ChIP-GSM: cell type-specific cis-regulatory module inference for active regulatory element prediction

Xi Chen

Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, 900 North Glebe Road, Arlington, VA 22203, USA.

Contact information: xichen86@vt.edu.

Abstract

Transcription factors (TFs) usually function as cis-regulatory modules (CRMs) to activate enhancers or promoters and regulate target gene transcription. Cell type-specific ChIP-seq profiling of multiple TFs makes it feasible to infer functional CRMs for a particular cell type. Directly using ChIP-seq peaks may only recover CRMs partially since some weak binding events are likely missed by peak callers, resulting in an incomplete identification of co-binding events. We have developed a novel sampling-based approach, ChIP-GSM, for CRM inference and active regulatory element prediction. The approach uses a mixture model of Power-Law and Gamma distributions to model read counts in the foreground and background regions that enable the capture of more foreground regions regulated by CRMs with weak binding affinity. Specifically, ChIP-GSM samples read counts of multiple TFs iteratively for the joint effect of each potential TF combination; using inferred CRMs as novel features, ChIP-GSM employs logistic regression to accurately predict active regulatory elements. We first benchmarked ChIP-GSM on simulated ChIP-seq data and real data with known CRMs. ChIP-GSM shows superior performance over existing module inference methods. We

then applied ChIP-GSM to ChIP-seq data acquired from breast cancer MCF-7 cells and revealed that CRMs are functional distinctly at enhancer and promoter regions. To broaden the applicability of ChIP-GSM, we further deployed ChIP-GSM to infer CRMs from 9 different cell types and used CRMs as features to predict active enhancers or target genes; the performance of ChIP-GSM is significantly improved over that of using individual TF peaks. Finally, we performed functional enrichment analysis of the target genes of CRMs inferred from K562 cells and demonstrated that CRMs are likely to be active at different time points to mediate distinct cellular functions.

1. Introduction

ChIP-seq profiling provides a quantitative measurement of binding of a specific transcription factor (TF) to DNA regions or sites [1]. TFs often cooperate by forming cis-regulatory modules (CRMs) to regulate target genes through activating enhancer or promoter regions [2], where one or two TFs may serve as major regulators (or factors) and recruit other co-factors to perform context specific regulation [3]. Joint analysis of multiple TFs' ChIP-seq profiles has become a powerful tool to study the cell type-specific regulatory mechanism. Several CRM identification approaches have been proposed based on ChIP-seq data. For example, ChromHMM [4] infers chromatin states using ChIP-seq peak files of multiple histone markers. jMOSAiCS [5] infers the most reliable module at each genomic segment. SignalSpider [6] better utilizes 'weak' binding signals during module inference by modeling read coverage instead of the read count. However, input data are not used in this method and the false positive rate in the results

cannot be ignored, especially when weak peaks are included in the analysis. As shown in [7], input data play a key role in differentiating weak peaks from background signals. To interpret the function of CRMs, functional studies like analyzing the function of individual TFs, target genes expression, or gene set functional enrichment are needed. Therefore, instead of working on broad genome regions, inferring CRMs at enhancer or promoter regions would better explain CRM functions because we can associate target genes with those regions based on existing biological knowledge.

It is also important to predict active regulatory regions based on CRMs since a regulatory region is usually activated by a CRM rather than a single TF [8-10]. A prediction method, EMERGE [11], was proposed to predict active regulatory elements based on TF bindings. For a cell type, EMERGE merges all cell type-specific ChIP-seq datasets and uses a logistic regression framework [12] to predict genomic regulatory regions. If TF ChIP-seq datasets are used, it assigns a weight to each TF, labels presence or not of each region in all ChIP-seq datasets and uses a weighted sum of TF bindings at each region for active regulatory element prediction. As regulatory regions are activated by CRMs rather than a single or all TFs, we expect that using CRMs to predict active regulatory elements should have a better prediction power than using a weighted sum of individual TFs.

We develop a Gibbs sampling-based approach, ChIP-GSM, for CRM inference and active regulatory region prediction from multiple ChIP-seq profiles. We assume that a foreground region (targeting by a specific TF's antibody during the 'ChIP' process) is

more likely to be sequenced multiple times during the 'seq' step than any background regions (non-specific binding area). For each TF, we use a mixture model of Power-Law and Gamma distributions [13] to model read counts sequenced from foreground or background regions, respectively. ChIP-GSM assigns each region a weight (a probability) representing the probability that this current region will be sequenced. This weight is calculated based on the observed read count and Power-Law or Gamma distribution, depending on whether it is a foreground or background region. The sum of all weights equals to 1. Then, under a Gibbs sampling framework, we infer the most reliable CRM (a combination of TFs) at each foreground region by updating the weight of each TF iteratively with the sampling of CRMs. The sampling frequency of each CRM represents the posterior probability. For each CRM, the average sampling frequency at all regions denotes its relative importance in the cell, which can be used to prioritize CRMs. Based on inferred CRMs, ChIP-GSM further uses a logistic regression approach [12] to predict whether cis-regulatory regions are active or not for a particular cell type.

To demonstrate the capability of ChIP-GSM on CRM inference, we applied ChIP-GSM to simulated ChIP-seq profiles and compared its performance with existing methods. The superior performance on capturing more foreground regions of weak enrichment makes ChIP-GSM more feasible to use in real ChIP-seq data analysis, especially in large-scale studies with many TFs and diverse TF-DNA binding strengths. We then benchmarked ChIP-GSM and competing methods on a H1-hESC cell ChIP-seq data set, including two TFs and two histone markers and known association relationships. Our results show that ChIP-GSM captures more regions under regulation of high order TF

combinations. As a real application, we used ChIP-GSM to infer CRMs from MCF-7 breast cancer cells. We found that CRMs acting at enhancer and promoter regions are quite different. We continued CRM inference on another 8 cell types. We aimed to examine whether TF combinations as CRMs have a stronger prediction power on active enhancers/genes than individual TFs, and whether this is true in different cell types. For each cell type, active/non-active enhancers or genes were selected from the FANTOM5 database based on their expression value. Using logistic regression, the prediction performance of using CRMs is significantly improved when compared with using binary TF peaks. Finally, we studied CRMs inferred from promoter regions of K562 cells, which have the best prediction performance on active regions. Based on TF similarity, we classified CRMs into several groups and found that each group regulates very distinct sets of target genes. Further validation using a time-course gene expression dataset shows that after a stimulus, each group of CRMs are active at a specific time points and involved in the regulation of a unique set of cellular processes.

2. Methods

2.1 Description of the ChIP-GSM approach

The workflow of ChIP-GSM is shown in Fig. 1. Given candidate genomic locations, we partition each region into 500 bps bins and count read tags in each bin from each TF's ChIP-seq profile using HOMER [14]. To infer CRMs, we apply a probabilistic Power Law-Gamma mixture model to assign the weights of foreground or background regions. ChIP-GSM iteratively samples foreground and background regions, updates weights of TFs at each region, and samples possible TF combinations under a Gibbs sampling

framework. CRMs and their associated foreground regions are identified after multiple rounds of sampling. At each region, the sampling frequency of a CRM represents the posterior probability of regulation. All inferred CRMs can further be ranked according to the average sampling frequency at all regions. Inferred CRMs are then used as features in a logistic regression approach to predict active regulatory regions. Details of the ChIP-GSM approach are provided in the Methods section.

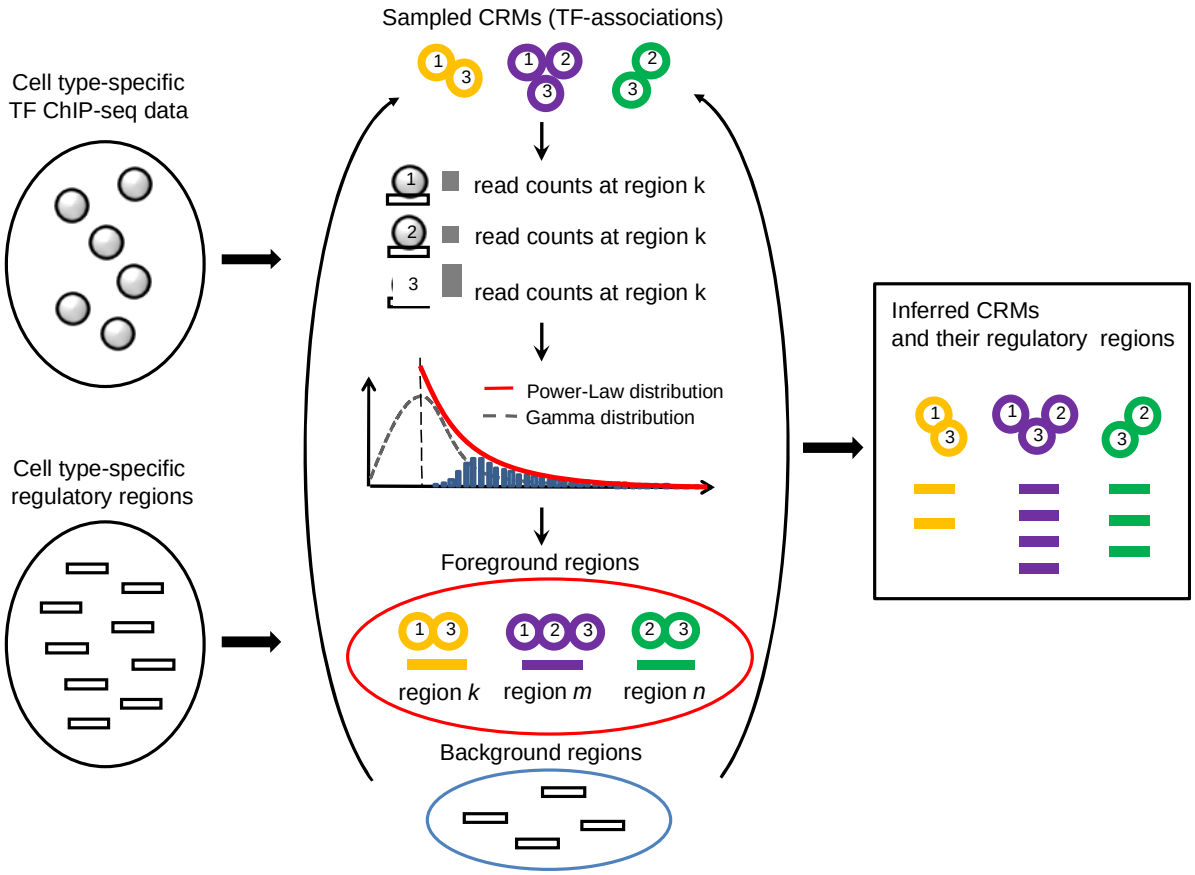


Figure 1. Flowchart of ChIP-GSM for CRM inference. A mixture model of Power-Law and Gamma distributions is first used to model read counts in the foreground and background regions. A Gibbs sampling approach is then used to infer the most reliable CRM at each foreground region.

The computational complexity of CRM inference is determined by the total number of TFs (T). In the past decade, substantial ChIP-seq data have been generated for many cell types, hence T continues to increase. The computational complexity for regulatory modules (TFs combinations) identification increases greatly (e.g., with 2^{20} combinations for T = 20) and it becomes impractical to explore all possible combinations exhaustively. For a number of TFs, given their read counts at annotated enhancer or promoter regions, ChIP-GSM automatically identifies a list of candidate CRMs. A detailed workflow is shown in *Supplementary Fig. S10*. A candidate CRM matrix $\mathbf{B}$ is defined with M rows ($M << 2^T$) and T columns. Each row is a constant binary vector $[b_{m,1}, b_{m,2}, ..., b_{m,t}, ..., b_{m,T}]$ representing a unique CRM with $\sum_{t=1}^{T} b_{m,t} \geq 2$. We add a zero row (row index m=0) to $\mathbf{B}$ since there are many background regions without any TF regulation.

Given K regions and M candidate CRMs, we define a matrix $\mathbf{C}$ with K rows and M columns with binary variables $c_{k,m}$. In each row, we assume that there is one and only one module regulating it, thus $\sum_m c_{k,m} = 1$. If the k-th region is regulated by the m-th CRM, then we have $c_{k,m} = 1$. At this time, for the t-th TF, the k-th region can be either a foreground region if $b_{m,t} = 1$ or a background region if $b_{m,t} = 0$. The observed read count $Y_{k,t}$ can be modeled by a mixture model as follows:

$$Y_{k,t} = \sum_m c_{b,m} b_{m,t} X_{k,t} + (1 - \sum_m c_{b,m} b_{m,t}) I_{k,t} + N_{k,t}, \quad (1)$$

where $X_{k,t}$ denotes the estimated read count given $c_{k,m}b_{m,t}=1$; $I_{k,t}$ denotes the estimated read count given $c_{k,m}b_{m,t}=0$; $N_{k,t}$ denotes the noise occurred during sequencing process. Histograms of observed read counts $Y_{k,t}$ at promoter regions for several general TFs are shown in *Supplementary Table S11*. Read counts in each sample ChIP-seq dataset have a long tail to the high value region, which can be fitted by a Power-Law distribution. Read counts in the input ChIP-seq dataset can be fitted by a Gamma distribution around low count area with a short tail.

To estimate $X_{k,t}$ or $I_{k,t}$, we define Power-Law and Gamma distributions as follows:

$$\begin{cases} P(X_{k,t}) = \text{Power-Law}(X_{\min}, \gamma_t) \\ P(I_{k,t}) = \text{Gamma}(\alpha_{I,t}, \beta_{I,t}) \end{cases}, \tag{2}$$

where $X_{k,t}$ is assumed to follow a Power-Law distribution with hyper-parameters $X_{\min}$ and γ_t; $I_{k,t}$ is assumed to follow a Gamma distribution with hyper-parameters $\alpha_{I,t}$ and $\beta_{I,t}$. $N_{k,t}$ is assumed to follow a Gaussian distribution with zero mean and variance σ_N^2. We assume a prior Inverse-Gamma distribution (conjugate prior of the variance of Gaussian distribution) on the variable σ_N^2 with hyper-parameters α_N and β_N. Using this Inverse-Gamma assumption we want to minimize the difference between observed read counts and estimated read counts to find an optimal solution.

Specially for promoter region study, as demonstrated in [15], there is an exponential decay effect on read enrichment along with the increase of relative distance of binding locations to TSS (promoter center). At background regions, the distribution of read

enrichment is more like uniform. Hence, to facilitate identifying foreground regions at gene promoters, for the t-th TF we define $d_{k,t}$ as the relative distance of the k-th region its original TSS. We propose another distance-based mixture model in the following equation to denote the regulatory effects of foreground or background regions on target gene regulation:

$$\begin{cases} P(d_{k,t} \mid \sum_m c_{b,m} b_{m,t} = 1) \sim \lambda_t \exp(-\lambda_t \left| d_{k,t} \right|) \\ P(d_{k,t} \mid \sum_m c_{b,m} b_{m,t} = 0) \sim \Delta d / d_p \end{cases}, \quad (3)$$

where λ_t is the exponential decaying parameter; Δd denotes region length; d_p is the length of promoter. λ_t is specific for the t-th TF and needs to be estimated from the location distribution of foreground regions. We assume a prior Gamma distribution (conjugate prior of exponential distribution [16]) on λ_t with hyper-parameters $\alpha_{\lambda,t}$ and $\beta_{\lambda,t}$. From *Supplementary Table S12*, it can be found that the enrichment of real ChIP-seq data along the promoter region is well fitted using the proposed distance-based mixture model. In enhancer study, for both foreground and background regions, this effect is uniformly distributed at the whole enhancer region.

Given read counts candidate modules matrix $\mathbf{B}$, observed read counts matrix $\{\mathbf{Y} \mid Y_{k,t}, k = 1 \sim K, t = 1 \sim T\}$ and binding location matrix $\{\mathbf{D} \mid d_{k,t}, k = 1 \sim K, t = 1 \sim T\}$ in order to estimate $\{\mathbf{X} \mid X_{k,t}, k = 1 \sim K, t = 1 \sim T\}$, $\boldsymbol{\lambda} = \{\lambda_t \mid t = 1 \sim T\}$ (for promoter focused study only), and $\mathbf{C}$ at foreground regions, or $\{\mathbf{I} \mid I_{k,t}, k = 1 \sim K, t = 1 \sim T\}$ at background regions, we define a posterior probability $P(\mathbf{X}, \mathbf{I}, \boldsymbol{\lambda}, \mathbf{C} \mid \mathbf{Y}, \mathbf{D}, \Phi)$. Φ includes $\boldsymbol{\gamma} = \{\gamma_t \mid t = 1 \sim T\}$, $X_{\min}$,

$\boldsymbol{\alpha}_I = \{\alpha_{I,t} \mid t = 1 \sim T\}$, $\boldsymbol{\beta}_I = \{\beta_{I,t} \mid t = 1 \sim T\}$, α_N, β_N, $\boldsymbol{\alpha}_\lambda = \{\alpha_{\lambda,t} \mid t = 1 \sim T\}$ and $\boldsymbol{\beta}_\lambda = \{\beta_{\lambda,t} \mid t = 1 \sim T\}$. γ_t and $X_{\min}$ can be learned from the sample ChIP-seq profile of the *t*-th TF by setting the minimum read count to a small number (e.g., 10 in our implementation). $\alpha_{I,t}$ and $\beta_{I,t}$ can be learned from the ChIP-seq input profile prepared for the *t*-th TF. We assume an informative prior on σ_N^2 with parameters $\alpha_N = 2$ and $\beta_N = 5$, and a non-informative prior on λ_t with parameters $\alpha_{\lambda,t} = 1$ and $\beta_{\lambda,t} = 1$. We convert the module identification problem to an optimization problem by maximizing the posterior probability $P(\mathbf{X}, \mathbf{I}, \boldsymbol{\lambda}, \mathbf{C} \mid \mathbf{Y}, \mathbf{D}, \mathbf{B}, \Phi)$.

Considering that the total number of read tags in each ChIP-seq dataset is limited, read counts at different regions are not independent. In our proposed method, for the *k*-th region we calculate a weight based on the estimated read count $X_{k,t}$ (or $I_{k,t}$) given binding state $c_{k,m} b_{m,t} = 1$ (or 0). Using a Gibbs sampling strategy, we assign read tags one by one to all regions (according to their weights by simulating the ChIP-seq generation process), update binding states, estimate all variables and achieve the final identification of CRM at each region.

2.2 Gibbs sampling process

I. Model initialization

For the *t*-th ChIP-seq profile, we first roughly estimate the total number of reads to be assigned to foreground (or background) regions. Initially, we assume that all regions are background regions (before a TF specific antibody is applied to the cell) and randomly

generate a read count $I_{k,t}$ for each region according to Eq. (2). Then, a weight is calculated as $p_{k,t} = I_{k,t} / \sum_{k'} I_{k',t}$. Based on the initial state of $\mathbf{C}$, for any binding event $c_{k,m} b_{m,t} = 1$ we amplify its weight $p_{k,t}$ by F times. Here, the amplification factor F represents the average fold change between foreground and background regions. Assuming that the total number of reads in the *t*-th TF ChIP-seq dataset is R_t, the ratio between the total numbers of reads assigned to foreground regions ($R_{X,t}$) and background regions ($R_{I,t}$) can be calculated as follows:

$$r_t = R_{t,X} / R_{t,I} = F \sum_{k,m} c_{k,m} b_{m,t} p_{k,t} \Big/ \sum_{k} (1 - \sum_{m} c_{k,m} b_{m,t}) p_{k,t} , \quad (4)$$

where $R_{X,t} = R_t \cdot r_t / (1 + r_t)$ and $R_{I,t} = R_t / (1 + r_t)$.

II. Sampling read counts

For foreground regions, given $R_{X,t}$ reads, we first assign $\sum_{k,m} c_{k,m} b_{m,t} X_{\min}$ read tags evenly to individual regions. We then assign the remaining read tags one by one to each region according to a conditional probability $P(X_{k,t} \mid Y_{k,t}, \mathbf{C}, \mathbf{B}, \Phi)$ defined in the following equation:

$$P(X_{k,t} \mid Y_{k,t}, d_{k,t}, \mathbf{C}, \mathbf{B}, \Phi) \propto P(Y_{k,t} \mid \sum_{m} c_{b,m} b_{m,t} = 1, \sigma_N^2, X_{k,t}) P(X_{k,t}) P(d_{k,t} \mid \sum_{m} c_{b,m} b_{m,t} = 1) . \quad (5)$$

For background regions, in total $R_{I,t}$ reads should be assigned to individual regions according to their weights as defined in the following equation:

$$P(I_{k,t} \mid Y_{k,t}, \mathbf{C}, \mathbf{B}, \Phi) \propto P(Y_{k,t} \mid \sum_{m} c_{b,m} b_{m,t} = 0, \sigma_N^2, I_{k,t}) P(I_{k,t}) . \quad (6)$$

III. Sampling model parameters

There are two unknown model parameters as exponential distribution parameter λ_t and the noise variance σ_N^2. The conditional probability of λ_t can be calculated as follows:

$$P(\lambda_t \mid \mathbf{D}, \mathbf{C}, \mathbf{B}, \Phi) \propto \prod_{k,m} P(d_{k,t} \mid \sum_m c_{b,m} b_{m,t} = 1) P(\lambda_t) . \tag{7}$$

As demonstrated in *Supplementary Material S10,* Eq. (7) is a Gamma distribution so we sample λ_t according to a Gamma distribution with parameters $\alpha_{\lambda,t} + \sum_{k,m} c_{k,m} b_{m,t}$ and $\beta_{\lambda,t} + \sum_{k,m} c_{k,m} b_{m,t} \left| d_{k,t} \right|$. σ_N^2 represents the overall fitting performance between the observed and estimated read counts. The conditional probability of σ_N^2 is defined as follows:

$$P(\sigma_N^2 \mid \mathbf{Y}, \mathbf{X}, \mathbf{I}, \mathbf{C}, \mathbf{B}, \Phi) \propto \prod_{k,m,t} P(Y_{k,t} \mid Y'_{k,t}, \sigma_N^2) P(\sigma_N^2) , \tag{8}$$

where $Y'_{k,t} = X_{k,t} \sum_m c_{b,m} b_{m,t} + (1 - \sum_m c_{b,m} b_{m,t}) I_{k,t}$. As detailed in *Supplementary Material S10*, Eq. (8) is an inverse-Gamma distribution so we directly sample σ_N^2 with updated parameters $\alpha_N + KT/2$ and $\beta_N + \sum_{k,t} (Y_{k,t} - Y'_{k,t})^2 / 2$.

IV. Sampling CRMs

The sampling of CRM for each region can be achieved by calculating a discrete probability density distribution of all candidate CRMs. At the k-th region, for each candidate CRM we calculate a conditional probability $P(c_{k,m} = 1 \mid \mathbf{Y}, \mathbf{X}, \mathbf{I}, \mathbf{D}, \mathbf{B}, \Phi)$ by assuming $c_{k,m} = 1$. We collect probabilities for all candidate CRMs at each region and then randomly sample a CRM according to the discrete probability distribution. After repeating CRM sampling at all regions, we bring an updated matrix $\mathbf{C}$ back to Eq. (4) to

start a new round of sampling. After drawing enough samples, we obtain a weighted matrix $\hat{\mathbf{C}}$ with each element $0 \le \hat{c}_{k,m} \le 1$ denoting the sampling frequency of the *m*-th CRM at the *k*-th region. $\hat{c}_{k,m}$ is an approximate of the posterior probability of module regulation. Note that for each region, there may exist multiple $\hat{c}_{k,m} > 0$.

2.3 Active regulatory region prediction

Given the inferred module results as $\hat{\mathbf{C}}$, we determine an optimal set of weights ($\boldsymbol{\beta}$) for CRMs so as to predict the activity of each regulatory region. The optimal weight determination requires a training process with true positive ($y_k = 1$) and true negative ($y_k = 0$) regulatory regions. For enhancer studies, the training dataset can be obtained from cell type-specific enhancer RNA (eRNA) expression data or from a validated enhancer list for each cell type. Active promoter regions can be selected from cell type-specific Pol II ChIP-seq data or mRNA expression data. We use the elastic net logistic regression method [12] to learn weights on the identified CRMs by training a binomial model. Elastic net is particularly suitable in this application because the features used here, CRMs, are highly correlated and have very strong group effects. Elastic net logistic regression assigns similar weights to correlated features or removes them together by assigning zero weights. Compared to linear regression, elastic net regression extends the method of least squares by adding a regularization (or penalty) term that includes the values of weights ($\boldsymbol{\beta}$) in the minimization process:

$$\min_{\beta_0,\beta_1,\ldots\beta_M} -\frac{1}{K'}\sum_k \left[y_k(\beta_0 + \sum_m \hat{c}_{k,m}\beta_m) + \log(1+\exp(\beta_0 + \sum_m \hat{c}_{k,m}\beta_m)) \right] + \lambda_e P_\alpha(\boldsymbol{\beta}) , \quad (9)$$

$$\text{with } P_\alpha(\boldsymbol{\beta}) = \sum_m \left(\frac{1-\alpha}{2}\beta_m^2 + \alpha|\beta_m| \right), \quad (10)$$

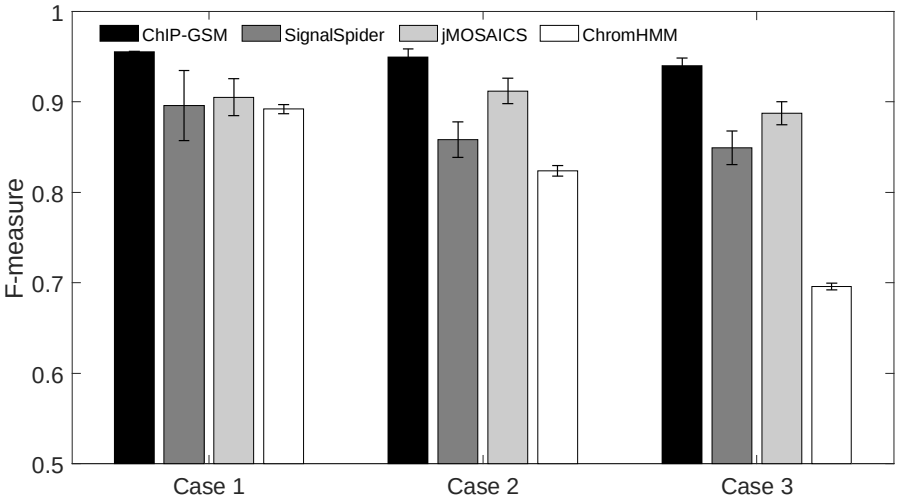
where K' is the total number of true positive and true negative regions; *m* is the index of modules; $\lambda_e \geq 0$ is a parameter controlling the model complexity; $0 \leq \alpha \leq 1$ controls a trade-off between the ridge and lasso penalty used in the elastic net regression method [17]. After the training, we obtain an optimal set of weights including β_0 and $\beta_1 \ldots \beta_m \ldots \beta_M$ for individual CRMs.

3. Results

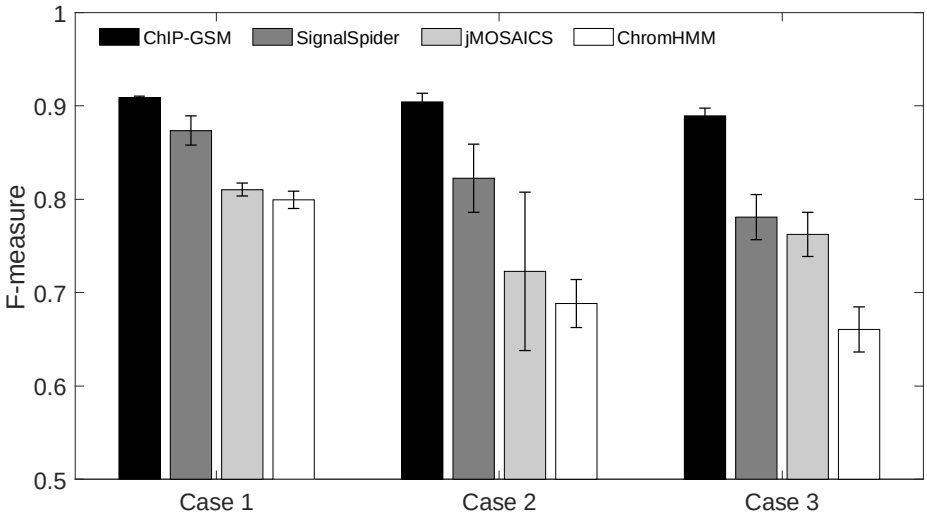
3.1 Realistic simulation

We performed realistic simulation by simulating ChIP-seq read counts according to real ChIP-seq data. The advantages of this approach are (1) in real data, each TF is unique in both ChIP-seq read count distribution and binding locations; (2) strong and weak binding events are included and (3) real cooperative relationships between TFs (co-binding events) are retained even though TFs are simulated individually. For simulation, we downloaded the ENCODE breast cancer MCF-7 ChIP-seq profiles of 20 TFs (http://genome.ucsc.edu/ENCODE/). A detailed simulation procedure is described in *Supplementary Material S1.* We simulated three cases with different numbers of TFs: 4 TFs in Case 1, 7 TFs in Case 2 and 20 TFs in Case 3 (*Supplementary Table S1*). Read count distributions of several TFs ChIP-seq profiles and the input profile are provided in *Supplementary Fig. S1.* For each case, we generated 10 replicates by randomly

shuffling each TF's read counts at different locations; we did not change overall read count distribution and binding locations. ChIP-GSM and competing methods (ChromHMM, jMOSAiCS, SignalSpider) were tested in all cases. Scripts or commands used to run each program are provided in *Supplementary Material S2.*



(A)



(B)

Figure 2. Performance comparison between ChIP-GSM and competing methods using realistically simulated ChIP-seq data. (A) F-measures of each method on all foreground regions in Cases 1, 2, and 3; (B) F-measure of each method on those foreground regions with weak bindings only.

Capability of the CRM inference is mainly reflected by the full recovery of all 'true' bindings at each region. A F-measure (F=2/(1/precision+1/recall)) is used to evaluate the performance of each method (Fig. 2). As can be seen from Fig. 2, ChIP-GSM provides improved performance over existing methods tested here. While for regions with at least one 'weak' binding event from any TF (read count < 50), ChIP-GSM provides a more significantly improved performance over competing methods.

3.2 Benchmarking ChIP-GSM on modules with known associations

We downloaded a ChIP-seq data set with four factors (H3K27me3, H3K4me3, EZH2, SUZ12) in H1-hESC cells from the ENCODE database. Associations among these four factors have been well demonstrated in previous biological studies. EZH2 and SUZ12 are Polycomb-group (PcG) proteins [18] that co-bind to 'bivalent' domains in H1-hESC cells. Binding is marked by H3K4me3 and H3K27me3 [19,20]. Hence, strong cooperation exists among all four factors and a large number of regions under their simultaneous regulation are expected. Using ChIP-GSM, we identified 14,195 regions with a PcG co-binding event only (EZH2 and SUZ12), 1640 regions with a bivalent signature (H3K4me3 and H3K27me3) only, and 6,564 regions with all four factors (Fig. 3(A)). ChIP-GSM has the largest proportion of foreground regions under the regulation of all four factors (the longest red bar as shown in Fig.3(B)). The read count distributions at foreground regions (predicted by ChIP-GSM) under regulation of all four factors are shown in Figs. 3(C)-(F). As shown in Figs. 3(C) and (F), binding signals (red bar) of EZH2 and H3K27me3 are very strong with a long and thick tail to high read count area. In contrast, in Figs. 3(D) and (E) binding signals (red bar) of SUZ12 and H3K4me3 are

relatively weak with a short tail. Note that regions with moderate read counts are sampled more frequently by ChIP-GSM than those with very high read counts.

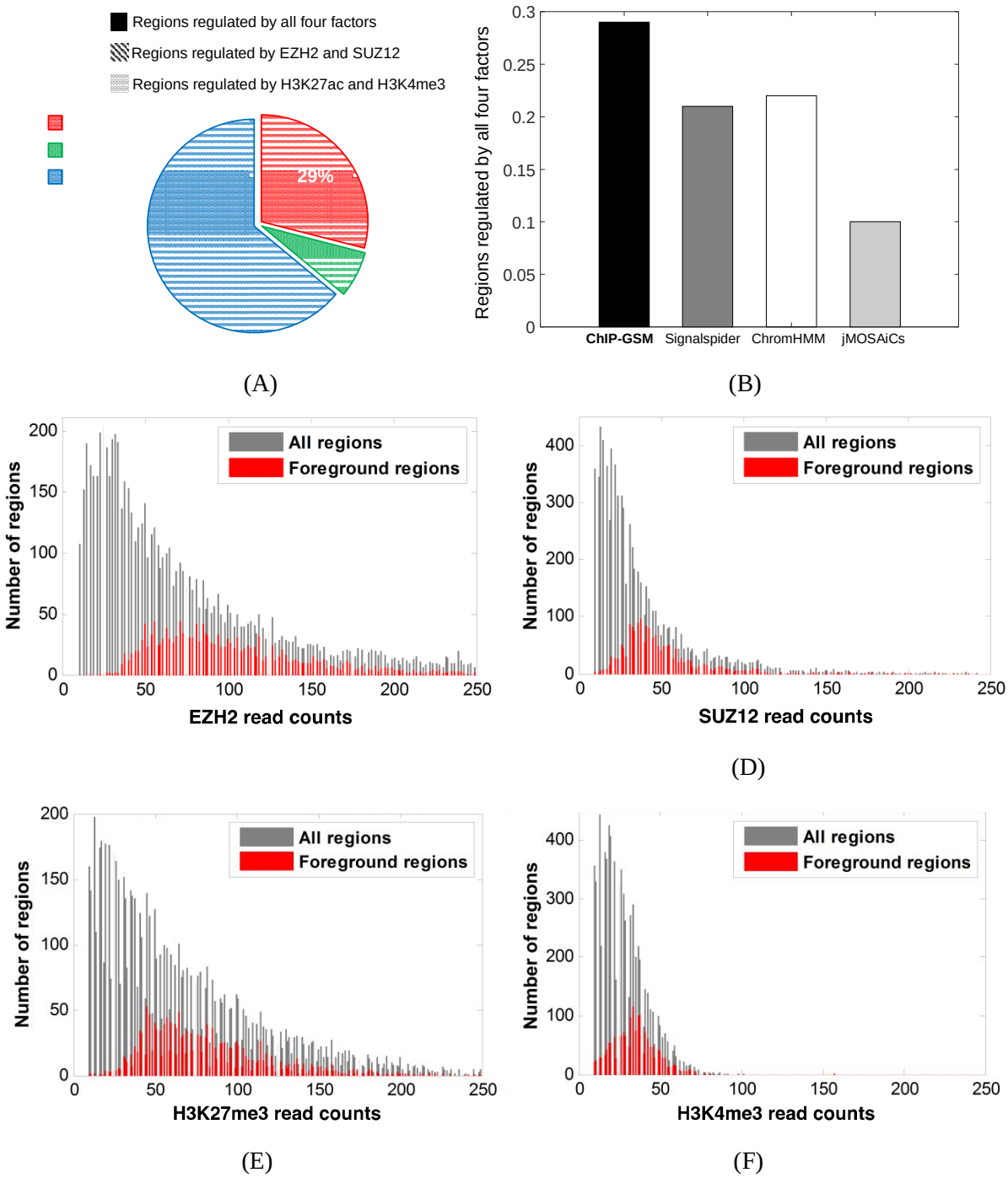


Fig. 3. Benchmarking ChIP-GSM results on ChIP-seq data generated from H1-hESC cells. (A) A pie chart of genomic regions with different co-binding events; (B) the proportion of regions under regulation of all four factors simultaneously; the read count distribution of (C) EZH2, (D) SUZ12, (E) H3K27me3, or (F) H3K4me3 at co-regulated foreground regions.

3.3 Inferring cell type-specific CRMs from enhancer or promoter regions

We further examined the TF-association differences in CRMs inferred from enhancer and promoter regions. We checked all cell types with ChIP-seq data in ENCODE database and selected 9 cell types for which the number of TF ChIP-seq profiles was larger than 20 and cell type-specific enhancer regions were also available (***Supplementary Tables S2 and S3***). We downloaded MCF-7 ChIP-seq data of 69 TFs and also its cell type-specific enhancer regions from the ENCODE database. Promoter regions were extracted from the human RefSeq file using the UCSC genome browser (human reference genome hg19). Details for this procedure can be found in ***Supplementary Section S5***. Using ChIP-GSM, we identified 52 CRMs including 34 TFs at enhancer regions, and 66 CRMs including 28 TFs at promoter regions. Similarities of ChIP-GSM identified TFs or CRMs between enhancer and promoters studies were respectively shown in **Fig.4A**. TF similarity (2*overlap/(no. of enhancer study + no. of promoter study)) is 77%, while CRM similarity is only 54%. We further checked TF similarity vs. CRM similarity in the other cell types (**Fig.4B**) and found that given a similar set of TFs at enhancer or promoter regions (average similarity 77%), the similarity of TF-associations (CRMs) is significant lower. In other words, CRMs are significantly different between enhancer and promoter regions, indicating their different functional categories in the whole genome. Therefore, for the same set of TFs, considering the specificity of CRMs at enhancer or promoter, it is necessary to infer CRMs respectively at these two different types of regions.

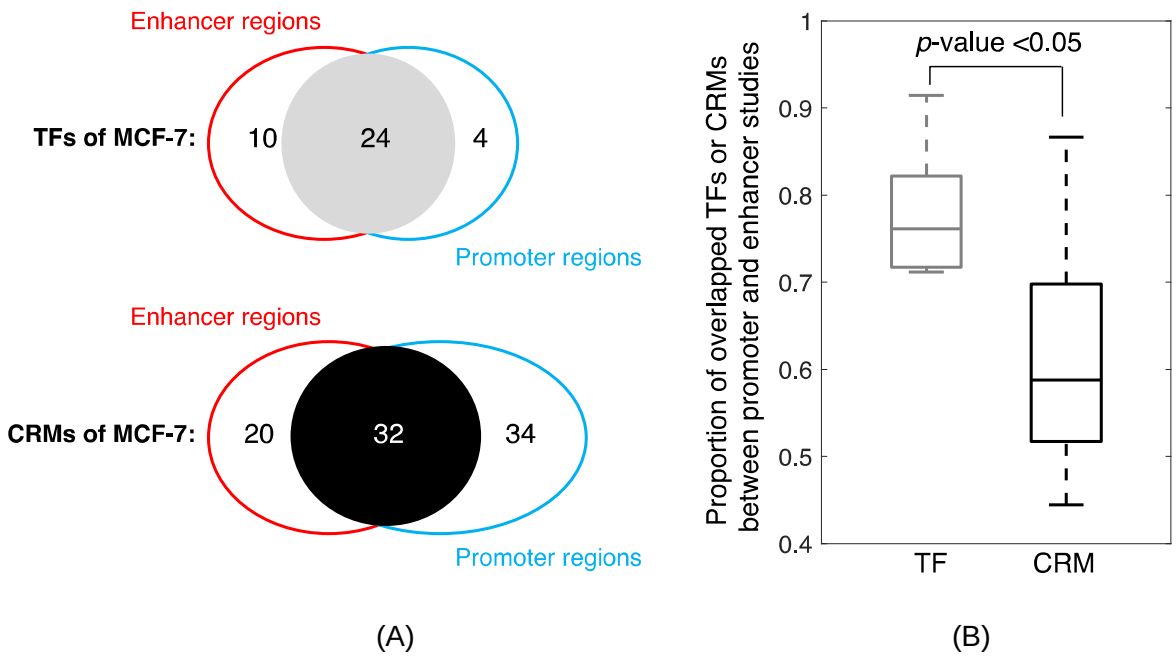


Fig. 4. Inferred cell type-specific CRMs are different between enhancer and promoter regions. (A) TF or CRM similarity between MCF-7 specific enhancer and promoter regions; (B) similarity of inferred CRMs are significantly lower than the similarity of enriched TFs between enhancer and promoter regions (in six cell types).

3.4 Predicting active enhancers or promoters using inferred CRMs

We extended our CRM inference to another 8 cell types with at least 20 TFs in the ENCODE ChIP-seq database (*Supplementary Table S3*). Since cell-type specific enhancer regions are available in the ENCODE database for only 5 cell types, we inferred enhancer CRMs for 5 cell types and promoter CRMs for all 8 cell types (*Supplementary Table S4*). If the enhancer or promoter regions are more likely to be activated by TF modules rather than individual TFs, we expect that CRMs will have a stronger predictive power on active enhancers or promoters. We applied a logistic regression approach [12], using CRMs as features, to predict active enhancers or promoters. We then compared performance to EMERGE [11], which applied the same regression method but used TF binding events as features. Cell type-specific 'positive'

and ‘negative’ enhancers or promoters were selected from the FANTOM5 database [21]. Specifically, we used enhancer RNA (eRNA) expression of the CAGE data in FANTOM5 to denote enhancer activity, and mRNA expression data to denote promoter activity. Regions were classified into ‘positive’ or ‘negative’ groups with the median expression value used as a cut-off (numbers of selected regions are listed in *Supplementary Table S5*). We performed 5-fold cross validation using CRMs and TFs as features, respectively. For CRMs, in the training matrix each row denotes posterior probabilities of each region under regulation of all candidate CRMs; for TFs, each row denotes binary binding events of each TF. Regression model parameter settings are provided in *Supplementary Material S5.* Receiver operating characteristic (ROC) curves of cross validation using CRMs (ChIP-GSM) or TFs (EMERGE) are shown in Fig. 5(A) (enhancers) and Fig. 5(B) (promoters) for the MCF7 cell line study. In Fig. 5(A), the area under ROC curve (AUC) of ChIP-GSM for active enhancer prediction is 0.85, while the AUC of EMERGE is 0.77. As shown in Fig. 5(B), a similar performance improvement is achieved for active promoter prediction. For the extended study of 8 cell types, the average AUC of 5-fold cross validation is computed for each cell type and a boxplot of AUC performances of all cell types is presented in Fig. 5(C) (for active enhancer prediction) and Fig. 5(D) (for active promoter prediction). Regardless of cell type specificity, the average AUC for active enhancer prediction is 0.81 using CRMs as compared to 0.74 using TFs (Fig. 5(C)). The overall AUC performance using CRMs inferred from ChIP-GSM is significantly better than using TFs (EMERGE). As shown in Fig. 5(D), the average AUC for active promoter prediction is 0.83, clearly higher than the average value 0.73 of using TFs. An improved performance for using CRMs on active

enhancer or promoter prediction motivates us to further interrogate CRMs (TF associations) for a better interpretation of the activation mechanism of enhancers or promoters.

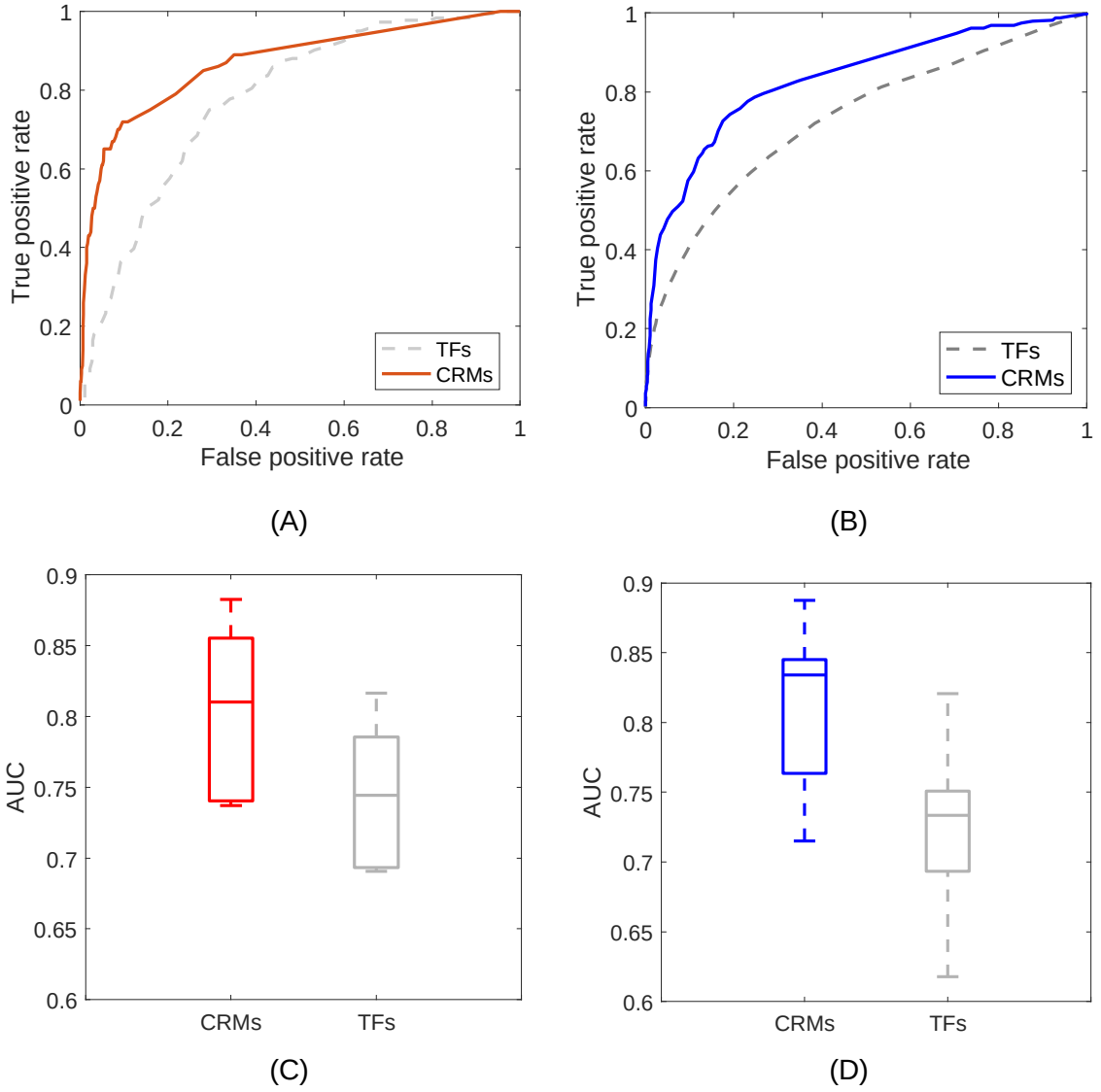
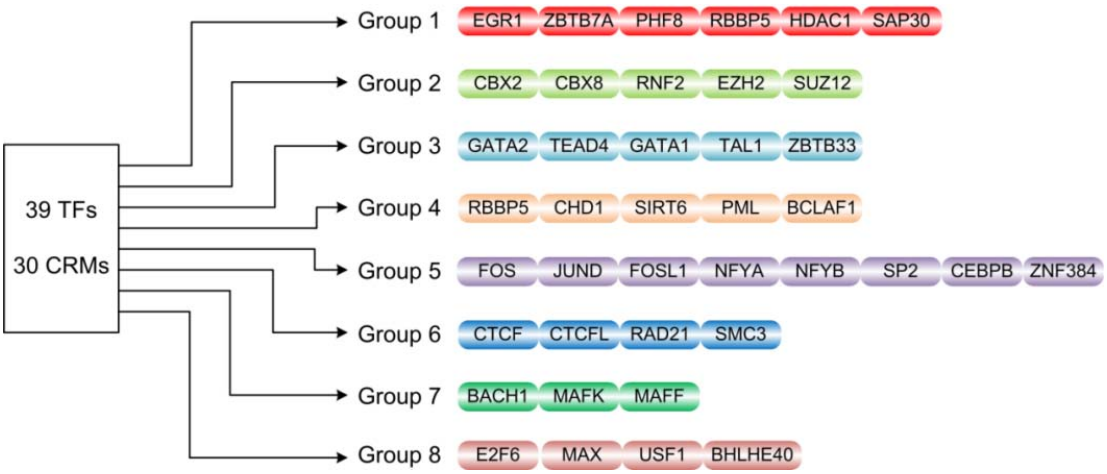


Figure 5. Prediction performance of active enhancers or promoters in 9 different cell types using ChIP-GSM results. (A) ROC curve for active enhancer prediction in MCF-7 cells; (B) ROC curve for active promoter prediction in MCF-7 cells; (C) AUC for active enhancer prediction in another 5 different cell types; (D) AUC for active promoter prediction in another 8 different cell types.

3.5 CRM function interpretation and target gene analysis

CRMs identified from promoter regions using ChIP-seq data of K562 cells have the best prediction performance on active promoters (green circle in Fig. 5(B)). It is important to determine whether target genes of different CRMs have distinct functions that may further explain why TFs work with each other in different ways. We identified 68 CRMs including 50 TFs (*Supplementary Table S6, Supplementary Fig. S4*). Compared with existing methods that can work on large-scale CRM inference using ChIP-seq data (*Supplementary Figs. S5~S7*), modeling raw ChIP-seq read counts rather than significant peaks provides an opportunity to discover some less dominant but biologically important modules. We filtered the 68 CRMs using two conditions: (1) each CRM regulates at least 500 foreground regions; (2) the posterior probability of each CRM is larger than 0.4. After filtering, 30 CRMs that include 39 TFs (~80% of all investigated TFs) were selected. We clustered these 30 CRMs using hierarchical clustering based on TF similarity (Fig. 6(A) and *Supplementary Fig. S4*) and identified 8 groups.



(A)

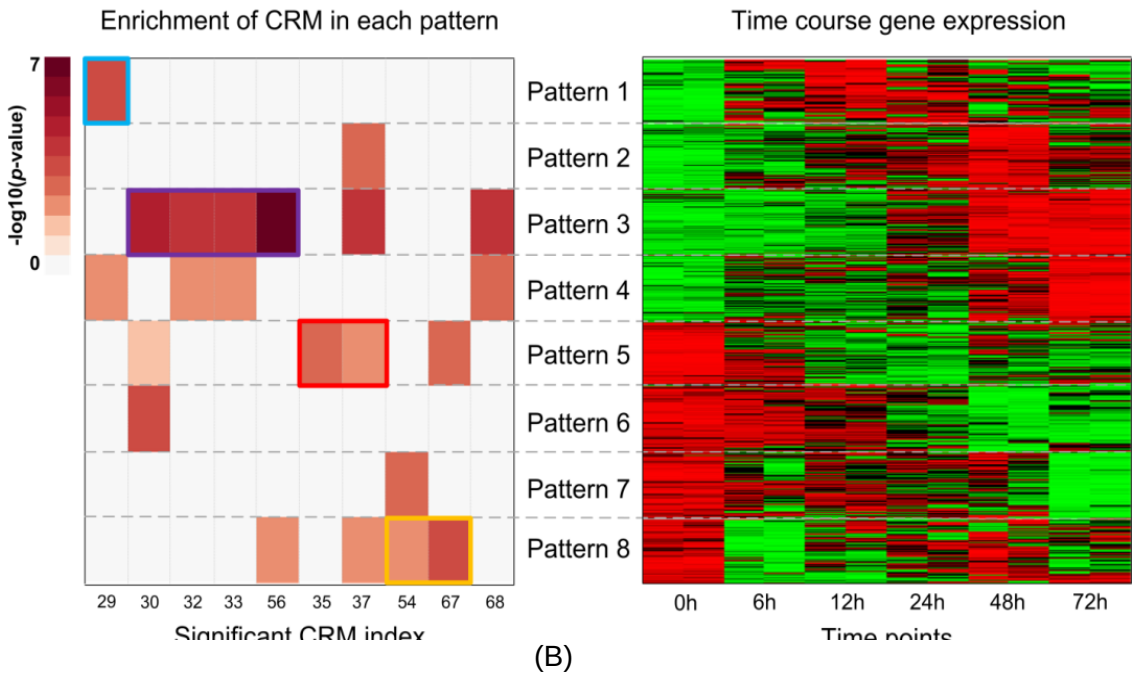


Figure 6. ChIP-GSM identified CRMs from the gene promoter region of K562 cells. (A) 8 groups of CRMs identified by ChIP-GSM; (B) enrichment of target genes of significant CRMs and gene clusters. 8 gene clusters are identified using K-means and the cut-off threshold for the color representation in the enrichment map is 0.001 (3 after -log10 transform).

For each group, we examined the functions of the TFs and found a significant enrichment of similar functions (*Supplementary Material S6*). We also checked co-expression between pairwise TFs in each group, since cooperative TFs are usually active simultaneously [22,23]. We downloaded time-course gene expression data (GSE1036) [24] and calculated the Pearson correlation coefficient for pairwise TFs (*Supplementary Fig. S8*). 6 out of 8 groups (75%) had significantly co-expressed TFs. TF co-expression in Group 2 or Group 4 was less significant but the functions of TFs in each group were closely related. For example, CBX2, CBX8, RNF2, EZH2 and SUZ12 in Group 2 are all Polycomb-group proteins [25].

We then analyzed the target genes of each CRM. Gene symbols can be found in *Supplementary Table S7.* The similarity of downstream target genes for the identified CRMs is shown in *Supplementary Fig. S9*. CRMs from the same group regulate distinct sets of target genes even though they share one or two common factors. For example, in Group 1, CRM 18 includes HDAC1 and SAP30 and regulates 815 genes. In CRM 20, these two TFs work together with PHF8 and RBBP5 regulating about 4,000 target genes with little overlap with the targets of CRM 18. These findings imply that main regulators recruit different co-factors, and function as CRMs to regulate different sets of target genes [26]. We further examined gene expression of those target genes regulated by each CRM [24]. An 'active' gene is selected if its fold change between any time point and the control (0h) is larger than 2. More details about active gene selection can be found in *Supplementary Material S7*. In total, we identified 1,569 active genes. For each CRM, we then calculated a hyper-geometric *p*-value to evaluate the active gene enrichment in its target gene list. There are 10 significant CRMs (~33%) with *p*-value < 0.05 (*Supplementary Table S8*). Functional genes of each significant CRM can be found in *Supplementary Table S9*. Using k-means clustering, we identified 8 clusters with a unique gene expression pattern in each. We assessed the functional gene enrichment for each pair of CRM and gene cluster. Numbers of genes in each cluster can be found in *Supplementary Table S10*. As shown in Fig. 6(C), each group of CRMs are more likely to be enriched in the same gene cluster. Since each gene cluster includes genes that are active at a specific time point, based on in Fig. 6(C), we can see that different groups of CRMs are rarely active at the same time. For instance, CRM 29 is highly enriched in gene cluster 1. CRMs 30, 32, 33 and 56 from Group 5 are significantly

enriched in gene cluster 3. CRMs 54 and 67 are very different and there is almost no overlap between their target gene lists. Nonetheless, CRM 54 and 67 are from module Group 4 and their target genes are enriched in gene cluster 8. We carried out gene set functional enrichment using DAVID [27], as summarized in *Supplementary Tables S11*. In Group 1, acute myeloid leukemia is uniquely and significantly enriched with *p*-value 0.014. In Group 4, there are more genes involved in cancer development, including chronic myeloid leukemia. In Group 5, cell survival related functions including cell death, apoptosis, and p53 signaling pathway are uniquely enriched. Functions of target genes regulated by different groups of CRMs are also different.

Discussion

Inferring CRMs among TFs is quite challenging because compared to histone modification signals, TF binding peaks are shorter, noisier, and more sensitive to the cell environment. We developed a probabilistic model, ChIP-GSM, for inferring cis-regulatory modules and predicting active regulatory elements. ChIP-GSM better models ChIP-seq read counts of strong and weak binding events so that many CRMs and regulatory regions containing weak binding TFs can be identified. ChIP-GSM is specifically designed to infer cell type-specific CRMs by integrating multiple TFs cell-type specific ChIP-seq binding signals at enhancer or gene promoter regions. Currently, there is very limited understanding of non-coding regions beyond regulatory elements, but ChIP-GSM is feasible to be extended and applied to inferring CRMs given any regions of interest.

We demonstrated that CRMs have a stronger prediction power on active regulatory elements than individual TFs. Further, using K562 cells as an example, we showed that target genes of different CRMs actively expressed at different time points and involved in distinct biological processes. This specific study revealed the functional diversity of CRMs. When TFs associated with different partners, they may regulate a different set of genes and their functional role within the cell may also change. We expected that CRMs at enhancer regions also have such function diversity. However, currently, very limited cell type-specific high-resolution 3D chromatin interaction data as ChIA-PET is available and the mapping between enhancers and genes has a lot of ambiguity, especially considering the fact that one enhancer can regulate multiple target genes and the nearest gene may not be one of them.

Currently we mainly applied ChIP-GSM to cell type-specific data as most ChIP-seq dataset were generated from cell line models. ChIP-GSM can also be applied to ChIP-seq data sequenced from heathy or disease tissues, in which the binding signals can be much noisier with very diverse binding strength. Applying ChIP-GSM to those cases has the potential to uncover more detailed TF-associations in CRMs that cannot be well captured by conventional methods. And the difference of the inferred CRMs may provide biological insights into protein functions of TF relevant to disease progression.

References

1. Blecher-Gonen, R. *et al.* High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. *Nat Protoc* **8**, 539-54 (2013).
2. Hardison, R.C. & Taylor, J. Genomic approaches towards finding cis-regulatory modules in animals. *Nat Rev Genet* **13**, 469-83 (2012).
3. Cheng, A.S. *et al.* Combinatorial analysis of transcription factor partners reveals recruitment of c-MYC to estrogen receptor-alpha responsive promoters. *Mol Cell* **21**, 393-404 (2006).
4. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-6 (2012).
5. Zeng, X. *et al.* jMOSAiCS: joint analysis of multiple ChIP-seq datasets. *Genome Biol* **14**, R38 (2013).
6. Wong, K.C., Li, Y., Peng, C. & Zhang, Z. SignalSpider: probabilistic pattern discovery on multiple normalized ChIP-Seq signal profiles. *Bioinformatics* **31**, 17-24 (2015).
7. Chen, X. *et al.* ChIP-BIT: Bayesian inference of target genes using a novel joint probabilistic model of ChIP-seq profiles. *Nucleic Acids Res* **44**, e65 (2016).
8. Chen, X. *et al.* CRNET: an efficient sampling approach to infer functional regulatory networks by integrating large-scale ChIP-seq and time-course RNA-seq data. *Bioinformatics* **34**, 1733-1740 (2018).
9. Chen, X. *et al.* Reconstruction of transcriptional regulatory networks by stability-based network component analysis. *IEEE/ACM Trans Comput Biol Bioinform* **10**, 1347-58 (2013).
10. Chen, X. *et al.* Statistical Identification of Co-regulatory Gene Modules using Multiple ChIP-Seq Experiments. in *International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS)* 109-166 (2014).
11. van Duijvenboden, K., de Boer, B.A., Capon, N., Ruijter, J.M. & Christoffels, V.M. EMERGE: a flexible modelling framework to predict genomic regulatory elements from genomic signatures. *Nucleic Acids Res* **44**, e42 (2016).
12. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301-320 (2005).
13. Zhang, Z.D., Rozowsky, J., Snyder, M., Chang, J. & Gerstein, M. Modeling ChIP sequencing in silico with applications. *PLoS Comput Biol* **4**, e1000158 (2008).
14. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-89 (2010).
15. Ouyang, Z., Zhou, Q. & Wong, W.H. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A* **106**, 21521-6 (2009).
16. Fink, D. *A Compendium of Conjugate Priors*, (See http://www.people.cornell.edu/pages/df36/CONJINTRnew%20TEX.pdf, 1997).
17. Hastie, T., Tibshirani, R. & Friedman, J.H. *The elements of statistical learning : data mining, inference, and prediction*, xxii, 745 p. (Springer, New York, NY, 2009).
18. Pasini, D., Bracken, A.P., Hansen, J.B., Capillo, M. & Helin, K. The polycomb group protein Suz12 is required for embryonic stem cell differentiation. *Molecular and Cellular Biology* **27**, 3769-3779 (2007).
19. Richly, H., Aloia, L. & Di Croce, L. Roles of the Polycomb group proteins in stem cells and cancer. *Cell Death Dis* **2**, e204 (2011).
20. Boyer, L.A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349-53 (2006).

21. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* **16**, 22 (2015).
22. Wang, W. *et al.* Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc Natl Acad Sci U S A* **102**, 1998-2003 (2005).
23. Wang, Y., Zhang, X.S. & Xia, Y. Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data. *Nucleic Acids Res* **37**, 5943-58 (2009).
24. Addya, S. *et al.* Erythroid-induced commitment of K562 cells results in clusters of differentially expressed genes enriched for specific transcription regulatory elements. *Physiol Genomics* **19**, 117-30 (2004).
25. Ram, O. *et al.* Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell* **147**, 1628-39 (2011).
26. Ji, Z. & Sharrocks, A.D. Changing partners: transcription factors form different complexes on and off chromatin. *Mol Syst Biol* **11**, 782 (2015).
27. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).