ChIP-GMM: A Gaussian Mixture Model for Inferring Binding Regions in ChIP-seq Profiles

Sharmi Banerjee, Xi Chen, Xiaowei Wu, Hehuang Xie, Jianhua Xuan, Wu-chun Feng
Virginia Tech
{bsharmi6, xichen86, xwwu, davidxie, xuan, wfeng}@vt.edu

Abstract

Chromatin immunoprecipitation (ChIP), followed by high-throughput DNA sequencing (ChIP-seq), enables genome-wide mapping of transcription-factor binding sites (TFBS). Several transcription factors (TFs) have been known to be able to differentiate tumor sub-types in diseases like cancer. For instance, the Luminal A and Luminal B sub-types of breast cancer tumors are high in estrogen receptor (ER) while human epidermal growth factor receptor 2 (HER2) tumors are high in HER2 protein. The accurate mapping of the DNA-protein loci is important in determining the causality of epigenetic regulation of gene expression under both normal and disease conditions in order to promote the development of targeted drug therapy. In this paper, we leverage the popular variational Bayes framework for Gaussian mixture models to demonstrate its effectiveness in identifying transcription-factor binding sites (TFBS) and common regions co-regulated by multiple TFs. We show that our method performs favorably when compared to existing peak calling and clustering methods. Our proposed method can both be used as a peak calling method as well as clustering co-regulated genomic regions acted upon by multiple TFs.

1 Introduction

Transcription factors (TFs) are proteins that regulate gene expression by binding to the DNA at specific locations along the genome. They play a significant role in many biological processes including cell growth, division, and signal transmission. These findings led to computational methods for identifying transcription-factor binding sites (TFBS) and leveraging the knowledge of genetics in the treatment of diseases through targeted drug therapy. The development of tools that can faithfully identify enriched binding sites and that are computationally inexpensive is necessary to analyze ChIP-seq data. Several peak calling methods have been developed in the past to identify TFBS. Some of the methods include MACS [8], PeakSeq [5], BCP [7] and Signal Spider [6].

PeakSeq relies on a local Poisson model while MACS uses a dynamic Poisson distribution to identify enriched regions. BCP uses Bayesian change point technology to identify read clusters. More recently, Signal Spider uses a Gaussian mixture model to approximate read intensity of ChIP-seq profiles to identify genomic regions co-regulated by multiple TFs but does *not* consider input ChIP-seq profiles. A second method [1] proposes a Gaussian mixture model for both sample and input ChIP-seq profiles to identify TF peaks and downstream genes, but uses only two components in the mixture. In addition, MACS and PeakSeq are read-count based and focus on strong TF-DNA bindings. However, the sequencing depth of ChIP-seq has been greatly improved leading to the discovery of weaker binding sites. Existing peak calling tools either do not consider or fail to identify these weak peaks.

We propose a variational Bayes (VB) structure with a multiple-component Gaussian mixture model (ChIP-GMM) that identifies TF-DNA binding sites by segregating ChIP from input. Adding more components to a Gaussian mixture increases its power to capture the finer properties of the data since each component has different eigenvalue spectra but similar eigenvector structure [9]. ChIP-GMM can cluster genomic regions co-regulated by multiple TFs. ChIP-GMM produces a list of region identified as peaks (for a single TF) or common TF binding regions (for multiple TFs) that are not based on p-value or any other significance test metric. We compare ChIP-GMM to PeakSeq and MACS and show that it performs favorably in terms of fold enrichment and the number of peaks identified.

2 Approach

Here we describe the workflow of ChIP-GMM and its associated mathematical model.

2.1 Workflow of ChIP-GMM

We leverage the variational Bayes (VB) approach [4], which has an inherent advantage over non-Bayesian methods in automatically selecting the model without

978-1-943436-05-7 / copyright ISCA, BICOB 2017
March 20-22, 2017, Honolulu, Hawaii, USA

over-fitting the data. In a variational approach, the posterior distribution is approximated over factorized parameter density and hidden variables. The convergence of the model is monitored by maximizing the negative free energy which consists of the average likelihood and Kullback-Leibler (KL) divergence between priors and posteriors. Clustering using ChIP-GMM is achieved by estimating the hidden parameters of the model and the probability of the Gaussian components associated with each point. To establish the association between the component and data points, we select the component with the highest weight since it is likely to have generated the data point.

As shown in Figure 1, input data to the model contains uniquely aligned reads from ChIP and the corresponding background/input signals. The pre-processing step removes low read coverage regions and normalizes the data. Next, we apply VB-GMM (short for *variational Bayes–Gaussian mixture model*) to the data, which generates a set of peak regions identified by the method. For multiple TFs, the input data are the ChIP signal profiles for each TF. Here the ChIP signal profile refers to the normalized value at each region.

Since ChIP-GMM clusters/segregates ChIP from input/control, we select those ChIP-seq data that had both ChIP and input reads. After aligning reads for ChIP and input, we use PeakSeq to get the accumulated read coverage for each region. PeakSeq uses a two-step peak calling tool that produces a directory of mapped reads for ChIP and input in the first step and a peak file with increasing Q-values in the second step. The parameters that determine the cut-off to select peaks are target FDR and max Q-value. Typically, these values are set to 0.05. Because we want to use PeakSeq as a pre-processing step to get the ChIP and input reads at each region, we set the FDR and Qvalue to 1.0 to ensure loose cut-off and pass most of the reads. Thus, from PeakSeq we have a list of accumulated read coverage for ChIP and input at each region, which we call the raw intensity values. Next, we transform the raw intensity values to a natural log scale and normalize each value by the length of the region. Input to ChIP-GMM then consists of the combined list of log-normalized read intensity values. We then leverage the MATLAB Statistical Parametric Model (spm) tool box [4] to fit a one-dimensional mixture of Gaussian distributions. The model generates the estimated mixing weight, mean, and precision for each component, the hyper-parameters, and the likelihood of the mixing weight for each data point. For every region, we check the components that are likely to generate the ChIP and input signals. If both ChIP and input have the same component, we conclude that the region does not have a peak. If the components are different, the region has a peak. Thus, the existence of peak at a region is based on the hypothesis that ChIP and input should be generated by different Gaussian distributions.

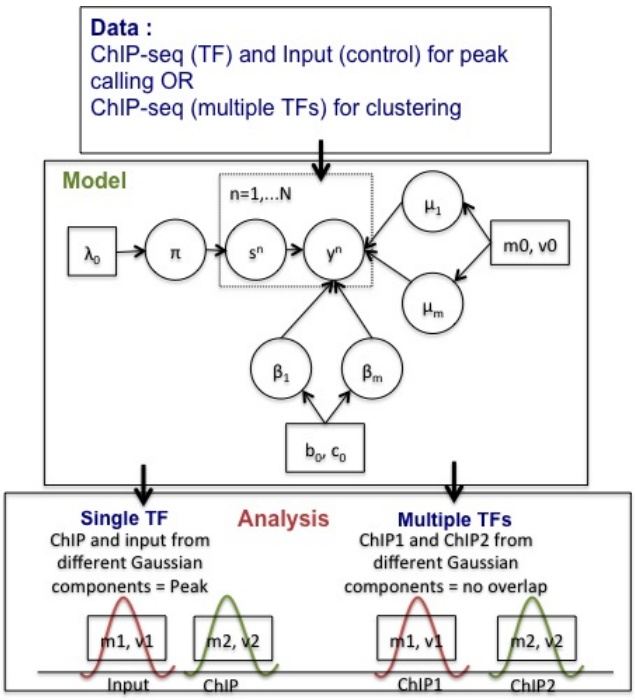

Figure 1: ChIP-GMM model in identifying peaks and co-regulated regions

2.2 Model description of ChIP-GMM

We show that the variational Bayes (VB) mixture model proposed by [4] can both (a) identify binding sites for a single TF and (b) cluster binding sites for multiple TFs. In [9], the authors describe the advantage of a GMM with multiple components compared to one with two components. Additional components that differ in eigenvalues capture the variability of the data better, which is not possible with a two-component GMM. Normalized ChIP signal profiles can be thought of as data points in one-dimensional space $D = y_n$ with $n = 1...N$. Let us consider the Gaussian mixture model to have m components. We denote the mixing probability, mean, and precision (inverse of variance) of each component s by $\pi_s, \mu_s,$ and β_s, respectively. We also introduce a parameter set of vectors $\boldsymbol{\theta} = \boldsymbol{\pi}, \boldsymbol{\mu},$ and $\boldsymbol{\beta}$ corresponding to the mixing weight, mean, and precision, respectively. The likelihood that a data point y_n is generated by the parameters is defined as follows:

$$p(y_n|\boldsymbol{\theta}) = \sum_{s=1}^{m} p(s_n = s|\boldsymbol{\pi}) p(y_n|\beta_s, \mu_s)$$

where s_n is the indicator variable denoting the source of a data point from m components and where $p(s_n = s|\boldsymbol{\pi}) = \pi_s$

The probability for each Gaussian component is defined as

$$p(y_n|\beta_s, \mu_s) = (2\pi)^{-1/2}\beta_s^{1/2} exp(-\frac{\beta_s}{2}(y_n - \mu_s)^2)$$

p($\boldsymbol{\theta}$) is the prior probability on the model parameter set, which can be written as a product of the prior probabilities of mixing weights with Dirichlet distribution (Eq. (1)), mean with a normal distribution (Eq. (2)), and precision with gamma distribution (Eq. (3)).

$$p(\boldsymbol{\pi}) = \frac{\Gamma(m\lambda_0)}{\Gamma(\lambda_0)^m}\prod_{s=1}^{m}\pi_s^{\lambda_0 - 1} \tag{1}$$

$$p(\mu_s) = N(\mu_s; m_0, v_0) \tag{2}$$

$$p(\beta_s) = \Gamma(\beta_s; b_0, c_0) \tag{3}$$

where λ_0 is the prior hyper-parameter of mixing weight $\boldsymbol{\pi}$; m_0 and v_0 are the prior mean and variance hyper-parameters of mean μ_s; and b_0 and c_0 are the prior shape and scale hyper-parameters of precision β_s.

In the prior settings, λ_0, m_0, v_0, b_0, and c_0 are fixed. For mixing weights, we set $\lambda_0 = 1$. For the mean, we use $m_0 = mean(y)$ and $\beta_0 = 1$. For precision, we define $c_0 = 1$ and $b_0 = 0.01$. When data with dimension d is normalized to zero mean and unit variance, then the prior for precision (or variance) can be set as $c0 = d$ and $b0 = 0.01 * d * eye(d)$, where $eye(d)$ is a diagonal matrix with size d x d. Also, the number of components m is fixed and does not vary during iterations.

The joint likelihood of the data and indicator, $p(Y, S|\theta)$, is defined as:

$$p(Y, S|\theta) = \prod_{n=1}^{N} p(s_n| = s|\boldsymbol{\pi})p(y_n| = \beta_s, \mu_s)$$

The posterior distribution of the parameter set, $q(\boldsymbol{\theta})$, is defined by the product of the density of each factor as follows: $q(\boldsymbol{\theta}) = q(\pi)q(\mu)q(\beta)$, and the posterior distribution of the hidden variables are approximated by $q(\boldsymbol{S})$. The Bayes framework maximizes the negative free energy similar to expectation maximization technique. The negative free energy is a composition of the average likelihood and the KL divergence between the prior and the posterior. The E-step and M-step equations are called iteratively and are summarized below.

2.2.1 E-step

In this step, we update the indicator posterior, $\widetilde{\gamma}_s^n$ which consists of the posterior hyper-parameters. Let us define $\widetilde{\gamma}_s^n = p(s_n = s)$. We expand $\widetilde{\gamma}_s^n$ in Eq. (4) using λ_s, which is the posterior hyper-parameter of $\boldsymbol{\pi}$; m_s and v_s, which are the posterior mean and variance hyper-parameters of μ_s, respectively; and b_s and c_s, which are the posterior shape and scale hyper-parameters of β_s.

$$\widetilde{\gamma}_s^n = \widetilde{\pi}_s\widetilde{\beta}_s^{\frac{1}{2}} exp[-\frac{1}{2}\bar{\beta}_s(y_n^2 + m_s^2 + v_s - 2m_s y_n)] \tag{4}$$

where

$$log\widetilde{\pi}_s = \Psi(\lambda_s) - \Psi(\sum_{s'}\lambda_{s'})$$

$$log\widetilde{\beta}_s = \Psi(c_s) + logb_s$$

$$\bar{\beta}_s = b_s c_s$$

Thus, Eq. (5) gives the probability that y_n is generated by s. (Ψ is the digamma function)

$$\gamma_s^n = \frac{\widetilde{\gamma}_s^n}{\sum_{s'}\widetilde{\gamma}_{s'}^n} \tag{5}$$

2.2.2 M-step

In this step, we update the posterior hyper-parameters. The posterior hyper-parameter of mixing weight, λ_s is updated as:

$$\lambda_s = \overline{N}_s + \lambda_0$$

The posterior hyper-parameters of precision, b_s and c_s, are updated as:

$$\frac{1}{b_s} = \frac{N}{2}\widetilde{\sigma}_s^2 + \frac{1}{b_0}$$

$$c_s = \frac{\overline{N}_s}{2} + c_0$$

The posterior hyper-parameters of mean, m_s and v_s, are updated as:

$$\tau_s = \tau_0 + \tau_{data}(s)$$

$$m_s = \frac{\tau_0}{\tau_s}m_0 + \frac{\tau_{data}(s)}{\tau_s}m_{data}(s)$$

where $\tau_0 = \frac{1}{v_0}$ and $\tau_s = \frac{1}{v_s}$ are the prior and posterior precision hyper-parameters, respectively.

$\overline{\pi}_s$ is the proportion of data in component s defined as:

$$\overline{\pi}_s = \frac{1}{N}\sum_{n=1}^{N}\gamma_s^n$$

$\overline{N}_s$ is the number of data points in component s defined as:

$$\overline{N}_s = N\overline{\pi}_s$$

$\overline{y}_s$ are the weighted data points defined as:

$$\overline{y}_s = \frac{1}{N}\sum_{n=1}^{N}\gamma_s^n y_n$$

$\widetilde{y}_s^2$ are the weighted square data points defined as:

$$\widetilde{y}_s^2 = \frac{1}{N}\sum_{n=1}^{N}\gamma_s^n y_n^2$$

$\widetilde{\sigma}_s^2$ is the average variance of component s defined as

$$\widetilde{\sigma}_s^2 = \widetilde{y}_s^2 + \overline{\pi}_s(m_s^2 + v_s) - 2m_s\overline{y}_s$$

The negative free energy F_m is given by

$$F_m = L_{av} - KL(q(\pi)||p(\pi)) - KL(q(\mu)||p(\mu)) - KL(q(\beta)||p(\beta))$$

where the Kullback-Leibler (KL) divergence is defined as a metric to determine the difference between two probability distributions, where $p(.)$ and $q(.)$ denote the prior probability distribution and the posterior distribution of parameters, respectively, and L_{av} denotes the average likelihood.

3 Results

We provide our results in three parts. The first two parts present the performance of ChIP-GMM in peak calling and TF clustering, respectively, and its comparison to existing peak calling tools while the third part presents the functional analysis of the peaks using gene annotation tools.

3.1 Peak calling for single transcription factor

We utilized three ChIP-seq data sets, listed below, to identify peaks for a single transcription factor (TF) and compared the performance of ChIP-GMM to MACS and PeakSeq in terms of average log fold change (FC) and the number of peaks identified. MACS and PeakSeq were run with default parameters.

(1) Two ER+ breast cancer samples: one with favorable response to treatment (SRR1021753) and another with poor response to treatment (SRR1021758) from GSE32222. We refer to these samples as ER_1 and ER_2

(2) JMJD3 and SMAD3 ChIP-seq data from neural stem cells of 12.5 embryonic, day-old, wild-type mice from GSE36673

(3) Znf335a ChIP-seq from lateral telencephalon of E14.5 mice from GSE36386

The threshold for peaks was selected as $5(-log_{10}p - value)$ for MACS and $-5(log_{10}p - value)$ for PeakSeq. Table 1 shows that ChIP-GMM performs favorably with PeakSeq and MACS in terms of fold change peaks in addition to identifying more peaks than either method. These additional regions identified by ChIP-GMM may contain genes functionally relevant to the TF, which are not captured by MACS or PeakSeq.

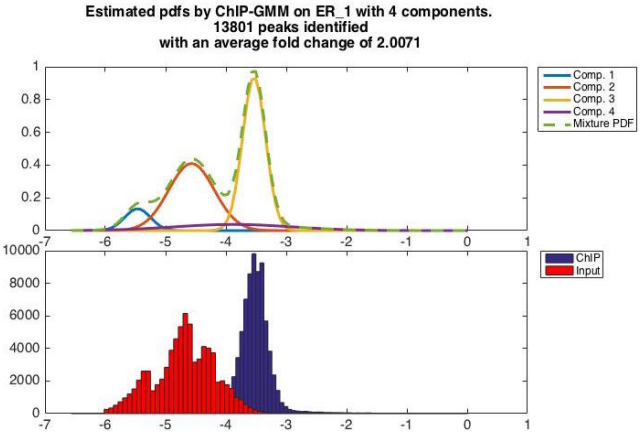

Figure 2: ChIP-GMM results for ER_1

Table 1: ChIP-GMM vs. other peak calling methods

Data	ChIP-GMM		PeakSeq		MACS	
	FC	Peaks	FC	Peaks	FC	Peaks
ER_1	2.01	13,801	2.16	3,485	2.62	2,723
ER_2	1.90	12,133	1.60	3,980	2.21	2,467
JMJD3	1.75	52,997	1.47	28,259	1.94	15,366
SMAD3	1.80	14,392	1.44	38,350	1.89	21,832
Znf335a	2.11	1,515	2.18	367	2.13	99

We show the estimated component pdfs and histogram of log-normalized ChIP and input intensities of ER_1 in Figure 2. As can be seen, the intensity values of ChIP are larger than the input. We also illustrate the distribution of log-normalized ChIP and input reads for ChIP-GMM estimated peaks ER_1 in Figure 3. It can be seen from the p-values that the ChIP read intensities are significantly higher than the input read intensities for the peaks identified by ChIP-GMM. This validates the hypothesis of ChIP-GMM that in order to have a peak at a particular region,

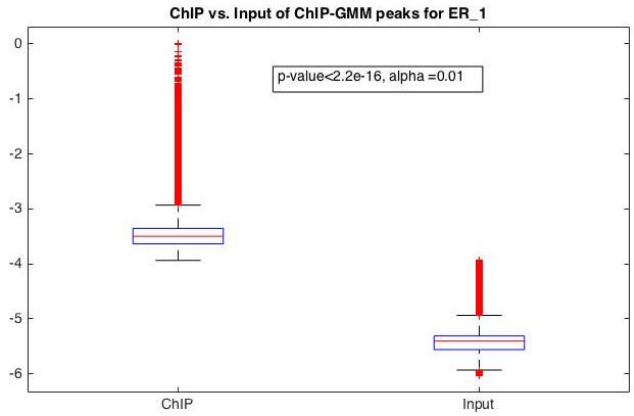

Figure 3: ChIP vs. input for ER_1

ChIP and input are generated from different Gaussian distributions.

3.2 Clustering for multiple TFs

We applied ChIP-GMM to cluster genomic regions co-regulated by multiple TFs for two cases. For the first case, we used the peak files of JMJD3 and SMAD3 produced by PeakSeq with default parameters as input to ChIP-GMM. An important aspect of ChIP-GMM is that it has different modes of operation between peak calling and TF clustering. For a single TF, a region will have peak if ChIP and input are generated by different Gaussian components. For TF clustering, an overlapped region will have reads from different TFs generated by the same Gaussian component. Thus we first list regions from both TFs having the same Gaussian component. Then we calculate the overlap among those regions. For each component, we have a list of overlapping regions. We then combine the regions for all components to obtain the final set of co-regulated regions. From PeakSeq, there were 28259 peaks for JMJD3 and 38350 peaks for SMAD3. Upon clustering, ChIP-GMM identified 15041 common regions.

We then compared the results with MACS2 in differential binding mode and with Signal Spider. With MACS2, for each treatment condition we first ran 'predictd' and averaged the fragment length from the two conditions. We used this averaged length in the 'callpeak' stage. Finally we used the pileup .bdg files and control .bdg files and ran 'bdgdiff'. For Signal Spider, we transformed the ChIP read-intensity values (from PeakSeq) to log scale. MACS2 and Signal Spider identified 6575 and 19175 common regions, respectively.

For the second case, we selected AUTS2, BMI1, P300 and RING1B from GSE60409 [2]. Using PeakSeq we obtained 172983, 76454, 377272 and 53651 peaks for AUTS2, BMI1, P300 and RING1B, respectively.

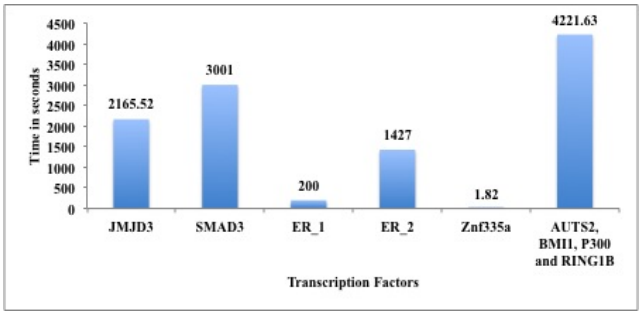

Figure 4: Time complexity of ChIP-GMM

Following the steps from the first case, we applied ChIP-GMM to find common regions regulated jointly by the fours TFs. ChIP-GMM identified 2508 common regions while Signal Spider identified 5305 common regions (numOfCombinations = 4,numOfBindingModes = 3). We compared ChIP-GMM to Signal Spider only for this case because both methods can handle multiple ChIP profiles (unlike MACS2 which can only be used to analyze a pair of TFs).

As shown in Figure 4, the execution time of ChIP-GMM depends on the size of the data set and ranges from a few seconds to over sixty minutes.

3.3 Binding regions to genes

In order to identify the genes regulated by TFs, we used the peaks identified by ChIP-GMM from ER_1 on GREAT [3] as a test case for functional analysis. We prepared a bed file with the peak regions as the test regions for GREAT, selected the whole genome (hg18) as the binding regions and then selected single nearest gene within 1.5 Kb in annotating genomic regions to genes.

The first list in Table 2 is a set of 96 genes with a significant binomial raw p-value and fold enrichment (FE) of 3.43. These genes are typically down-regulated in basal sub-type of breast cancer. Basal sub-type tumors have ER-negative status compared to the Luminal A and Luminal B tumors that have ER-positive status. The second list of 37 genes are up-regulated in Luminal B sub-type. These two sets indicate that the identified peak regions for ER contain genes that might help to classify ER+ vs. ER- tumor sub-types leading to the development of targeted drug therapy. In addition the motifs have been targets of miR-19a and miR-19b which are known tumor suppressors in breast cancer.

4 Conclusion

We present ChIP-GMM, a variational Bayes framework on a Gaussian mixture model that can be used

Table 2: Functional analysis of ER_1 using GREAT

Type of Genes	**Gene Count**	**p-val**	**Fold Change**
Down-regulated in basal subtype of breast cancer samples	96	3.7e-27	3.43
Up-regulated in the luminal B subtype of breast cancer	37	1.3e-17	5.34
Between two groups of breast cancer according to basal (ESR1- AR-) and luminal (ESR1+ AR+)	50	4.1e-16	3.68
Up-regulated in breast cancer samples positive for ESR1 compared to the ESR1 negative tumors	22	8.9e-12	5.28
Targets of MicroRNA TTTGCAC,MIR-19A,MIR-19B	52	8.2e-11	2.79

both to identify protein-DNA binding sites as well as clustering genomic regions that are co-regulated by multiple TFs. An important difference between our proposed method and existing peak calling methods is in the selection of peaks. Based on the hypothesis that log-normalized ChIP and input read intensities should be generated from different Gaussian distributions, the method determines the existence of a peak if the ChIP and input are assigned to different Gaussian components. ChIP-GMM also clusters co-regulated genomic regions using the reverse hypothesis that for overlapped genomic regions ChIP signals for each TF are likely to be generated by the same Gaussian distribution. Analysis of ChIP-GMM peaks for ER+ data revealed genes that are known to be up/down-regulated in breast cancer. ChIP-GMM performs favorably to others in terms of fold enrichment of peaks. Future work might use a Dirichlet prior mixture model to auto-select the number of Gaussian components in the mixture.

5 Acknowledgment

This research was supported in part by a Virginia Tech ICTAS grant for interdisciplinary research.

References

[1] Xi Chen, Jin-Gyoung Jung, Ayesha N Shajahan-Haq, Robert Clarke, Ie-Ming Shih, Yue Wang, Luca Magnani, Tian-Li Wang, and Jianhua Xuan. Chip-bit: Bayesian inference of target genes using a novel joint probabilistic model of chip-seq profiles. *Nucleic acids research*, 44(7):e65–e65, 2016.

[2] Zhonghua Gao, Pedro Lee, James M Stafford, Melanie von Schimmelmann, Anne Schaefer, and Danny Reinberg. An auts2-polycomb complex activates gene expression in the cns. *Nature*, 516(7531):349–354, 2014.

[3] Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. Great improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5):495–501, 2010.

[4] William Penny. Variational bayes for 1-dimensional mixture models. `http://www.fil.ion.ucl.ac.uk/~wpenny/publications/vbmog.ps`, 2000.

[5] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nature biotechnology*, 27(1):66–75, 2009.

[6] Ka-Chun Wong, Yue Li, Chengbin Peng, and Zhaolei Zhang. Signalspider: probabilistic pattern discovery on multiple normalized chip-seq signal profiles. *Bioinformatics*, 31(1):17–24, 2015.

[7] Haipeng Xing, Yifan Mo, Will Liao, and Michael Q Zhang. Genome-wide localization of protein-dna binding and histone modification by a bayesian change-point method with chip-seq data. *PLoS Comput Biol*, 8(7):e1002613, 2012.

[8] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):1, 2008.

[9] Daniel Zoran and Yair Weiss. Natural images, gaussian mixtures and dead leaves. In *Advances in Neural Information Processing Systems*, pages 1736–1744, 2012.