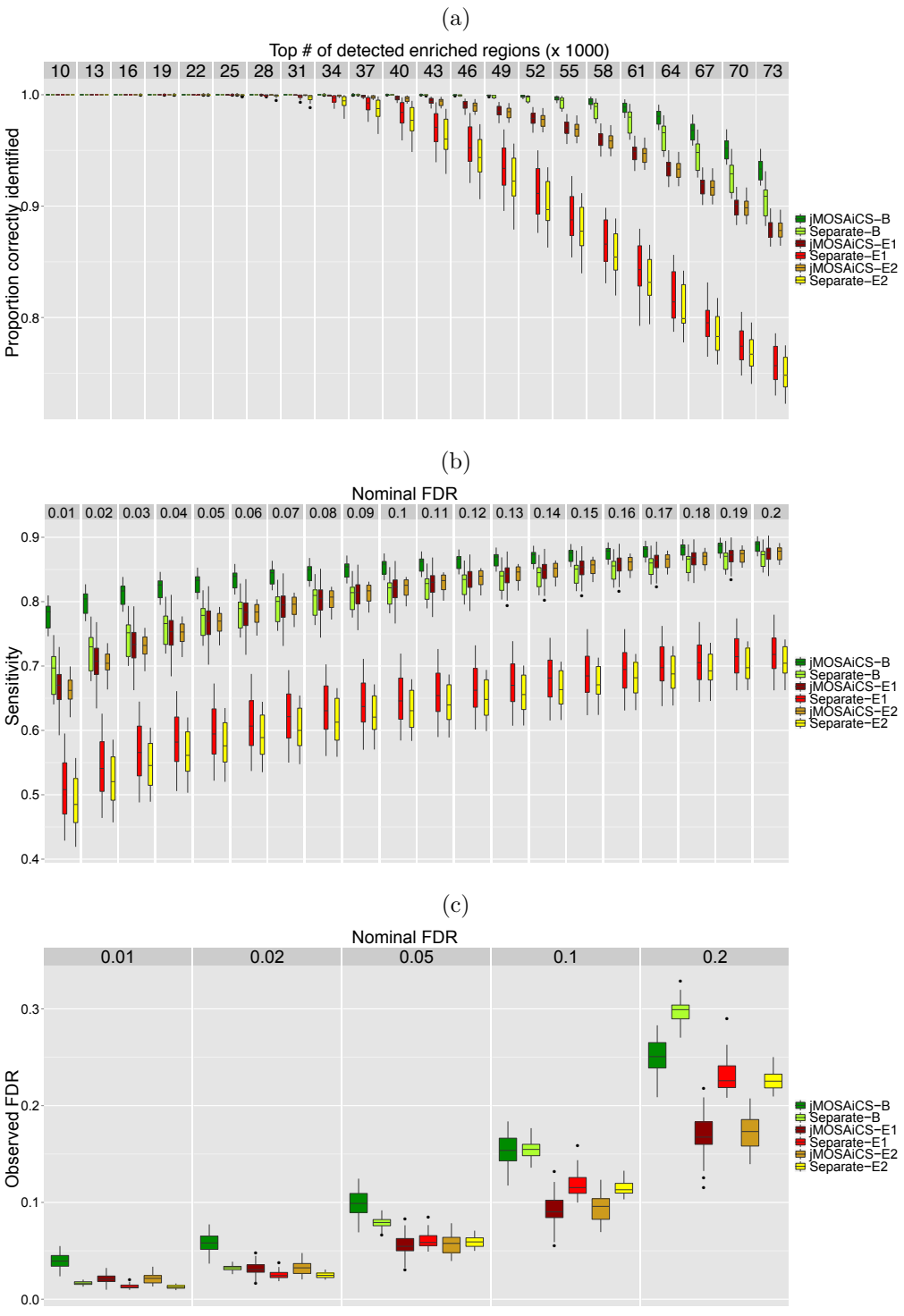

Figure 2 Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from the STAT1 ChIP-seq experiment. jMOSAiCS-B, jMOSAiCS-E1, and jMOSAiCS-E2 are results derived from posterior probability inferences of the B, E_1, and E_2 variables. Separate-B, Separate-E1, and Separate-E2 are results derived from separate analysis of each dataset. **(a)** Proportion of top ranking enriched regions that are true positives. **(b)** Sensitivity by nominal FDR. **(c)** Observed FDR by nominal FDR. ChIP: chromatin immunoprecipitation; FDR: false discovery rate; jMOSAiCS: joint model-based one- and two-sample analysis and inference for ChIP-seq

cutoff, jMOSAiCS identifies more true enrichments than the separate analysis. Next, we show how well the FDR is controlled by the two approaches in Figure 2(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 overestimation of FDR by jMOSAiCS; however, this still presents 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 Additional file 3, 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 an assessment of variability. Prior research suggests that non-specific biases such as GC content can vary significantly between biological replicates [19,22]. 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 for a computational experiment based on MeCP2 binding in mice. The data consisted of two biological replicates with five and six lanes of sequencing reads, respectively. The number of usable reads within a lane varied between 6.8 and 19.7 million. MOSAiCS provided adequate fits on each dataset and the simulation parameters were set according to estimates from the MOSAiCS fits. Details on the parameter settings are available in Additional file 3, Table S1. Within this simulation, we varied the sequencing depth of one of the replicates (replicate 2) at one, three, and six lanes while keeping the other replicate at five 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 3, 4, and 5 summarize the results for these experiments.

Figure 3(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 5(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 3(a), 4(a), and 5(a)). The differences in the sensitivities of the two approaches vary significantly with the number of lanes of replicate 2 (Figures 3(b), 4(b), and 5(b)). Overall, jMOSAiCS consistently identifies 10% to 15% more of the true enrichments when replicate 2 has lower depth. In Figure 4, 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 5(b)). In this setting, jMOSAiCS identifies 50% more of the true enrichment regions across all the nominal FDR levels. In Figures 3(c), 4(c), and 5(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 a fixed number of top ranking regions are considered (Figures 3(a) and 5(a)).

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

Comparison with chromHMM

chromHMM [8] 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 bp 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 two-dimensional 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

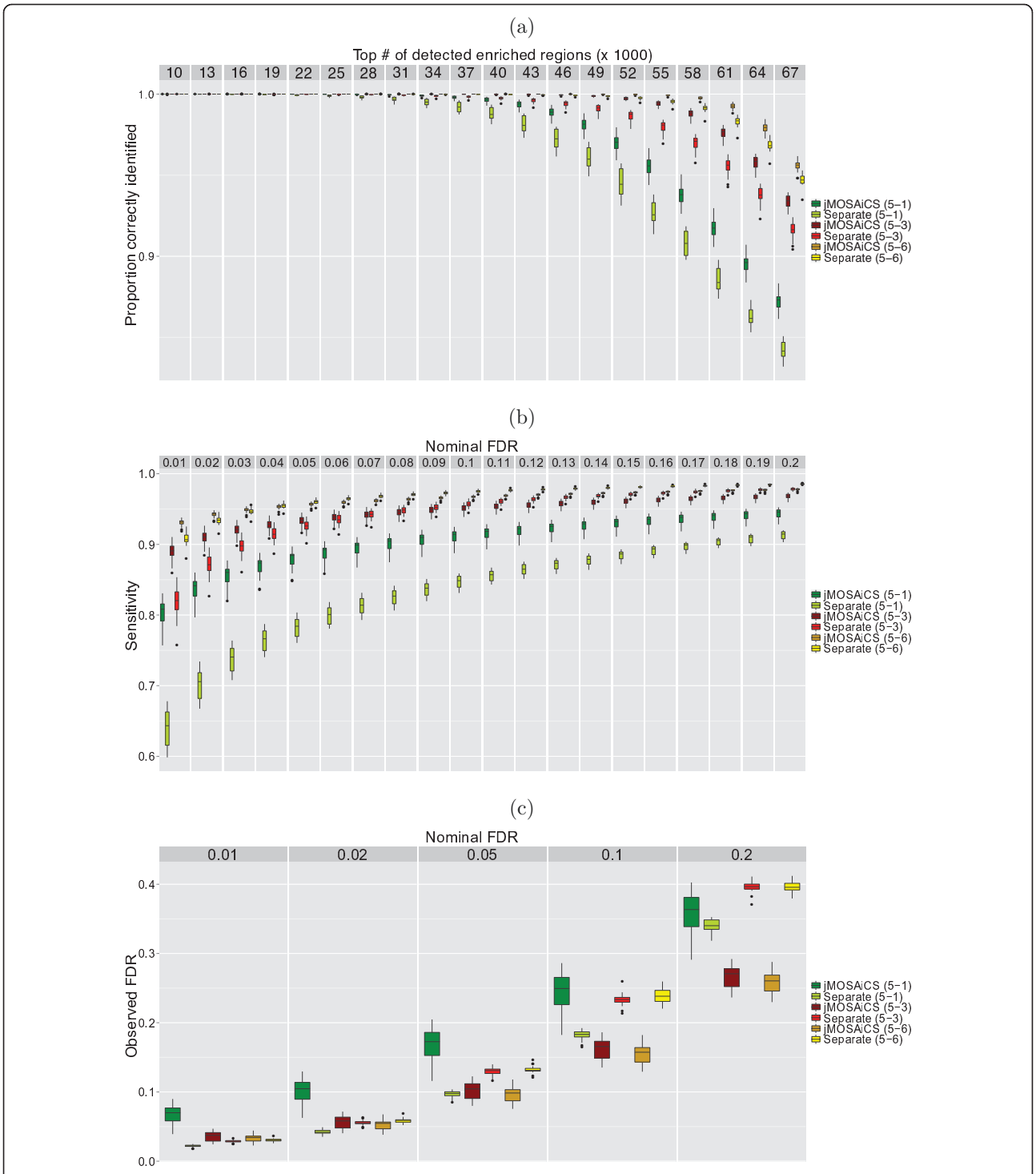

Figure 3 Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from the MeCP2 ChIP-seq experiment. Comparisons of region level (*B*) results of jMOSAiCS and separate analysis. jMOSAiCS (*x-y*) and Separate (*x-y*) refer to jMOSAiCS and separate analysis of *x* lanes of replicate 1 with *y* lanes of replicate 2. **(a)** Proportion of top ranking enriched regions that are true positives. **(b)** Sensitivity by nominal FDR. **(c)** Observed FDR by nominal FDR. ChIP: chromatin immunoprecipitation; FDR: false discovery rate; jMOSAiCS: joint model-based one- and two-sample analysis and inference for ChIP-seq

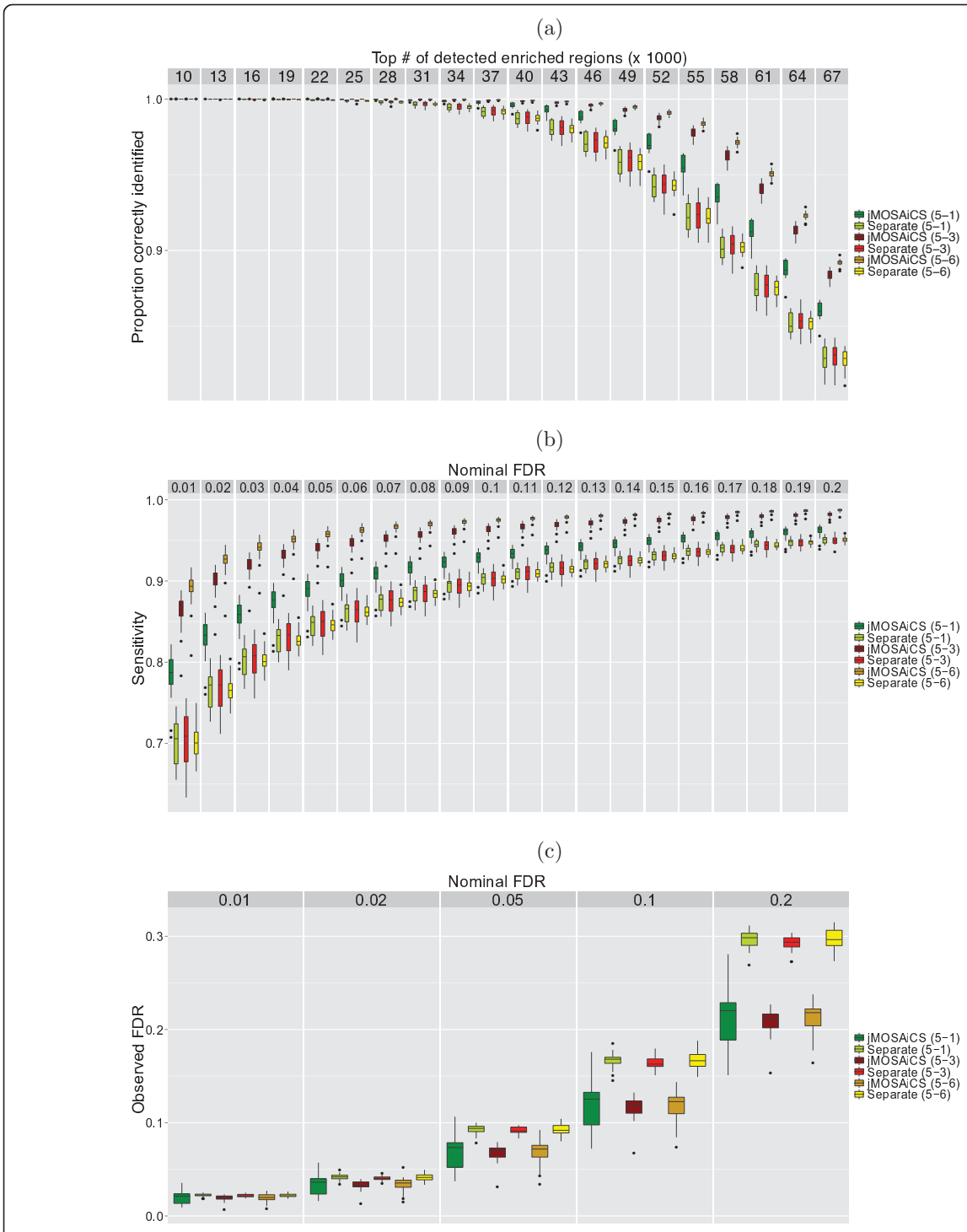

Figure 4 Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from the MeCP2 ChIP-seq experiment. Comparison of dataset-specific region-level enrichment detection (E_1) by jMOSAiCS and separate analysis on replicate 1. jMOSAiCS (*x-y*) and Separate (*x-y*) refer to jMOSAiCS and separate analysis of *x* lanes of replicate 1 with *y* lanes of replicate 2. **(a)** Proportion of top ranking enriched regions that are true positives. **(b)** Sensitivity by nominal FDR. **(c)** Observed FDR by nominal FDR. ChIP: chromatin immunoprecipitation; FDR: false discovery rate; jMOSAiCS: joint model-based one- and two-sample analysis and inference for ChIP-seq

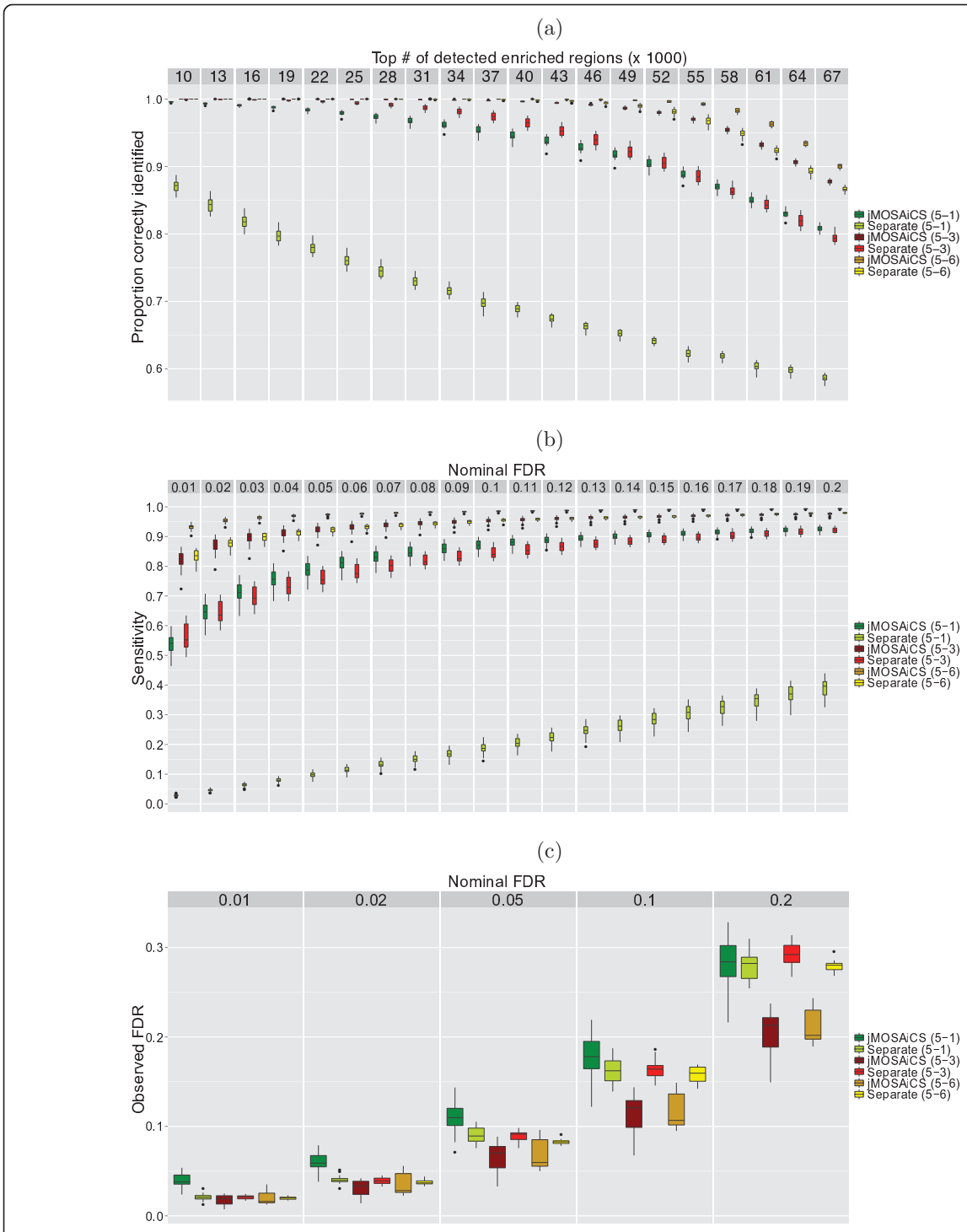

Figure 5 Computational experiments comparing jMOSAiCS with the separate analysis approach on data simulated from MeCP2 ChIP-seq data. Comparison of dataset-specific region-level enrichment detection (E_2) by jMOSAiCS and separate analysis on replicate 2 for which the number of data lanes varies. jMOSAiCS (*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)** Proportion of top ranking enriched regions that are true positives. **(b)** Sensitivity by nominal FDR. **(c)** Observed FDR by nominal FDR. ChIP: chromatin immunoprecipitation; FDR: false discovery rate; jMOSAiCS: joint model-based one- and two-sample analysis and inference for ChIP-seq

of histone data. The specific simulation settings are as follows:

SE1: Same as the STAT1 simulation described in the earlier section.
SE2: η_2 lowered from 0.9 to 0.5 to increase the number of regions with ten patterns.
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 a 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 6. Figure 6(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

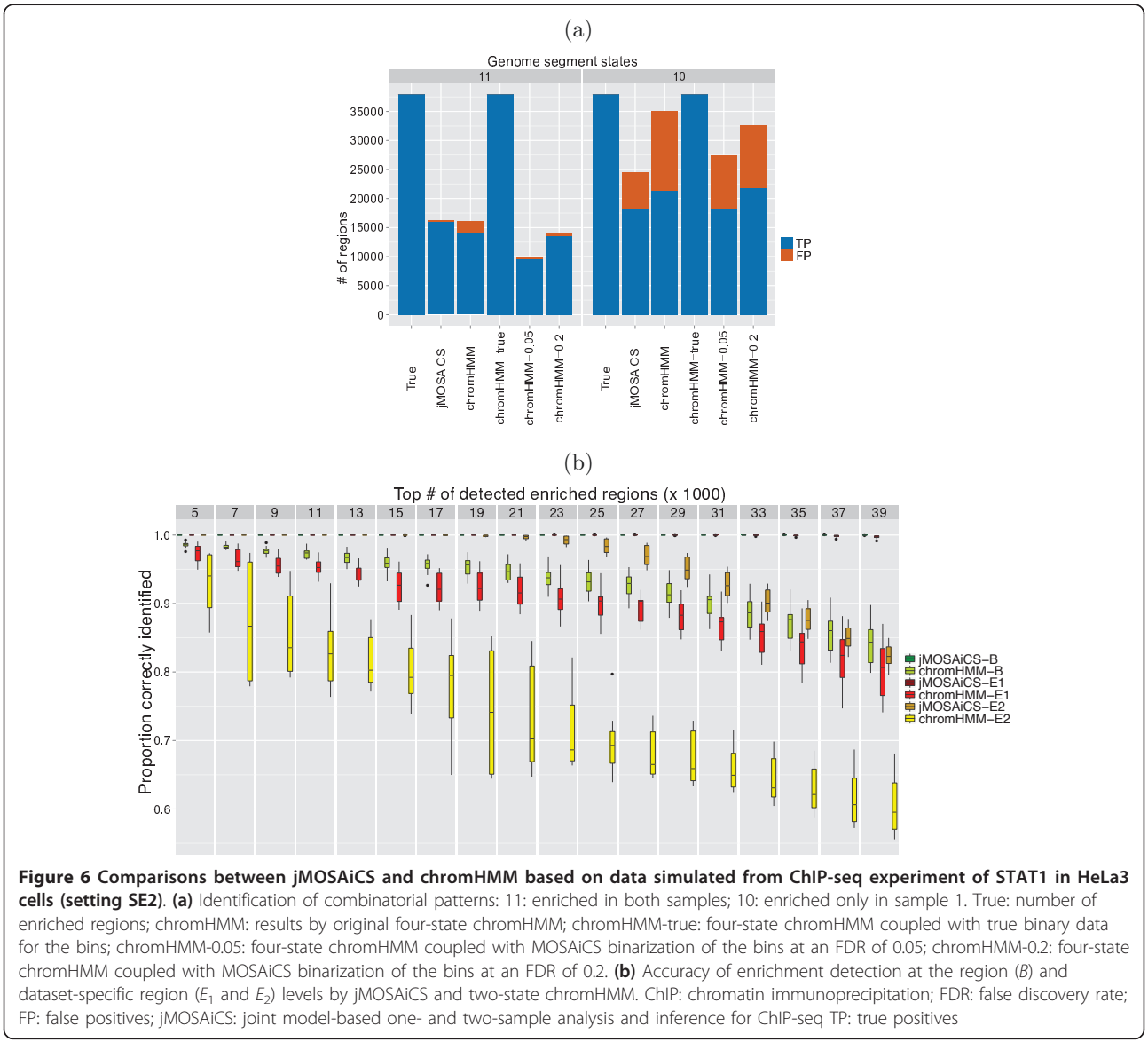

Figure 6 Comparisons between jMOSAiCS and chromHMM based on data simulated from ChIP-seq experiment of STAT1 in HeLa3 cells (setting SE2). **(a)** Identification of combinatorial patterns: 11: enriched in both samples; 10: enriched only in sample 1. True: number of enriched regions; chromHMM: results by original four-state chromHMM; chromHMM-true: four-state chromHMM coupled with true binary data for the bins; chromHMM-0.05: four-state chromHMM coupled with MOSAiCS binarization of the bins at an FDR of 0.05; chromHMM-0.2: four-state chromHMM coupled with MOSAiCS binarization of the bins at an FDR of 0.2. **(b)** Accuracy of enrichment detection at the region (B) and dataset-specific region (E_1 and E_2) levels by jMOSAiCS and two-state chromHMM. ChIP: chromatin immunoprecipitation; FDR: false discovery rate; FP: false positives; jMOSAiCS: joint model-based one- and two-sample analysis and inference for ChIP-seq TP: true positives

misannotated by chromHMM. Overall, these results illustrate that jMOSAiCS outperforms four-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 the original four-state chromHMM but identified fewer correct regions in the 11 state. Figure 6(b) provides detailed comparison of jMOSAiCS with the two-state chromHMM where chromHMM approximates the full state space of dimension 4 by only two states. A similar comparison between jMOSAiCS and four-state chromHMM is provided in Additional file 3, Figure S7. We observe that approximating the state space of dimension of 4 by two dimensions leads to significant loss in accuracy for chromHMM. At the individual dataset level, the difference in accuracy between two-state chromHMM and jMOSAiCS can be as large as 20% (comparing jMOSAiCS-E2 with chromHMM-E2 in Figure 6(b)). The results for simulation settings SE1 and SE3 are similar and provided as Additional file 3, Figures S8 and S9.

Scalability with large numbers of datasets

We evaluated how well jMOSAiCS scales up to large numbers of datasets by extending our simulation setting SE3 with $5b_1$ and $5b_2$. We generated $D = 20$ datasets and assigned each region of size 250 bp to one of the 76 states out of 2^{20} possible states, including the 0...0, non-enrichment state. The number of regions assigned to each individual enrichment state ranged between 1,048 and 1,212 and, on average, each state had 1,129 assigned regions. These experiments revealed that because the dataset-specific background and signal read distributions are estimated separately in the jMOSAiCS framework, it scales up easily to large numbers of datasets. Thus, for $D = 20$ datasets, jMOSAiCS runs in 2 hours on a 64-bit machine with an Intel Xeon 3.0 GHz processor after the background and signal read distributions are obtained for each dataset. Dataset-specific estimation with MOSAiCS requires 30 minutes to an hour; however, since each dataset can be handled separately, this process can be run in parallel using multiple CPUs.

Although jMOSAiCS can generate any number of states for large D, it is important to summarize key states for any given collection of datasets. One key advantage of jMOSAiCS is that the model fitting can be carried about without *a priori* setting the maximum number of allowed states. After the model is fit, each region is initially assigned to the state for which it has the largest posterior probability. However, if a much smaller number of states is desired, we consider the top K states with the largest number of region assignments and reassign each region to these K states. From the perspective of jMOSAiCS, summarizing the set of possible states with K states is analogous to choosing the number of clusters in a clustering problem. We have implemented a penalized average silhouette based criterion [23], which is widely used in clustering and seems to work well for our large D simulations (Additional file 3, Figure S10).

We next evaluated the accuracy and sensitivity for both jMOSAiCS and chromHMM by varying the maximum number of allowed states, K_{max}, as 30, 76, and 100 (Figures 7(a) and 7(b)). Overall, jMOSAiCS has very good accuracy and sensitivity when the number of states is chosen optimally or overestimated. However, when the number of states is grossly underestimated, the accuracy is comparable to those of large numbers of states for the top set of detected enriched regions (approximately 35%) and then it rapidly deteriorates since the small number of states simply does not capture the state of many regions. In order to evaluate

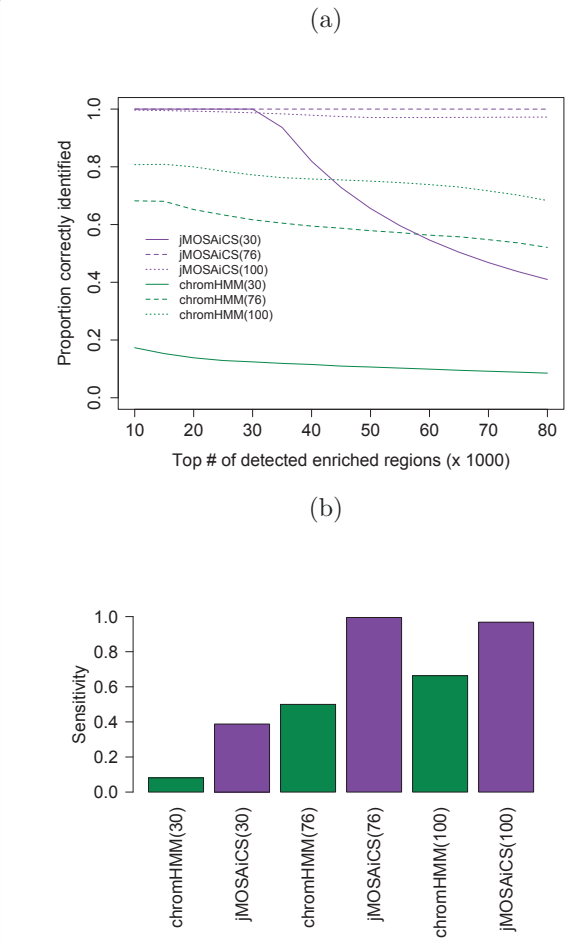

Figure 7 Computational experiments for evaluating scalability of jMOSAiCS to large numbers of datasets using data simulated from ChIP-seq experiment STAT1 in HeLa3 cells (extension of setting SE3). **(a)** Accuracy of enrichment detection at the combinatorial pattern (state) level for different numbers of states. **(b)** Sensitivity at varying numbers of states.

chromHMM under similar conditions, we performed chromHMM analysis by requiring 30, 76, and 100 states and thresholded the resulting emission probabilities to generate distinct states. This resulted in a total of 27, 63, and 85 distinct states.

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 G1E and G1E-ER4+E2 cells [6]. 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 [6]. The original analysis by [6] 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 a $2^4 = 16$ dimensional state space with only six states. Our jMOSAiCS application explored the full state space and, in addition to the six states identified by chromHMM, identified five more states to which a significant number of GATA1-occupied segments were assigned. Figure 8(a) enumerates the state space for jMOSAiCS and Figure 8(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 8(c), we provide normalized read data for the 311 GATA1-occupied peaks (with width less than 1,400 bp out of a total of 366) identified to switch from state 1101 in G1E to state 1111 in G1E-ER4+E2. We note that the chromHMM output does not include the 1101 or the 1111 pattern and distributes these loci over the six 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 location and identified that a large subset of them (48%) map to the immediate 5' or 3' end, or within introns of known genes. We studied expression profiling data from GATA1-null erythroid 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 [24] 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, [25] 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 Leukemia-1 (TAL1) gene [26,27]. 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 (Additional file 3, Table S2). We provide detailed read coverage plots of these regions in Additional file 3, Figures S11 to S14 along with their chromHMM annotations to further support their jMOSAiCS annotation.

In addition to the above direct comparison of jMOSAiCS analysis with the results of [6] for a six-state chromHMM, we repeated the chromHMM analysis by requiring 16 states. We cross-tabulated the numbers of GATA1-occupied segments assigned to each of the 16 states by the two methods in Additional file 3, Figures S15(a) and (b) for the G1E and G1E-ER4+E2 cell lines, respectively. Overall, 59% of regions are assigned to the same chromatin state by both methods in both cell lines (G1E: 6774/ 11485 and G1E-ER4+E2: 6752/11485). The largest discordances between the two methods are due to the modification of the H4K3me3 mark. Most of the regions that are identified as unenriched for the H4K3me3 in chromHMM are identified as enriched by jMOSAiCS. In order to quantify this further, we marginally tabulated the regions according to their enrichment classification by both chromHMM and jMOSAiCS (Additional file 3, Table S4). Next, we investigated the raw data (both ChIP read counts and log base 2 ratios of ChIP over scaled input read counts) within each of these classes. Additional file 3, Figures S16(a) and (b) display boxplots of log base 2 ratios for 11 (detected as enriched by both jMOSAiCS and chromHMM), 10 (detected as enriched only by jMOSAiCS), 01 (detected as enriched only by chromHMM), and 00 (not detected as enriched by either method) groups within each mark across the two cell lines. For each GATA1-occupied segment, we used the bin with the maximum log base 2 ratio to generate these plots. Overall, we observed that regions declared as enriched by both methods showed the most enrichment of ChIP compared to input and the regions identified as unenriched by both methods

(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
16	Active	1	0	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	16
1	7209	314	11	0	1	3	0	152	1	6	3	24	305	305	86	1
2	313	100	6	0	0	2	0	0	0	0	0	13	28	29	8	0
3	20	2	1	0	0	0	1	0	0	0	0	1	3	17	6	0
4	1	0	1	1	0	1	0	0	0	1	0	0	0	2	1	0
5	10	1	0	0	3	0	0	0	2	1	2	2	6	1	3	0
6	99	50	7	0	7	11	1	3	0	2	2	6	5	16	3	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
8	40	5	1	0	0	4	1	53	2	8	2	2	6	27	7	0
9	9	0	0	0	0	0	1	0	25	1	23	0	11	5	17	0
10	3	0	0	0	0	0	0	0	0	2	0	0	0	13	4	0
11	1	0	0	0	0	0	0	0	0	0	1	0	0	0	5	0
12	3	3	0	0	1	0	0	0	0	0	0	2	13	2	4	0
13	166	7	1	0	0	0	0	0	31	2	72	20	290	27	366	0
14	282	7	4	0	0	0	0	1	0	5	1	1	20	357	103	0
15	22	0	0	0	0	0	0	0	0	0	4	1	19	26	119	0
16	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

(c)

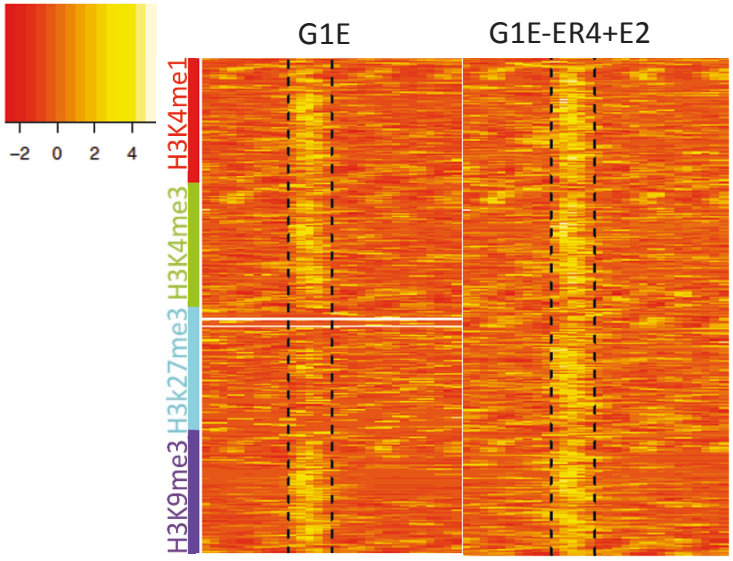

Figure 8 Analysis of mouse ENCODE histone ChIP-seq datasets. **(a)** List of combinatorial patterns identified by jMOSAiCS. Patterns 1 to 6 are also identified by chromHMM. **(b)** Changes in chromatin states between G1E and G1E-ER4+E2 cells for DNA segments occupied by GATA1 in the latter cells. **(c)** Heatmap of normalized raw data for a group of 311 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 1,400 bp in size) identified across different marks are aligned and depicted in between the dashed lines.

showed the least enrichment. Although the differences between 10 (jMOSAiCS specific) and unenriched regions did not appear as pronounced as the differences between 11 (enriched according to both methods) versus 00 (unenriched according to both methods), the observed differences were highly significant (P values < $1 \times e^{-10}$ with both the t-test and Wilcoxon rank-sum test), indicating that these regions are enriched for H4K3me3. In addition, we observed a relatively small number of regions identified as enriched exclusively by chromHMM. Although some of these appear to be true false negatives for jMOSAiCS, overall, they tend to be

regions with relatively low ChIP counts. Additional file 3, Figure S16(c) displays ChIP versus normalized input read counts across all the regions, stratified with respect to their enrichment status by jMOSAiCS and chromHMM, for the H3K4me1 mark in G1E-ER4+E2 cells. We also evaluated how the performance of chromHMM changed for the segments that were identified as changing from 1101 in G1E to 1111 in G1E-ER4 +E2 by jMOSAiCS. Although the 16-state chromHMM produced concordant results with jMOSAiCS for 58.5% of these GATA1-occupied loci, including the validated Atp6v1e and Cmas loci, it still mislabeled both Elf1 and Extl3 (Additional file 3, Table S5).

Discussion

Integrative analysis of multiple ChIP-seq datasets for enumerating enrichment patterns is an emerging need. We 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 [18]. Diagnostics is an important component of probabilistic model-based approaches. jMOSAiCS inherited 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 (for example, [11]), jMOSAiCS can efficiently handle multiple datasets and is accurate at both obtaining global and local structures. A comparison of jMOSAiCS with chromHMM revealed 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 was identified by chromHMM. Our computational experiments indicated that jMOSAiCS scales up well with large numbers of datasets and it can summarize key states with a penalized average silhouette criterion [23].

Our analyses illustrated 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 the 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 host and virus proteins has the potential to boost power for detecting regions enriched for the virus protein (for example, [28]).

Meta-analysis of multiple samples is another integrative approach to multiple ChIP-seq samples. However, the focus of such meta approaches (for example, MM-ChIP [29] and ChIPMeta [30]) 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 the 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 the marginal distributions of read counts in each dataset are modeled 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 [31] are two appealing, flexible candidates that can encode for increasingly complex dependencies. Furthermore, they can be tailored for specific characteristics of analyzed samples, for example, 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 (2 hours) to identify combinatorial patterns of four (twenty) datasets across the whole mouse genome with a single CPU on a 64-bit machine with an Intel Xeon 3.0 GHz 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:

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

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 number of enriched bins and $S_{id} \in \{1, \ldots, L_i\}$ the starting position of the set of enriched bins in region i. Then, we have:

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

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, that is, $\Pr(S_{id} = s \mid E_{id} = 1, B_i = 1, V_{id} = v) = 1/(L_i - v + 1)$, $s = 1, \ldots, L_i - v + 1$. The likelihood of full data is a product over I regions:

$$\begin{aligned}
\Pr(Y, E, B) &= \prod_{i=1}^{I} \Pr(Y_i, E_i, B_i) \\
&= \prod_{i=1}^{I} \Pr(Y_i|E_i, B_i) \Pr(E_i|B_i) \Pr(B_i) \\
&= \prod_{i=1}^{I} \left(\left[(1-\tau_1) \prod_{d=1}^{D} (1-E_{id}) p_{0id} \right]^{1-B_i} \left[\tau_1 \prod_{d=1}^{D} ((1-\eta_d) p_{01d})^{1-E_{id}} (\eta_d p_{1id})^{E_{id}} \right]^{B_i} \right).
\end{aligned} \tag{1}$$

We estimate f_{0d} and f_{1d} for each individual dataset separately using the MOSAiCS algorithm. Therefore, the quantities p_{0id} and p_{1id}, $i = 1, \ldots, I$, $d = 1, \ldots, D$ are fixed given $\hat{f}_{0d}$ and $\hat{f}_{1d}$. Because B, E, S, and V are unobserved variables, we derive an expectation-maximization [32] algorithm to obtain maximum likelihood estimators of τ_1 and $\eta = (\eta_1, \ldots, \eta_d)$ based on the likelihood in (1). The full data log likelihood can be written as:

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

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^t_1, η^t denote parameter estimates from the tth iteration:

E step:

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

M step:

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

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 $\Pr\left(B_i|Y_i, \hat{\tau}_1, \hat{\eta}\right)$ and $\Pr\left(E_{id}|Y_i, \hat{\tau}_1, \hat{\eta}\right)$ for false discovery rate control with a direct posterior probability approach [33] 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. Consecutive $n \in \{3, 5\}$ bins were organized into non-overlapping regions to facilitate B-level data generation. For each region i, $i = 1\ldots$, 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 $B_i = 1$, the E variable was simulated at the dataset level, for example, 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 in a region at random and allowed the number of consecutive bins with enrichment to vary within each region. For non-enriched bins, 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 (Additional file 3, 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 [19]. This allowed us to quantify the gain due to the joint modeling approach rather than differences in modeling 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 [33]. 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 the 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 the 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 the genome into TF binding/chromatin states based on the jMOSAiCS fit. We calculated the joint posterior probability of the E variables Pr($E_{i1} = r_1$, ..., $E_{iD} = r_D \mid Y_i, \tau_1, \eta$) for each combination of r_1, ..., 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 [6]

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 11,485 GATA1-occupied segments defined by [6] and enumerated H3K4me3, H3K4me1, H3K27me3, and H3K9me3 modification patterns of these regions across the two cell types. The median width of the GATA1-occupied segments was 800 bp and only 0.75% of the segments were wider than 2,000 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 [34]. The relative levels of the specific histone marks are indicated in the Additional file 3, Table S2. The PCR primers used to analyze the four loci are provided in Additional file 3, Table S3.

Additional material

Additional file 1: R package for jMOSAiCS.

Additional file 2: Vignette for the R package jMOSAiCS.

Additional file 3: Supplementary materials. This file contains further details on and additional results from the computational experiments presented as supplementary text and figures.

Abbreviations

bp: base pair; ChIP: chromatin immunoprecipitation; ENCODE: Encyclopedia of DNA Elements; ER: estrogen receptor; FDR: false discovery rate; GC: guanine-cytosine; jMOSAiCS: joint model-based one- and two-sample analysis and inference for ChIP-seq; MeCP2: methyl CpG binding protein 2; MOSAiCS: model-based analysis and inference for ChIP-seq data; NGS: next generation sequencing; PBMC: peripheral blood mononuclear cell; PCR: polymerase chain reaction; Pol II: polymerase II; TF: transcription factor.

Authors' contributions

SK conceived and designed the method, computational experiments, and the data analysis and wrote the paper. XZ designed and implemented the method, computational experiments, and the data analysis, and wrote the

paper. RS and EHB performed experimental validation of the selected targets. HL and QC contributed data. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
H3K9me3 ChIP-seq was performed by Henriette O'Geen in the lab of Peggy Farnham (University of Southern California), using PBMCs provided by Emery Bresnick (University of Wisconsin, Madison) and Swee Lay Thein (King's College London). We thank Jason Ernst PhD (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. We also thank Sushmita Roy PhD (Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison) for her insightful comments on an earlier version of the manuscript. This research is supported by a National Institutes of Health Grant (HG006716) to SK.

Author details
[1]Department of Statistics, University of Wisconsin-Madison, 1220 Medical Sciences Center, 1300 University Avenue, Madison, WI 53706, USA. [2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, K6/446 Clinical Sciences Center, 600 Highland Avenue, Madison, WI 53792-4675, 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, 325 Services Memorial Institute, 1300 University Avenue, Madison, WI 53706, USA. [4]Genetics Training Program, Waisman Center, University of Wisconsin-Madison, 1500 Highland Avenue, Madison, WI 53705-2280, USA. [5]Department of Genetics and Neurology, University of Wisconsin-Madison, 425 Henry Mall Madison, WI 53706, USA.

Received: 2 October 2012 Revised: 12 April 2013
Accepted: 29 April 2013 Published: 29 April 2013

References
1. 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**:R80+.
2. 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**:232-235.
3. Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M: **Genetic analysis of variation in transcription factor binding in yeast.** *Nature* 2010, **464**:1187-1191.
4. 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, *et al*: **Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project.** *Science* 2010, **330**:1775-1787.
5. 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.
6. 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**:1659-1671.
7. Wilbanks EG, Facciotti MT: **Evaluation of algorithm performance in ChIP-seq peak detection.** *PLoS ONE* 2010, **5**:e11471.
8. Ernst J, Kellis M: **Discovery and characterization of chromatin states for systematic annotation of the human genome.** *Nature Biotechnology* 2010, **28**:817-25.
9. 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**:Article 1.
10. 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**:e35-e35.
11. 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**:1000-1006.
12. Song Q, Smith AD: **Identifying dispersed epigenomic domains from ChIP-Seq data.** *Bioinformatics* 2011, **27**:870-871.
13. Taslim C, Huang T, Lin S: **DIME: R-package for identifying differential ChIP-seq based on an ensemble of mixture models.** *Bioinformatics* 2011, **27**:1569-1570.
14. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS: **Unsupervised pattern discovery in human chromatin structure through genomic segmentation.** *Nature Methods* 2012, **9**:473-476.
15. 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.
16. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of *in vivo* protein-DNA interactions.** *Science* 2007, **316**:1749-1502.
17. 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**:13765-13769.
18. Wu H, Ji H: **JAMIE: joint analysis of multiple ChIP-chip experiments.** *Bioinformatics* 2010, **26**:1864-1870.
19. Kuan PF, Chung D, Pan G, Thomson J, Stewart R, Kele 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.
20. Schwarz G: **Estimating the dimension of a model.** *The Annals of Statistics* 1978, **6**:461-464.
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**:e72.
23. Rousseeuw PJ: **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.** *Computational and Applied Mathematics* 1987, **20**:53-65.
24. Fujiwara T, O'Green H, Kele¸s S, Blahnik K, Linneman AK, Kang YA, Choi K, Farnham PJ, Bresnick EH: **Discovering hematopoietic mechanisms through genomewide analysis of GATA factor chromatin occupancy.** *Molecular Cell* 2009, **36**:667-681.
25. 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, **39**:7415-7427.
26. 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**:1908-1916.
27. 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.
28. 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**:14902-14907.

29. 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**:R11.
30. 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**:1715-1721.
31. Friedman N, Geiger D, Goldszmidt M: **Bayesian network classifiers.** *Machine Learning* 1997, **29**:131-163.
32. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm.** *Journal of the Royal Statistical Society: Series B* 1977, **39**:1-38.
33. Newton MA, Noueiry A, Sarkar D, Ahlquist P: **Detecting differential gene expression with a semiparametric hierarchical mixture model.** *Biostatistics* 2004, **5**:155-176.
34. 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.

doi:10.1186/gb-2013-14-4-r38
Cite this article as: Zeng *et al.*: jMOSAiCS: joint analysis of multiple ChIP-seq datasets. *Genome Biology* 2013 **14**:R38.