Summer 2012

A Statistical Model to Determine Multiple Binding Sites of a Transcription Factor on DNA Using ChIP-seq Data

Rasika Jayatillake
Old Dominion University

A STATISTICAL MODEL TO DETERMINE MULTIPLE BINDING SITES OF A TRANSCRIPTION FACTOR ON DNA USING CHIP-SEQ DATA

by

Rasika Jayatillake
B.Sc. September 2005, University of Colombo, Sri Lanka
M.S. May 2010, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

MATHEMATICS AND STATISTICS

OLD DOMINION UNIVERSITY
August 2012

Approved by:

Nak-Kyeong Kim (Director)

N. Rao Chaganty (Member)

Dayanand Naik (Member)

Jing He (Member)

ABSTRACT

A STATISTICAL MODEL TO DETERMINE MULTIPLE BINDING SITES OF A TRANSCRIPTION FACTOR ON DNA USING CHIP-SEQ DATA

Rasika Jayatillake
Old Dominion University, 2012
Director: Dr. Nak-Kyeong Kim

Protein-DNA interaction is vital to many biological processes in cells such as cell division, embryo development and regulating gene expression. Chromatin Immunoprecipitation followed by massively parallel sequencing (ChIP-seq) is a new technology that can reveal protein binding sites in genome with superior accuracy. Although many methods have been proposed to find binding sites for ChIP-seq data, they can find only one binding site within a short region of the genome. In this study we introduce a statistical model to identify multiple binding sites of a transcription factor within a short region of the genome using the ChIP-seq data. Mapped sequence reads from the ChIP-seq experiments are modeled as the sum of observations from unknown number of Poisson distributions. The rate parameters of these Poisson distributions are considered as a function of the underlying distribution of the tags that depends on the locations of the binding sites and their intensity parameters. For the parameter estimation of the model, two major approaches are discussed: one is a Bayesian method, the other, the EM algorithm. For the Bayesian method the reversible jump Markov chain Monte Carlo (RJMCMC) method is used for computation. An extensive simulation study was performed for the selection of proposal methods and priors in RJMCMC as well as for the comparison of model selection criteria in the EM algorithm. Real ChIP-seq datasets for transcription factors STAT1 and ZNF143 were used to demonstrate the performance of the proposed model. The results from the multiple binding sites model were compared with existing peak-calling programs.

DEDICATED TO MY PARENTS.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my thesis advisor Dr. Nak-Kyeong Kim. I am grateful for the opportunity to work as his student and engage in research in the application of statistics in genomics study. His guidance, support and patience was instrumental in completing this research and thesis.

I am very much grateful to Dr. Rao Chaganty for serving as a member of my dissertation committee. Moreover, I am deeply grateful to him for his inspirational guidance and kind support as the graduate program director for Statistics throughout my graduate study.

I am also grateful to Dr. Dayanand Naik for his support and advice not only as a dissertation committee member but also as a lecturer throughout my graduate study. I also wish to thank Dr. Jing He for her invaluable support and for serving as a member of my dissertation committee.

I am grateful to Dr. Mark Dorrepal, Chair of the Mathematics and Statistics department and Dr. Richard Noren, the graduate program director and all the professors of the department for their kind support and encouragement throughout the graduate program. I also thank all the staff members for their generous support in numerous ways.

Last but not least, I thank all my friends for their loving and caring friendship, continuous encouragement, and inspiring discussions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

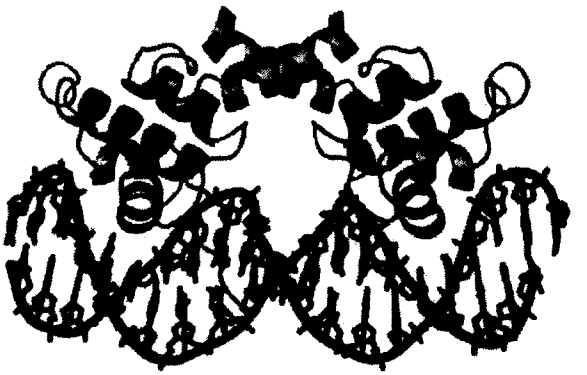
1.1 DNA-PROTEIN INTERACTION

Genomic studies have become instrumental in investigation of many biological processes. With evolving technologies, these studies generate massive volumes of data that can be analyzed effectively using statistical methodologies. In this thesis we present a statistical model to analyze data from ChIP-seq (Chromatin Immunoprecipitation followed by sequencing) experiments to identify multiple binding sites of a transcription factor protein on short regions of DNA (Deoxyribonucleic acid).

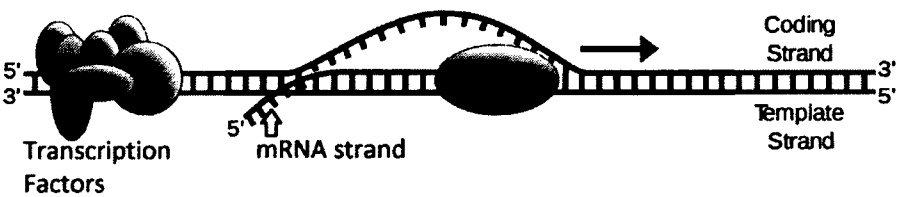
Protein-DNA interaction is vital to many biological processes in cells, especially in regulating gene expression (Semenza 1998, Fields 2007, and Park 2009). Gene expression is the process of using information in DNA to synthesize proteins and other gene products such as RNA(Ribonucleic acid). Although many other factors contribute to the regulation of genes, it is mainly controlled at the transcription phase by a specific type of proteins referred to as transcription factors (Semenza 1998). These transcription factors (TFs) bind to specific regulatory sequences of the DNA. They then control the gene expression by either promoting or suppressing transcription. In this thesis these specific DNA sequences are referred to as *transcription factor binding sites* (TFBSs) and the binding of the protein on DNA is referred to as a *binding event.* The process of transcription, which includes the interaction of TFs and their impact, is complex and details of the process can be found in Semenza (1998) and Yilmaz and Grotewold (2010). When a TF is bound to DNA, it interacts with other DNA bound TFs and mediates the RNA polymerase, an enzyme, to bind to the promoter region of the DNA. Once the RNA polymerase is bound, it traverses through the DNA segment, usually a gene, shearing its double helix structure (Yilmaz and Grotewold 2010). Concurrently it reads the base pairs of the template strand of the DNA and generates a messenger RNA, a complementary copy of the DNA sequence read. This process of generating the mRNA is called

This dissertation follows the style of *Journal of the American Statistical Association.*

transcription. In a later phase, known as *translation*, the mRNA is used to synthesize proteins. Figure 1 gives a graphical representation of a protein bound to the DNA and the transcription process. In humans it is estimated that there are about 3000 TFs that are responsible for controlling gene expressions (Babu et al. 2004).



(a) Protein bound to DNA.



(b) Transcription process controlled by transcription factors.

Figure 1. Protein-DNA interaction.

1.2 CHROMATIN IMMUNOPRECIPITATION FOLLOWED BY SEQUENCING (CHIP-SEQ) METHOD

There are many techniques developed to study protein-DNA interaction. Among these, Chromatin Immunoprecipitation (ChIP) followed by genomic tiling microarray hybridization (ChIP-chip) and ChIP followed by massively parallel sequencing (ChIP-seq) are two of the most commonly used approaches. The ChIP-seq method (Johnson et al. 2007) has several advantages over the ChIP-chip method. The following list contains some facts that have lead to the rapid adaptation of the ChIP-seq

method over ChIP-chip (Park 2009 and Ho et al. 2011):

- Ability to produce profiles with higher spatial resolution, dynamic range, and genomic coverage.
- Can be used to analyze virtually any species with a sequenced genome, since it is not constrained by the availability of an organism-specific microarray.
- Can work with a smaller amount of initial material.
- More cost effective as the sequencing techniques continues to be cheaper.

Ideally the identification of these binding sites under various conditions and for all the different TFs need to be performed using biological or biochemical experiments. However, these experimental techniques are yet to be matured. Therefore, predicting the TFBSs relies on statistical models that use data from available techniques such as ChIP-seq and ChIP-chip. These models may reveal combined binding sites of a transcription factor and its co-factors. They identify binding sites in species for which experimental binding data is not available, and explain variation in binding affinities that can have a functional effect (Reid et al. 2010). Therefore, in this thesis we present a statistical model to analyze ChIP-seq data. In the next section we present a brief overview of the ChIP-seq method to better understand the characteristics of ChIP-seq data.

1.2.1 WORK FLOW OF CHIP-SEQ APPROACH

The ChIP-seq method consists of several steps as illustrated in Figure 2. The steps can be categorized into two main parts: ChIP and sequencing. Following is a summary of these steps (Fields 2007, Park 2009, and Kuznetsov, Singh, and Jenjaroenpun 2010).

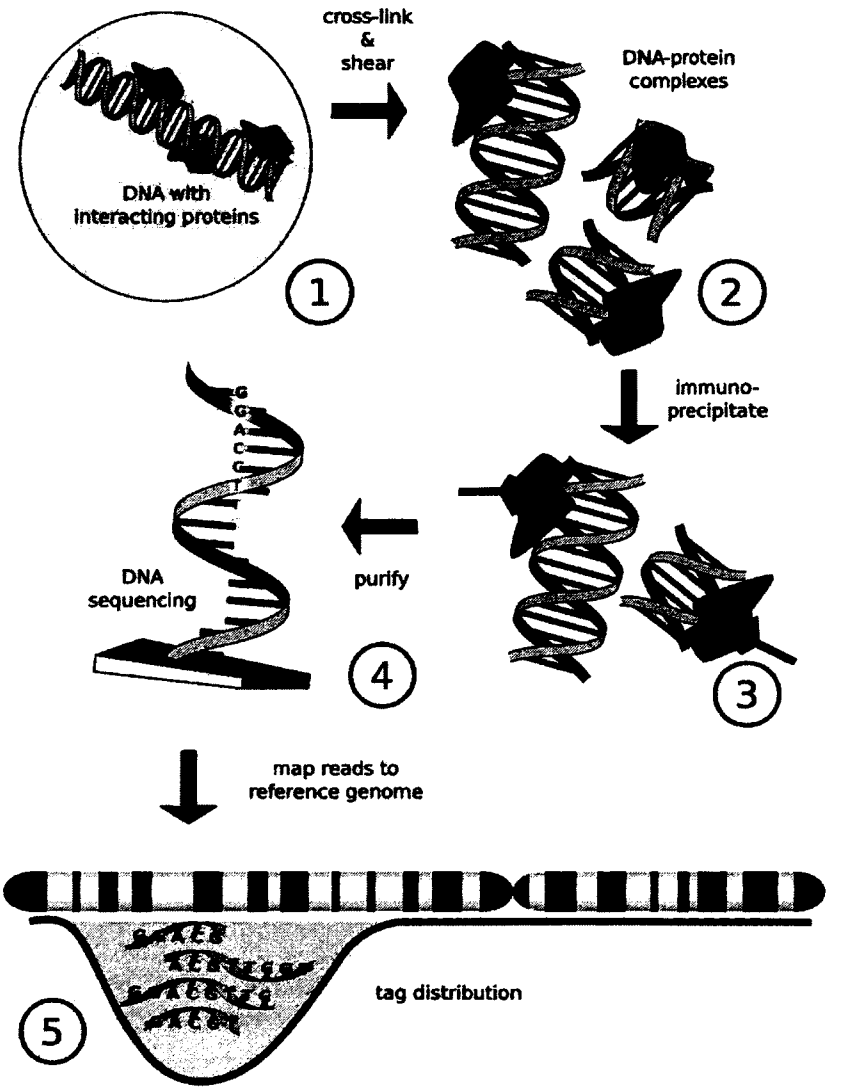


Figure 2. Workflow of ChIP-seq.

- ChIP

  1. Using about 10^8 cells, the transcription factors are cross-linked to their DNA by treating the cells with formaldehyde.
  2. Cells are lysed and the chromatin is isolated. The DNA is sheared into small fragments by ultrasound sonication.
  3. The protein and its associated DNA fragments are isolated by using a protein specific antibody. The DNA fragments are separated from the protein by reverse cross-linking.

- Sequencing

  4. The released DNA fragments are directly sequenced in series of 20~80 bp reads producing millions of short read sequences.
  5. Short read sequences are mapped back to a reference genome.

After collecting these short sequence reads, they are mapped to a reference genome using a mapping software/algorithm, such as MAQ (Li, Ruan, and Durbin 2008) and Bowtie (Langmead et al. 2009). Mapped reads are usually referred to as tags (Jothi et al. 2008). Most of the existing ChIP-seq data are generated by the Illumina Genome Analyzer, but other platforms such as SOLiD and Helicos are also available and can generate 100-400 million tags in a single run and 60-80% of these tags can be mapped uniquely to the genome (Park 2009). When the sequences are mapped, peaks of tags can be observed over the genome. These peaks may correspond to the protein-DNA binding sites.

1.3 FEATURES AND CHALLENGES IN ANALYZING CHIP-SEQ DATA

In this section we discuss some features and challenges associated with the analysis of ChIP-seq data.

1. Mappability

   For further analysis, raw sequence reads from a ChIP-seq experiment are mapped back to the genome. These sequence reads can be mapped to unique segments of the genome (with or without several mismatches), or they can be

mapped to more than one segment of the genome. The segments of the genome that the tags cannot be mapped uniquely are referred to as *unmappable* and segments that can be uniquely mapped are referred to as *mappable.* In the analysis if one decides to use only the uniquely mapped tags, some true sites will be invisible because they are located in repeats or recent duplicated region (Pepke, Wold, and Mortazavi 2009). On the other hand, including reads with multiplicity and multi-reads can increase false positive peaks.

2. Strand specific information

   The backbone of the double helix structure of the DNA is made from alternating phosphate and sugar bases. These alternating bonds between the sugar bases and phosphate bases gives each of the two strands of the DNA directionality. In the double helix structure the direction of the nucleotides in one strand is opposite to their direction in the other strand. These asymmetric ends of the DNA strands are called 5′ (five prime) and 3′ (three prime) ends. The 5′ end have a terminal of phosphate base whereas the 3′ end have a sugar base. The DNA strand that has the directionality of 5′-3′ is referred to as left (positive, forward) strand and the other with directionality of 3′-5′ is referred to as right (negative, backward) strand. As given by Park (2009), when the DNA sequence reads are mapped to the genome they results in two peaks, one on each strand (see Figure 3). Furthermore, in ChIP-seq experiments the DNA fragments are sequenced from 5′ end. Therefore, when the sequence reads are mapped to the left strand, they are mapped with a shift to the left from the binding cross-link and when they are mapped to the right strand, they are mapped with a shift to the right from the binding cross-link. In the analysis of the data, one could consider combining the two peaks from the two strands or use this directionality information to detect the locations of the peaks with higher precision.

3. Background noise

   ChIP-seq involves background noise which results in spikes of tag counts due to factors other than protein-DNA binding. Kharchenko, Tolstorukov, and Park (2008) describe some of these background noises and their causes. The first are singular peaks of tag density at a single chromosome position that is due to the non-uniform shearing of DNA around chromosomes. The

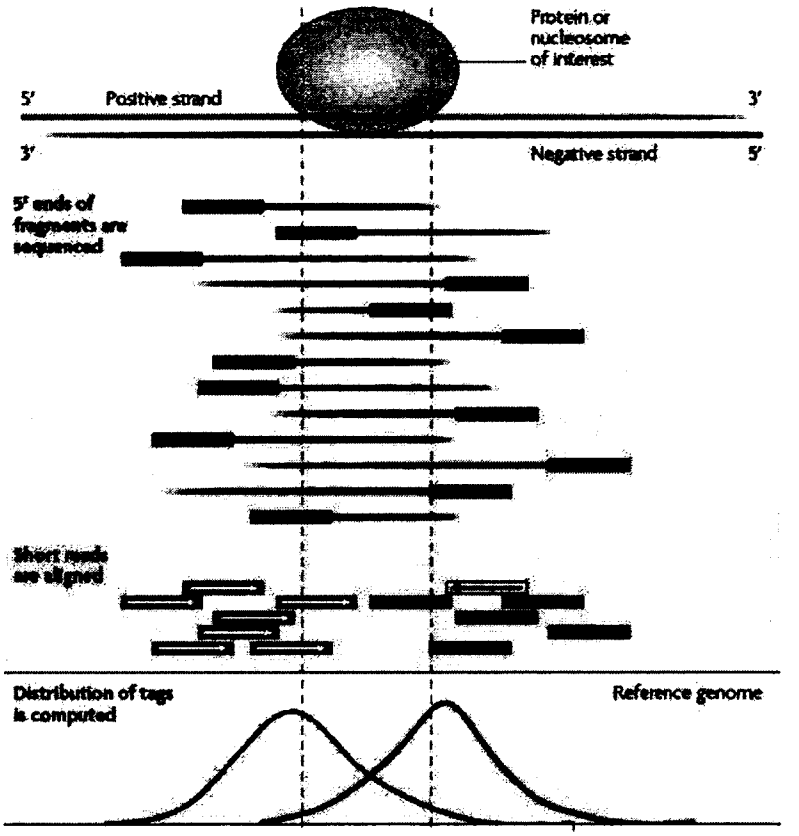


Figure 3. Shifted tag peaks on the left (positive) and the right (negative) strands (Park 2009).

second are non-uniform wide clusters of increased densities and the third are small clusters of strand specific tag density resembling the pattern of protein-binding site but with smaller separation between strand peaks.

4. Ranking peaks

   Due to the varying strength of the protein-binding, the height (intensities) of the peaks will vary. In addition to detecting the tag peaks, it is also necessary to determine the intensities of the peaks. These intensities can then be used to rank or score the peaks allowing more in-depth analysis.

1.4 OVERVIEW OF EXISTING PEAK-CALLING ALGORITHMS

Since the introduction of ChIP-seq method many algorithms and software programs were developed to detect protein-DNA binding sites. A list of available programs and evaluations of these methods are given by Wilbanks and Facciotti (2010), Pepke, Wold, and Mortazavi (2009), and Laajala et al. (2009). According to these evaluations, no single method has a significant advantage over others and different methods perform in varying degree of precision for different experimental data. Therefore, developing new approaches to detect TFBS is an ongoing research and new methods continue to be developed and published by the scientific community. In this section, we present an overview of some popular methods in finding tag peaks.

- Hidden Markov model based Peak-finding algorithm (HPeak)
  HPeak (Qin et al. 2010) method introduced by Qin et al. (2010) is based on a Hidden Markov model (HMM). This procedure has four main steps. In the initial step, it imports genomic coordinates of all mapped sequenced reads. The short reads are extended directionally from its start position to form a hypothetical DNA fragment (HDF), mimicking the ChIP-DNA fragment from which the sequencing read was generated. In the second step, the entire mapped genome is partitioned into small bins of fixed length (25 bp) and counts the HDF's that fall within the bins. In the third step, the two state HMM is applied to the HDF counts to distinguish blocks of consecutive ChIP enriched bins from the background. Authors have chosen the HMM approach due to the observed strong correlation of HDF coverage in adjacent bins. Due to the wide dynamic range of the ChIP-seq data, the number of HDF's falling into the ChIP-enriched bins varies dynamically and show significant over-dispersion. Therefore, a generalized Poisson (GP) distribution that accounts the over-dispersion is used in estimating the emission probabilities of the HMM model. Since there are more empty bins in background data, zero inflated Poisson (ZIP) distribution is used for control data. For experiments with control data and experimental data, the authors have used the bivariate GP and ZIP. Parameters of the HMM are estimated using the Viterbi algorithm.

- Quantitative Enrichment of Sequence Tags (QuEST)
  QuEST (Valouev et al. 2008) starts analyzing the data by constructing two profiles for the left and right strands. These profiles are fitted using the kernel

density estimation (KDE) method with the Gaussian kernel density. The distance between the adjacent tag peaks of the two strands is estimated by using a particularly robust subset of the data and the half distance between the two peaks is referred to as *peak shift*. Then the left and the right profiles are combined through out the genome to form the combined density profile (CDP) by shifting the tags by *peak shift*. QuEST then searches the CDP's for enriched loci as positions in the genome corresponding to local maxima of the CDP with sufficient enrichment compared to the background. Thus, a threshold for the peak calling is required. To determine the threshold, the negative control data is separated into two sets. One is used as a pseudo-ChIP sample in which peaks are to be determined and the other is used as the background data for the sample. Any peak that is predicted in this comparison is considered as false positive. The false discovery rate (FDR) is calculated as the ratio of the number of peaks predicted in the pseudo-ChIP sample to the number of peaks identified in the real ChIP experiment data. This allows users to set specific values for thresholds or vary the threshold until a desired FDR is achieved. QuEST reports a score quantifying the tag enrichment at the peak and uses it to rank the peaks.

- Site Identification from Short Sequence Reads (SISSRs)
  In SISSRs (Jothi et al. 2008), the entire genome is scanned using a window of size w (20 bp wide) with an overlap of w/2. The net tag count is computed by subtracting the number of antisense (left strand) tags from the number of sense (right strand) tags. Each time the net tag count changes from positive to negative, that location (t) is considered as a candidate for binding location. For these locations to be confirmed, the number of tags (p) in the right strand between $[t - F, t]$ must be at least E, the number of tags n in the right strand between $[t, t + F]$ must be at least E and the total tags $p + n$ must be at least R. The value of R is estimated with the user defined FDR.

- Model-based Analysis of ChIP-Seq data (MACS)
  MACS (Zhang et al. 2008) takes the advantage of the bimodal pattern on the left and the right strand to empirically model the shifting size to better locate the binding site. Therefore, the initial step is an estimation of the peak shift. Given a sonication size/bandwidth and a high-confidence fold-enrichment

(mfold), MACS slides a window of size of 2 times the bandwidth size across the genome to find regions with tags counts exceeding mfold times the counts observed under random distribution. MACS randomly samples 1000 of these high quality peaks and separates their left and right strand peaks. These peaks are aligned by the midpoint between the left and right strand peak centers if the left strand peak is to the left of the right strand peak. The distance between the modes of the left and the right strand peaks in the alignment is defined as "d", and all the tags are shifted by d/2 toward the 3′ end and use the shifted tags for peak detection. In peak detection, for experiments with control data, MACS linearly scales the total control tag count to be the same as the total ChIP count. Some of these tags, that may be sequenced repeatedly more times than expected from a random genome-wide tag distribution, are removed. For the shifted tags, MACS count the number of tags by using a sliding window of two bandwidths across the genome to locate candidate sites with significant enrichment based on a Poisson distribution p-value that depends on the background rate. Overlapping enriched peaks are merged and each tag position is extended by d bases from its center. The tag distribution along the genome is modeled by a Poisson distribution. In the control sample, the fluctuations that are often observed are accommodated by using a dynamic rate parameter for the Poisson distribution defined for each peak.

- SPP

  As mentioned in section 1.2, ChIP-seq reads are mapped to the genome in varying accuracy. SPP (Kharchenko, Tolstorukov, and Park 2008) method use the length of the matched read and the number of nucleotides covered by mismatches and gaps to classify the quality of the tag alignment. This method then uses the strand cross-correlation profile to decide whether to include that tag in the analysis. The strand-cross correlation is the Pearson correlation coefficient between genome-wide profiles of tag density of left and right strands, shifted relative to each other by a specific distance. The SPP adjusts for background anomalies by removing extremely deviated peaks and subtracting the re-scaled background tag density. Within the SPP algorithm there are two main sub-methods that are used in peak detection; they are the WTD (widow tag density) method and the MTC (mirror tag correlation) method. The WTD method scores peaks based on strand-specific tag counts

upstream and downstream of the examined position. The MTC method scans the genome to identify positions exhibiting pronounced positive and negative strand tag patterns that mirror each other. The statistical threshold for FDR is obtained by accounting for the degree of clustering present in the background. Authors have used a randomization that maintains tag occurring at the same or nearby positions together, instead of assigning them independent positions as in the Poisson model.

- CisGenome

  CisGenome (Ji et al. 2008) first calculates FDR that will later be used in peak detection. When only ChIP-seq experiment data is present and control data is unavailable, it computes FDR by dividing the genome into non-overlapping windows of length w (100 bp). Then the number of tags, n_i, in each i^{th} window is counted. The tag counts are modeled as $n_i|\lambda_i \sim Poisson(\lambda_i)$, $\lambda_i \sim Gamma(\gamma, \delta)$ and $n_i \sim Negative\ Binomial(\alpha, \beta)$. Parameters α and β are estimated by fitting a negative Binomial distribution to the number of windows with small number of tag counts ($\leq$ 2 reads). This estimated null distribution is then used in computing the FDR for each level of read counts. In the presence of control data, also referred to as negative control sample, the genome as in the previous case, is divided into windows of size w. For each window i, the number of reads k_{1i} in ChIP sample, number of reads k_{2i} from the control sample and the total count $n_i = k_{1i} + k_{2i}$ is counted. Authors assume that $k_{1i}|n_i \sim Binomial(n_i, p_0)$. The parameter p_0 is estimated from windows with small total counts and uses it to estimate the FDR associated with each level of n_i and k_{1i}/n_i. Binding regions are detected by scanning the genome using a sliding window of width w to detect the windows with FDR smaller than user specified cut-off. Overlapping windows are merged and the minimum FDR among the merged windows is considered as the FDR for the merged windows. For each window, fold enrichment score is also computed by $\frac{y_i+1}{r_0 Z_i+1}$, where y_i is the number of ChIP tags, z_i is the number of control tags, and $r_0 = \frac{p_0}{1-p_0}$.

When considering the given peak calling methods most of them detect peaks using sliding widow and counting the number of tags within the window. For the

calculation of FDR, most methods use a Poisson model or negative binomial distributions. Furthermore, these methods cannot identify multiple binding sites separated by short distances.

1.5 CHIP-SEQ DATASETS

The model presented in this thesis is applied to two published ChIP-seq datasets for two transcription factors: STAT1 and ZNF143. STAT1 belongs to the Signal Transducer and Activator of Transcription (STAT) family of proteins that regulate many aspects of growth, survival, and differentiation in cells. The raw data or unmapped sequenced reads for STAT1 (Robertson et al. 2007) were mapped to the human genome (NCBI Build 36.1) using Bowtie (Langmead et al. 2009) software. When mapping reads, only the sequences with 27 bp length were used enabling the *mappability* information to be used in the analysis. Furthermore, mismatches up to 2 were used in the analysis as long as it produced a unique mapping in the genome. There were about 15.1 million mapped tags from this dataset.

Zinc finger protein 143 (ZNF143) is a transcription factor that positively regulates many cell-cycle-associated genes and is highly expressed in multiple solid tumors (Izumi et al. 2010). The raw sequences from the ZNF143 ChIP-seq dataset (Wanga et al. 2011) were approximately 36 bp in length and there were about 27 million mapped tags from this dataset.

1.6 ASSESSING THE BINDING SITES IDENTIFIED FROM PEAK CALLING ALGORITHMS

One drawback in assessing binding sites for transcription factors is that there are no complete lists of true binding sites for the transcription factors including STAT1 and ZNF143. However, these transcription factors have known binding motifs. Motifs are DNA sequence patterns that characterize binding sites. By scanning through the genome using a position specific scoring matrix (PSSM) of the motif, hits or matches for the motif can be collected. These hits are called motif sites. We could use these motif sites as surrogates of true binding sites and assess the estimated or predicted binding sites by the peak calling programs (Wilbanks and Facciotti 2010). This

has been the standard method of assessing peak calling algorithms. Therefore, the estimated or predicted binding sites by the proposed model is validated with respect to motif sites.

However, it should also be noted that a single transcription factor may have more than one motif. An alternative approach, as conducted by Valouev et al. (2008), is to enter a long sequence (about 200 bp) around the identified peaks to a canonical motif search algorithm and detect the percentages of known motifs present within these sequences. This method of assessment can also reveal a new motif that is not experimentally verified but is significantly detected by the peak calling algorithms.

1.7 ORGANIZATION OF THE THESIS

This thesis is organized into five chapters. The second chapter introduces and describes the normal-exponential model for the tag distribution in the presence of a protein binding event. The Poisson model for a single binding event as well as its extension to multiple binding events are also described in details in this chapter. Chapter 2 also gives details of the simulated datasets that are used in the subsequent chapters. Chapter 3 gives a Bayesian model for estimating parameters of the model. This chapter also contains the results obtained from the simulated datasets using the RJMCMC scheme as well as a discussion on the limitations and strengths of the RJMCMC scheme based on the simulated results. Section 3.6 contains the results obtained by applying the RJMCMC scheme on STAT1 and ZNF143 ChIP-seq datasets. Details of the application of the EM algorithm in estimating the parameters of the multiple binding sites model is given in chapter 4, where the results from simulation studies as well as the results from the STAT1 and ZNF143 ChIP-seq datasets are also given in details. In chapter 4, we also present a comparison of performances of the multiple binding sites models with other existing peak calling algorithms.

CHAPTER 2

BASIC STATISTICAL MODEL FOR THE CHIP-SEQ DATA

Although experimental techniques to determine DNA binding information of various transcription factors are being developed at a rapid speed they are a long way from determining the binding sites for all transcription factors in all conditions (Reid et al. 2010). Therefore, statistical and heuristic models for predicting TFBSs are vital in the advancement of the studies of transcription regulation and in the construction of gene pathways.

The objective of this thesis is to present a more sophisticated statistical model to estimate the TFBSs of a given transcription factor. In this chapter we present the derivation of this statistical model that can estimate multiple binding sites within a short region of the genome using the ChIP-seq data. In section 2.1, we present some characteristics of the ChIP-seq data that lead to the introduction of the dual normal-exponential model for the underlying distribution of the observed tags. The Poisson model for a single and multiple binding events are described in sections 2.2 and 2.3 respectively. In the final section we describe the simulation datasets that are used in chapters 3 and 4 for assessing the performance of the model.

2.1 BASIC MODEL FOR THE CHIP-SEQ TAG DISTRIBUTION

The data that is considered in this study are the mapped sequence reads from the ChIP-seq experiment, where main variables of interest are the mapped location of the tags on the genome, the strand information and number of mismatches in the tag alignment to the genome. In the analysis rather than considering the genome as a whole, we partitioned it into approximately 150$\sim$1500 bp long regions. We first fit a statistical model for the mapped tags in these shorter regions. Subsequently, this model can be applied to all the partitioned regions of the genome. Since a ChIP-seq experiment generates millions of short sequence reads, it is expected that

there will be multiple tags mapped to the same location of the genome. Also it is expected that significant number of tags will be around DNA-protein binding sites. Figure 4 illustrates the tag distribution in a region starting at position 22122563 of chromosome 3 for STAT1 ChIP-seq data. Here we can clearly observe (as described in section 1.3) two peaks of tags mirroring each other; one on the left strand and the other on the right strand.

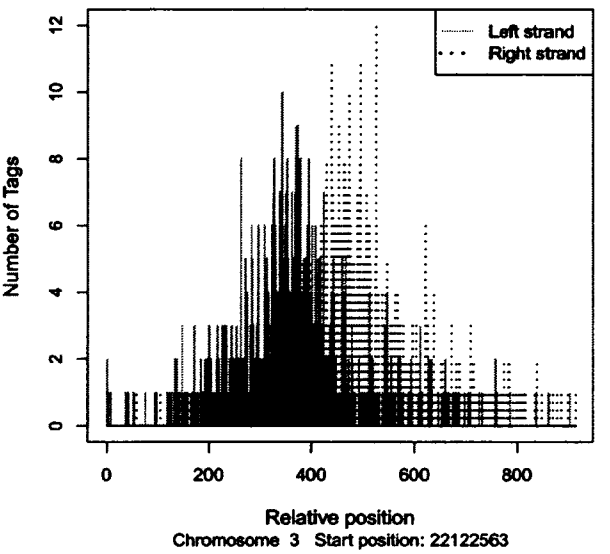


Figure 4. Distribution of tags in a genomic region from STAT1 ChIP-seq data.

To formulate our model let $x_{ij}^L \in \{1, ..., w_i\}$ be the j^{th}, $j = \{1, ..., n_i^L\}$, *mappable* tag location relative to the start location of the i^{th} region, $i = \{1, ..., N\}$, of the left strand. Here, w_i is the width of the region, and n_i^L is the number of *mapped* tags in the region. Similarly, let $x_{ij}^R \in \{1, ..., w_i\}$ be the j^{th}, $j = \{1, ..., n_i^R\}$ *mapped* tag location of the i^{th} region, $i = \{1, ..., N\}$, of the right strand where, $j = \{1, ..., n_i^R\}$ and n_i^R is the number of *mapped* tags in the region. Note that n_i^L and n_i^R are not the same in general, that is, n_i^R and n_i^L are not observed in pairs. Furthermore, to incorporate the *mappability* information described in section 1.3, let X^0 denote the *unmappable* locations in the region. Therefore, the observed tags for a given short region i can be represented by the vector $\mathbf{X} = (x_{i1}^L, x_{i2}^L, ..., x_{in_i^L}^L, x_{i1}^R, x_{i2}^R, ..., x_{in_i^R}^R, X^0)$.

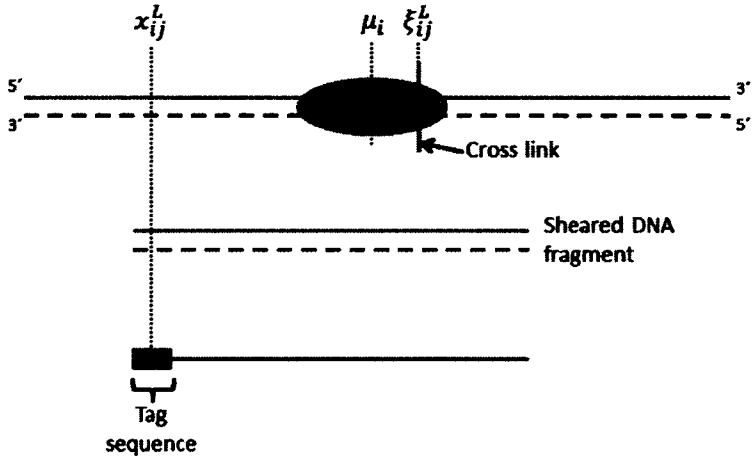
In the first step of a ChIP-seq experiment, the TFs are cross-linked to the DNA. As illustrated in Figure 5, we assume these unobservable cross-link locations of the TF, denoted by ξ_{ij}, to have a random shift from the center of the binding site. For mathematical convenience, we assume $\xi_{ij} \sim N(\mu_i, \sigma^2)$, where μ_i is the binding site location varying for regions, and σ^2 is the variance of the shift remaining the same across the regions.

After the cross-links are established, the DNA is randomly sheared into millions of fragments, most likely, several base pairs long to a couple of thousand base pairs long. Average fragment sizes are 100$\sim$500 bp depending on experiments. We assume that this shearing follows a Poisson process over the whole genome. The mapped tags of the output of the ChIP-seq experiment are the shorter (about 20$\sim$80 bp long) end segments of the fragments. Usually, these end reads are somewhere near the corresponding cross-link location, but the shearing process causes randomness in the exact distance between tag ends and cross-link location, and the short reads on the two DNA strands show different systematic biases in their average position relative to the cross-link location. That is, the short reads mapped to the right strand are expected to demonstrate a shift from the cross-link location to the right and left tags are expected to have a shift to the left of the cross-link location.

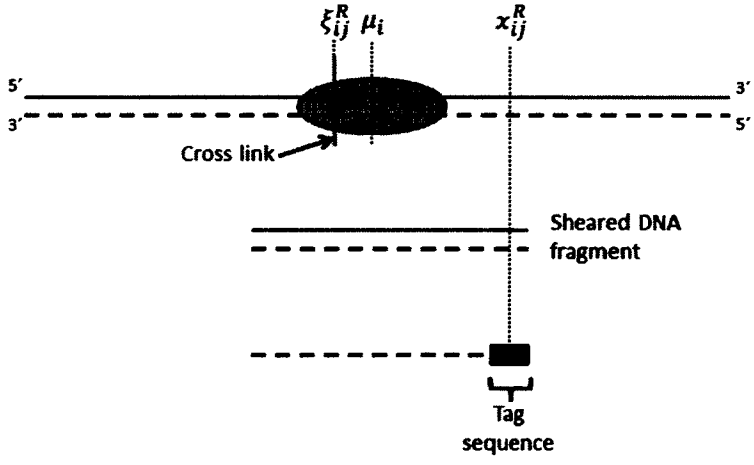
This shifting of the tags can be assumed to be an exponential distribution with mean β, which is assumed to be the same for all regions from the Poisson process assumption. Therefore, for the right tag location x_{ij}^R, with the cross-link location ξ_{ij}^R , the density can be denoted by

$$\pi(x_{ij}^R|\xi_{ij}^R, \beta) = \frac{1}{\beta} \exp\left(-\frac{(x_{ij}^R - \xi_{ij}^R)}{\beta}\right) I\left(x_{ij}^R > \xi_{ij}^R\right),$$

where $I(\cdot)$ is an indicator function.



(a) Tag sequence mapped to the left strand.



(b) Tag sequence mapped to the right strand.

Figure 5. Cross-link locations and mapped tag locations.

The joint density of the (x_{ij}^R, ξ_{ij}^R) is

$$\begin{aligned}\pi(x_{ij}^R, \xi_{ij}^R) &= \pi(x_{ij}^R|\xi_{ij}^R, \beta) \cdot \pi(\xi_{ij}^R|\mu_i, \beta) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\left(\xi_{ij}^R - \mu_i\right)^2}{2\sigma^2}\right) \frac{1}{\beta} \exp\left(-\frac{\left(x_{ij}^R - \xi_{ij}^R\right)}{\beta}\right) I\left(x_{ij}^R > \xi_{ij}^R\right).\end{aligned}$$

By integrating over ξ_{ij}^R,

$$\begin{aligned}\pi\left(x_{ij}^R|\beta, \mu_i, \sigma^2\right) &= \int \pi\left(x_{ij}^R, \xi_{ij}^R|\beta, \mu_i, \sigma^2\right) \mathrm{d}\xi_{ij}^R \\ &= \Phi\left(\frac{x_{ij}^R - (\mu_i + \sigma^2/\beta)}{\sigma}\right) \frac{1}{\beta} \exp\left\{-\frac{1}{\beta}\left(x_{ij}^R - (\mu_i + \sigma^2/2\beta)\right)\right\}, \quad (1)\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. This density function is a normal exponential density (Kim, Jayatillake, and Spouge 2012).

Similarly, the left tag location x_{ij}^L can be denoted with a given cross-link location ξ_{ij}^L as

$$\pi\left(x_{ij}^L|\xi_{ij}^L, \beta\right) = \frac{1}{\beta} \exp\left\{\frac{x_{ij}^L - \xi_{ij}^L}{\beta}\right\} I\left(x_{ij}^L < \xi_{ij}^L\right).$$

Hence, after integrating over ξ_{ij}^L, the density of x_{ij}^L is

$$\begin{aligned}\pi\left(x_{ij}^L|\beta, \mu_i, \sigma^2\right) &= \int \pi\left(x_{ij}^R, \xi_{ij}^L|\beta, \mu_i, \sigma^2\right) \mathrm{d}\xi_{ij}^L \\ &= \left[1 - \Phi\left(\frac{x_{ij}^L - (\mu_i - \sigma^2/\beta)}{\sigma}\right)\right] \frac{1}{\beta} \exp\left\{\frac{1}{\beta}\left(x_{ij}^L - (\mu_i - \sigma^2/2\beta)\right)\right\}. \quad (2)\end{aligned}$$

Note that for the left and the right tags of the i^{th} region, the model parameters μ_i, ν_i, σ, and β are the same. Therefore, the complete density can be given by

$$\begin{aligned}\pi\left(x_{i1}^L, \ldots, x_{in_i^L}^L, x_{ij}^R, \ldots, x_{in_i^R}^R|\beta, \mu_i, \sigma^2\right) &= \prod_{j \in mappable}^{w_i} \pi\left(x_{ij}^L|\beta, \mu_i, \sigma^2\right) \\ &\quad \times \prod_{j \in mappable}^{w_i} \pi\left(x_{ij}^R|\beta, \mu_i, \sigma^2\right). \quad (3)\end{aligned}$$

The overlayed plot of the densities given in (1) and (2) is shown in Figure 6. These density curves clearly reflects the duality of the kernel as well as the mirror image feature. As described in section 1.3, this is one of the main features of the tag

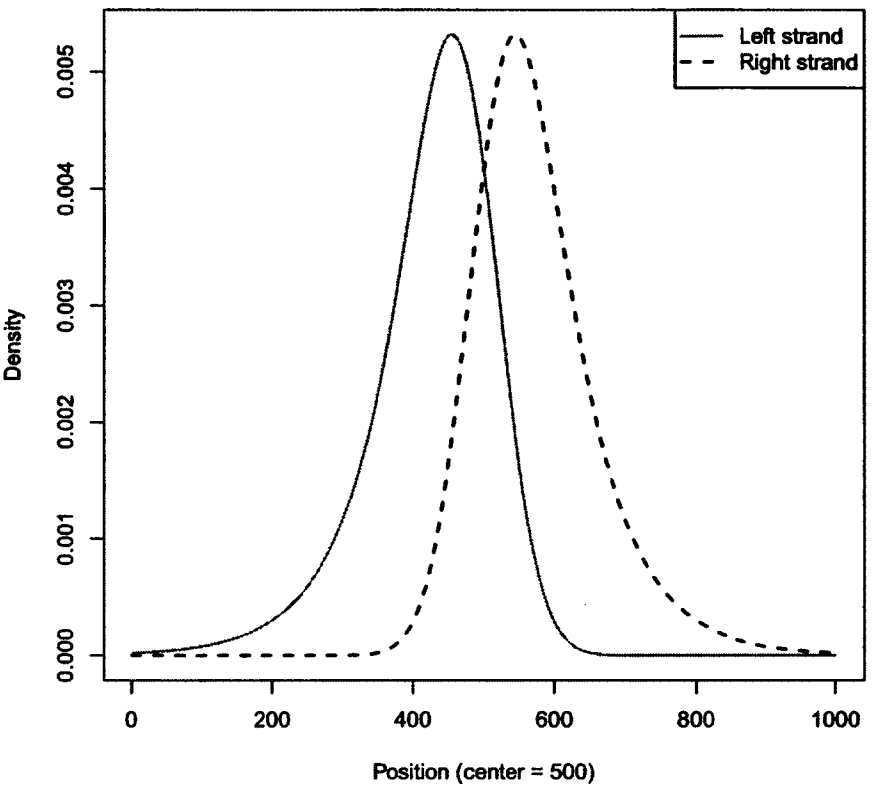


Figure 6. Dual normal-exponential density.

peaks at binding sites in ChIP-seq data. In fact, any peak that does not reflect this mirror image characteristic can be considered as peaks only due to background noise (Kuan et al. 2009).

Furthermore, the validity of the model can be assessed by investigating the distribution of the tags around the high scoring locations of a known motif. A motif is a sequence pattern where its matches called motif sites can be observed throughout the genome. Many transcription factors, including STAT1 and ZNF143 TFs analyzed in this study, exhibit known binding sequence specificities or motif (Kharchenko, Tolstorukov, and Park 2008, Reid et al. 2010, and Izumi et al. 2010). These motif sites can be identified and scored using a position specific scoring matrix (PSSM)

based algorithm (Staden 1989). Among the candidate motif sites for a particular motif, high scoring motif sites are selected using a user-specific cutoff value. The distribution of the tags from ChIP-seq experiment around motif sites can then be investigated by accumulating tag counts from various regions anchored at motif sites. Figures 7 and 8 present frequency plots of the tags mapped to the left and the right strand around the high-scoring motif sites for STAT1 and ZNF143, respectively. The overlayed curve is the proposed dual normal-exponential kernel. For the STAT1 data $\hat{\beta} = 74.1$ and $\hat{\sigma} = 52.5$. For the ZNF143 data $\hat{\beta} = 42.3$ and $\hat{\sigma} = 44.0$. For details, see Kim, Jayatillake, and Spouge (2012).

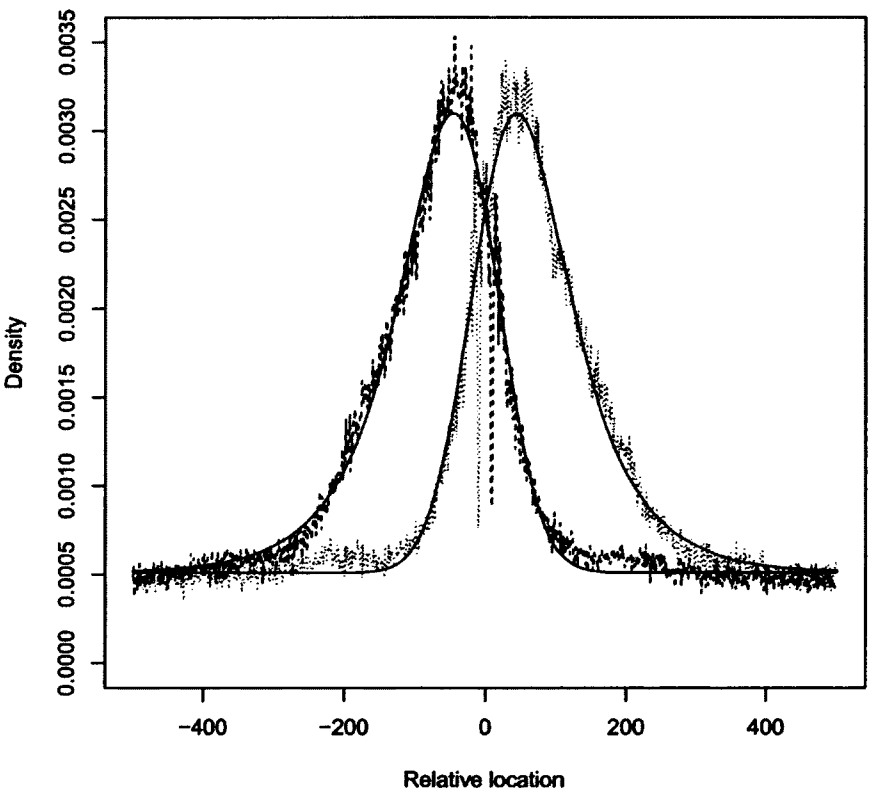


Figure 7. Anchored tag distribution of STAT1.

The fit of the dual normal-exponential kernel is good for both STAT1 and ZNF143

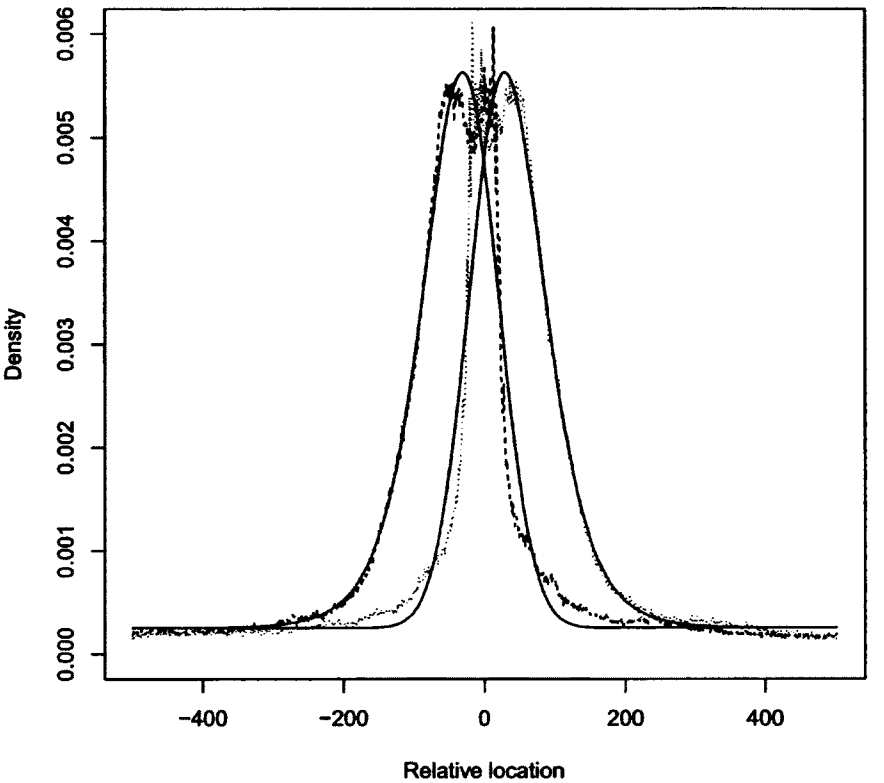


Figure 8. Anchored tag distribution of ZNF143.

as the distribution of the tags closely follows the dual kernel. When observing the two graphs the tag distributions are smoother for ZNF143 than STAT1, which may indicate that the noise level for ZNF143 data is lower than that for the STAT1 data. For both datasets, slight deviations from the dual normal-exponential kernel can be observed. The motif sites for both datasets were identified using the p-value cutoff of 5.0e-6 using the PSSM-based algorithm.

2.2 POISSON MODEL FOR SINGLE BINDING EVENT

An alternative representation of the tags is useful in proposing a Poisson model for the tag distribution. Let y_{ij}^L be the number of left tags observed at location j; y_{ij}^R the number of right tags observed at location j. Tags cannot be observed at *unmappable* locations. Thus for all practical purposes, $(y_{i1}^L, y_{i2}^L, ..., y_{iw_i}^L, y_{i1}^R, y_{i2}^R, ..., y_{iw_i}^R, X^0)$ is an equivalent representation of the data $\mathbf{X}$. Therefore, in this approach we assume the observed tag counts y_{ij}^L and y_{ij}^R to follow a Poisson model.

For the ChIP-seq data, we assume the observed tags to be generated by a Poisson model with the rate parameter as a function of the location μ_i and the intensity parameter ν_i. Consider a tag location j on the left strand and let y_{ij}^L be the number of tags at the location j. Then, the probability mass function is

$$P(Y = y_{ij}^L) = \frac{(\lambda_{ij}^L)^{y_{ij}^L} e^{-\lambda_{ij}^L}}{y_{ij}^L!},$$

where

$$\lambda_{ij}^L = \nu_i \pi(j|\mu_i, \beta, \sigma^2),$$

with $\pi(j|\mu_i, \beta, \sigma^2)$ as given in (2) and $y_{ij}^L \in \{0, 1, \ldots\}$. Similarly for the tag location j in the right strand with a tag count of y_{ij}^R

$$P(Y = y_{ij}^R) = \frac{(\lambda_{ij}^R)^{y_{ij}^R} e^{-\lambda_{ij}^R}}{y_{ij}^R!},$$

where

$$\lambda_{ij}^R = \nu_i \pi(j|\mu_i, \beta, \sigma^2),$$

with $\pi(j|\mu_i, \beta, \sigma^2)$ as given in (1). Here ν_i is the mean number of tags per strand due to the binding in the i^{th} region.

2.3 POISSON MODEL FOR MULTIPLE BINDING EVENTS

The model described in the previous section can only estimate parameters for a single binding event in a given region. However, in some regions there can be multiple binding events separated by about 300 bp or less. In this section, we introduce a model to detect these multiple binding events within short regions.

Let us consider a short region i in the left strand. Let us assume there are k number of binding events within the region. In our model we propose that the number

of tags y_{ij}^L at the *mappable* location j on the left strand is the sum of unobserved tag counts z_{ijh}^L, $h = 0, \ldots, k$, belonging to the k number of binding events. Therefore,

$$\begin{aligned} y_{ij}^L &= z_{ij0}^L + z_{ij1}^L + \ldots + z_{ijk}^L \\ &= \sum_{h=0}^{k} z_{ijh}^L. \end{aligned} \tag{4}$$

Furthermore, we propose $z_{ijh}^L \sim Poisson(\lambda_{ijh}^L)$ and the rate parameter λ_{ijh}^L for each of the binding event or component is a function of the overall distribution of the tags of the binding event in that region and the intensity of the binding event ν_{ih}. That is

$$\lambda_{ijh}^L = \nu_{ih} f_L\left(j|\mu_{ih}, \sigma^2, \beta\right),$$

where

$$f_L\left(j|\mu_{ih}, \sigma^2, \beta\right) = \left[1 - \Phi\left(\frac{j - (\mu_{ih} - \sigma^2/\beta)}{\sigma}\right)\right] \frac{1}{\beta} \exp\left\{\frac{1}{\beta}\left(j - (\mu_{ih} - \sigma^2/2\beta)\right)\right\},$$

for $h = 1, \ldots, k$. In addition, to the components for the binding events, we introduce z_{ij0}^L to denote the tag counts from the background noise. We propose

$$\lambda_{ij0}^L = \nu_0 \rho,$$

where $\rho = \frac{1}{w_i}$. That is, we assume a uniform background noise.

Similarly, for region i of the right strand, the number of tags y_{ij}^R at j^{th} location can be modeled as the sum of unobserved tag counts such that

$$y_{ij}^R = z_{ij0}^R + z_{ij1}^R + \ldots + z_{ijk}^R = \sum_{h=0}^{k} z_{ijh}^R. \tag{5}$$

The rate parameter is modeled as the function of binding event intensity and the distribution of the tags corresponding the right strand,

$$\lambda_{ijh}^R = \nu_{ih} f_R\left(j|\mu_{ih}, \sigma^2, \beta\right),$$

where

$$f_R\left(j|\mu_{ih}, \sigma^2, \beta\right) = \Phi\left(j - (\mu_i + \sigma^2/\beta)\sigma\right) \frac{1}{\beta} \exp\left\{-\frac{1}{\beta}\left(j - (\mu_i + \sigma^2/2\beta)\right)\right\},$$

for $h = 1, \ldots, k$. Similar to the left strand, z_{ij0}^R denotes the tag counts for the background component from background noise in the right strand. By design, the

sum of z_{ijh}^L is always y_{ij}^L. Therefore, the conditional distribution of the observed tags given the corresponding unobserved tags can be expressed as

$$f\left(y_{ij}^L|z_{ij0}^L,\ldots,z_{ijk}^L\right)=\begin{cases}1 & \text{if } \sum_{h=0}^{k} z_{ijh}^L = y_{ij}^L \\ 0 & \text{otherwise}\end{cases}$$

and similarly,

$$f\left(y_{ij}^R|z_{ij0}^R,\ldots,z_{ijk}^R\right)=\begin{cases}1 & \text{if } \sum_{h=0}^{k} z_{ijh}^R = y_{ij}^R \\ 0 & \text{otherwise.}\end{cases}$$

Let $\boldsymbol{\theta}=(\mu_{i1},\ldots,\mu_{ik},\nu_{i0},\ldots,\nu_{ik})$ and $\mathbf{z}_{ij}^L=\left(z_{ij0}^L,\ldots,z_{ijk}^L\right)$. Using the known result that the sum of Poisson distributions follows a Poisson distribution with the rate parameter being the sum of the individual rate parameters, distributions of y_{ij}^L and y_{ij}^R are

$$y_{ij}^L|\boldsymbol{\mu},k,\boldsymbol{\nu}\sim Poisson\left(\sum_{h=0}^{k}\lambda_{ijh}^L\right),$$
$$y_{ij}^R|\boldsymbol{\mu},k,\boldsymbol{\nu}\sim Poisson\left(\sum_{h=0}^{k}\lambda_{ijh}^R\right).$$

The conditional distribution of the unobserved tag counts can be obtained by

$$\begin{aligned}f\left(\mathbf{z}_{ij}^L|y_{ij}^L,\boldsymbol{\theta},k\right)&=\frac{f\left(y_{ij}^L|\mathbf{z}_{ij}^L,\boldsymbol{\theta},k\right)f\left(\mathbf{z}_{ij}^L|\boldsymbol{\theta}\right)}{f\left(y_{ij}^L|\boldsymbol{\theta},k\right)}\\&=\frac{1\times\prod_{h=0}^{k}\frac{e^{-\lambda_{ijh}^L}(\lambda_{ijh}^L)^{z_{ijh}^L}}{z_{ijh}^L!}}{\frac{e^{-\sum_{h=0}^{k}\lambda_{ij}^L}\left(\sum_{h=0}^{k}\lambda_{ijh}^L\right)^{y_{ij}^L}}{y_{ij}^L!}}\\&=\frac{y_{ij}^L!}{z_{ij0}^L!\ldots z_{ijk}^L!}\left(\frac{\lambda_{ij0}^L}{\sum_{h=0}^{k}\lambda_{ijh}^L}\right)^{z_{ij0}^L}\cdots\left(\frac{\lambda_{ijk}^L}{\sum_{h=0}^{k}\lambda_{ijh}^L}\right)^{z_{ijk}^L}.\end{aligned}$$

Therefore,

$$\mathbf{z}_{ij}^L|y_{ij}^L,\boldsymbol{\theta},k\sim Multinomial\left(y_{ij}^L,\frac{\lambda_{ij0}^L}{\sum_{h=0}^{k}\lambda_{ijh}^L}\cdots\frac{\lambda_{ijk}^L}{\sum_{h=0}^{k}\lambda_{ijh}^L}\right). \tag{6}$$

Similarly,

$$\mathbf{z}_{ij}^R|y_{ij}^R,\boldsymbol{\theta},k\sim Multinomial\left(y_{ij}^R,\frac{\lambda_{ij0}^R}{\sum_{h=0}^{k}\lambda_{ijh}^R}\cdots\frac{\lambda_{ijk}^R}{\sum_{h=0}^{k}\lambda_{ijh}^R}\right). \tag{7}$$

The joint distribution of y_{ij}^L and $\mathbf{z}_{ij}^L$ is

$$f\left(y_{ij}^L, \mathbf{z}_{ij}^L | \boldsymbol{\theta}, k\right) = \prod_{h=0}^{k} \frac{e^{-\lambda_{ijh}^L} (\lambda_{ijh}^L)^{z_{ijh}^L}}{z_{ijh}^L!} \cdot I\left(\sum_{h=0}^{k} z_{ijh}^L = y_{ij}^L\right).$$

Similarly, for the right strand,

$$f\left(y_{ij}^R, \mathbf{z}_{ij}^R | \boldsymbol{\theta}, k\right) = \prod_{h=0}^{k} \frac{e^{-\lambda_{ijh}^R} (\lambda_{ijh}^R)^{z_{ijh}^R}}{z_{ijh}^R!} \cdot I\left(\sum_{h=0}^{k} z_{ijh}^R = y_{ij}^R\right).$$

Let $\mathbf{y}_{ij} = \left(y_{ij}^L, y_{ij}^R\right)$ and $\mathbf{z}_{ij} = \left(\mathbf{z}_{ij}^L, \mathbf{z}_{ij}^R\right)$. Assuming the left and the right strand tag counts are independent, the joint distribution of tags at j^{th} position can be given by

$$L\left(\boldsymbol{\theta}, k | \mathbf{y}_{ij}, \mathbf{z}_{ij}\right) = \prod_{h=0}^{k} \frac{e^{-\lambda_{ijh}^L} (\lambda_{ijh}^L)^{z_{ijh}^L}}{z_{ijh}^L!} \frac{e^{-\lambda_{ijh}^R} (\lambda_{ijh}^R)^{z_{ijh}^R}}{z_{ijh}^R!}.$$

Therefore, the likelihood function for the tag counts in region i over the *mappable* locations can be given by

$$L\left(\boldsymbol{\theta}, k | \mathbf{y}_i, \mathbf{z}_i\right) = \prod_{j \in mappable}^{w_i} \prod_{h=0}^{k} \frac{e^{-\lambda_{ijh}^L} (\lambda_{ijh}^L)^{z_{ijh}^L}}{z_{ijh}^L!} \frac{e^{-\lambda_{ijh}^R} (\lambda_{ijh}^R)^{z_{ijh}^R}}{z_{ijh}^R!}, \tag{8}$$

where $\mathbf{y}_i$ is the vector of observed tag counts and $\mathbf{z}_i$ is the vector of unobserved tag counts for the *mappable* locations from both strands.

The Likelihood function in equation (8) of the proposed model for multiple binding sites can be used to estimate the intensities (ν_i's) and locations (μ_i's) of the binding sites. However, since this likelihood also include unobserved tags from multiple binding events, the usual maximization methods cannot be applied. The current estimation problem also differs from the usual set-up of mixture models or missing data models since the number of components, in this case the number of binding events itself is unknown and has to be estimated. In this thesis we investigate two different approaches for the estimation of the parameters as follows:

1. A fully Bayesian approach considering k as a variable.

2. Expectation-maximization (EM) algorithm.

In the Bayesian paradigm the number of components, k, can be considered as a variable with a suitable prior. However, updating the number of components causes the dimension of the variable space to vary. For example, in a proposed move from

k number of components to $k+1$ number of components the parameter space is increased by the parameters of the added components. Methods in the Bayesian paradigm that accommodate change of variable space have been introduced in the *birth-death* method by Stephen (1998) and in the reversible jump Markov chain Monte carlo (RJMCMC) method by Green (1995). In chapter 3 we present in detail a Bayesian model and an application of RJMCMC in estimation of the parameters.

On the other hand, the EM algorithm is applied to likelihoods with missing data, where the number of components is known and is fixed. Therefore, we propose the EM algorithm to be applied for the same region with different number of components, say $k = 1, \ldots, 3$, and then choose the best model via a model selection criterion. Details of the EM algorithm approach and its results are given in Chapter 4.

2.4 SIMULATION DATA

Simulation datasets were generated to investigate the performance of the model, for selecting suitable priors, and to tune in parameters of the priors for the Bayesian model (discussed in chapter 3). In our model, the main parameters of interest are the number of components, k, other than the background component, the locations (μ_{ih})'s, and the intensities (ν_{ih})'s. Therefore, for the simulation of the data, we considered several values for each of these parameters as well as combinations of the values mimicking scenarios that can be encountered in the real ChIP-seq data. In this section we present a brief description of the simulated datasets that will be used throughout the study. The simulation datasets can be categorized into twelve main groups labeled *group 1-group 12* that have varying values for location parameters and number of components. The datasets in *group 1-group 4* have two peaks with equal intensities separated by 200 bp, 150 bp, 100 bp, and 75 bp respectively. Each one of these groups have 6 subgroups of datasets with corresponding intensity values of 150, 125, 100, 75, 50, and 25 for the two peaks (see Table 43 in Appendix). These simulated datasets illustrate the increasing difficulty in detecting the peaks and estimating the parameters as the distances between the peaks decrease and the intensities decrease. Several examples of simulated data under these scenarios are presented in Figure 9.

The datasets in *group 5*, *group 7*, and *group 8* consist of two peaks separated by distances of 200 bp, 150 bp, and 100 bp, respectively. Unlike the groups described

previously, for datasets in these groups the intensities of the two peaks are set to be different from each other. Furthermore, each of these groups has seven subgroups of datasets with different combinations of peak intensities (see Table 44 in Appendix A). These combinations are (200, 50), (150, 50), (150, 75), (125, 50), (100, 25), (75, 25), and (50, 25). The dataset labeled *group 7* consist of two peaks separated by 200 bp and its seven subgroups of datasets have the peak intensities in the reverse order as (50, 200), (50, 150), (50, 125), (25, 100), (25, 75), and (25, 50). Some examples of these simulated data are presented in Figure 10. The final four groups labeled *group 8-group 12* are simulated to have 3 peaks. For the *groups 9-group 11*, the distances between the peaks are set to be equal but vary with values of 200 bp, 150 bp and 100 bp, respectively. Again each group has six subgroups where the three peaks have the intensities 150, 125, 100, 75, 50, and 25 (see Table 45 in Appendix A). Examples of simulated data with three peaks are given in Figure 11.

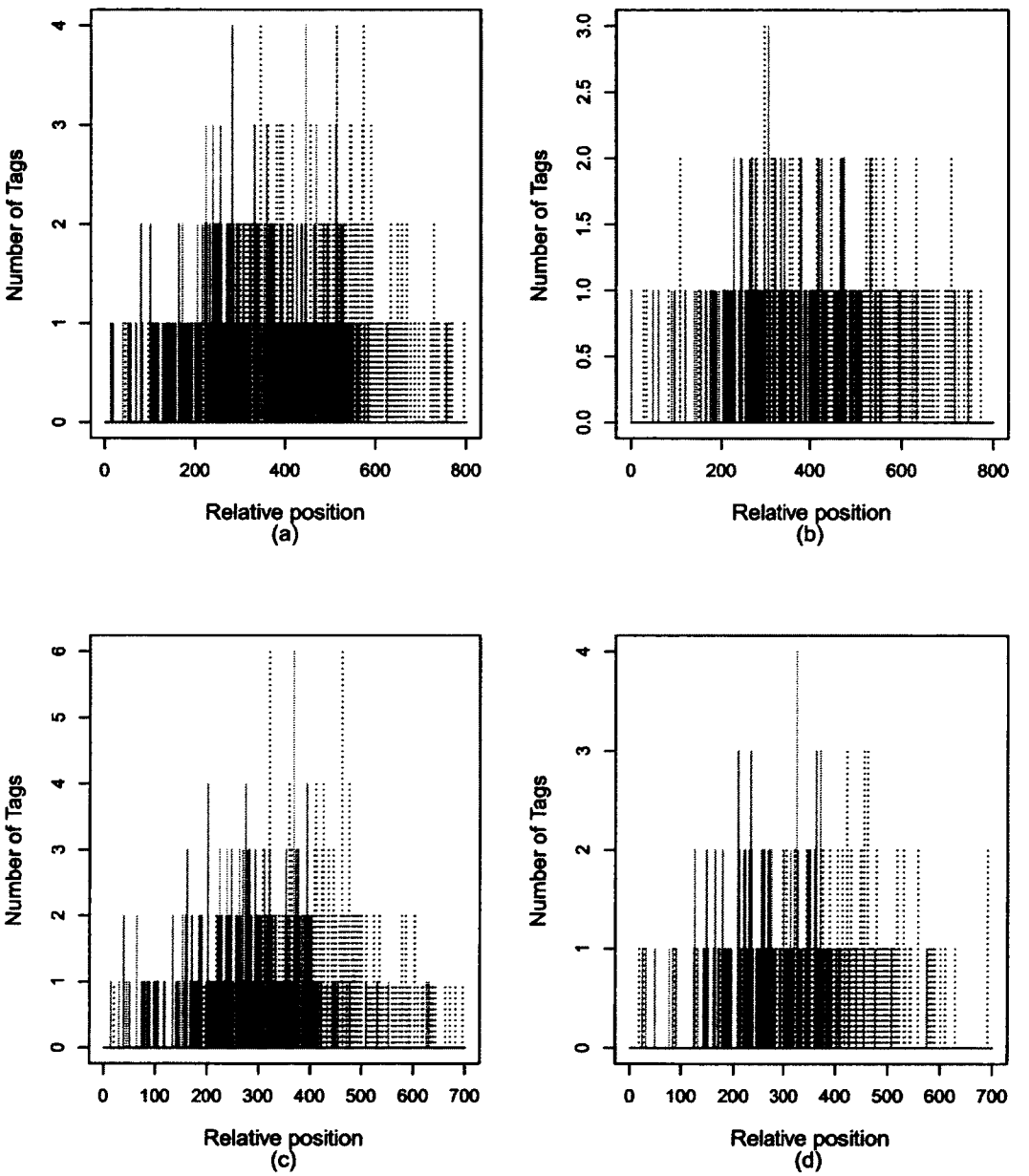


Figure 9. Examples of simulated data for two peaks with equal intensities: (a) $(\mu_1, \mu_2) = (300, 500)$ and $(\nu_0, \nu_1, \nu_2) = (10, 150, 150)$. (b) $(\mu_1, \mu_2) = (300, 500)$ and $(\nu_0, \nu_1, \nu_2) = (10, 75, 75)$. (c) $(\mu_1, \mu_2) = (300, 400)$ and $(\nu_0, \nu_1, \nu_2) = (10, 150, 150)$. (d) $(\mu_1, \mu_2) = (300, 400)$ and $(\nu_0, \nu_1, \nu_2) = (10, 75, 75)$.

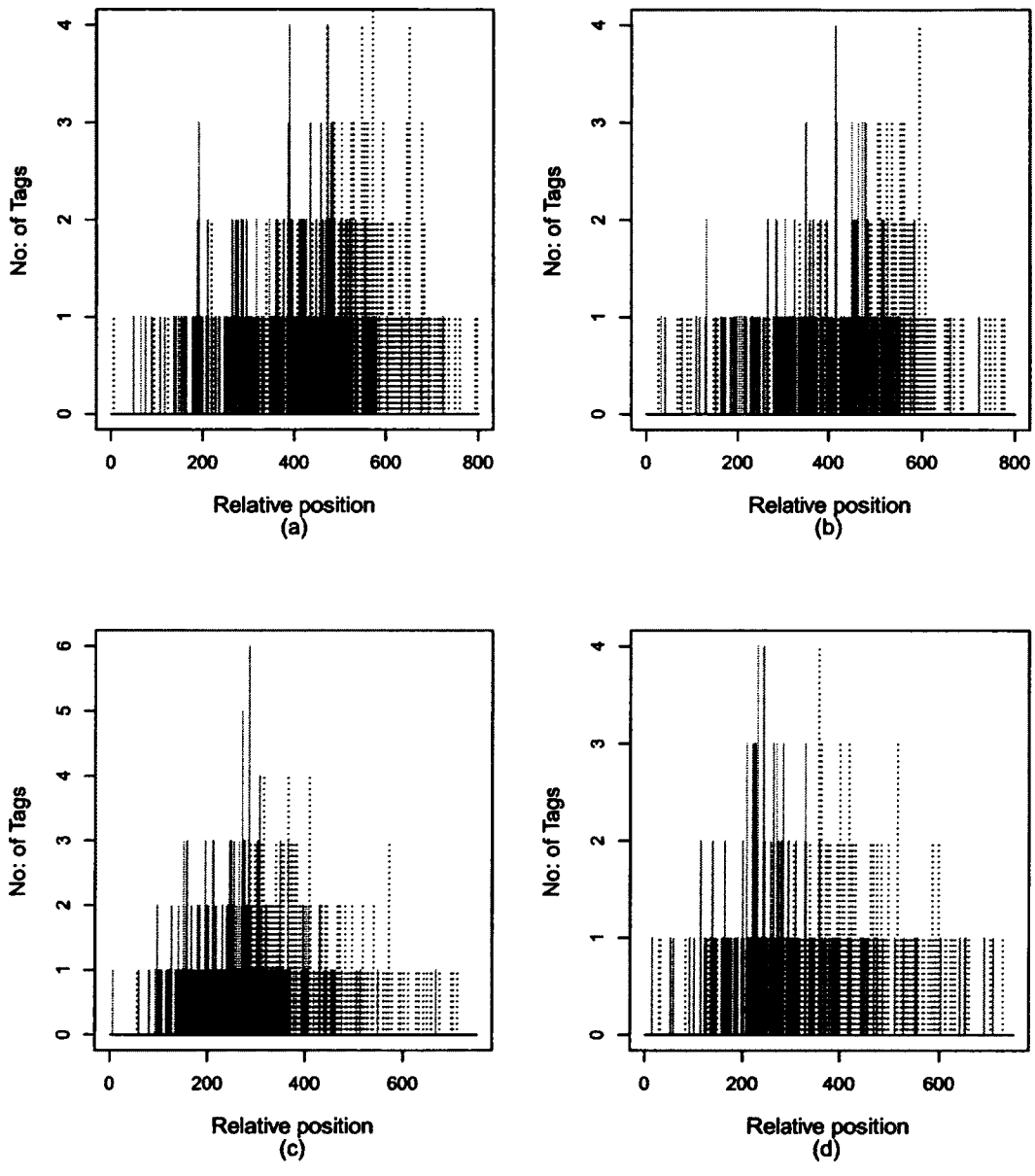
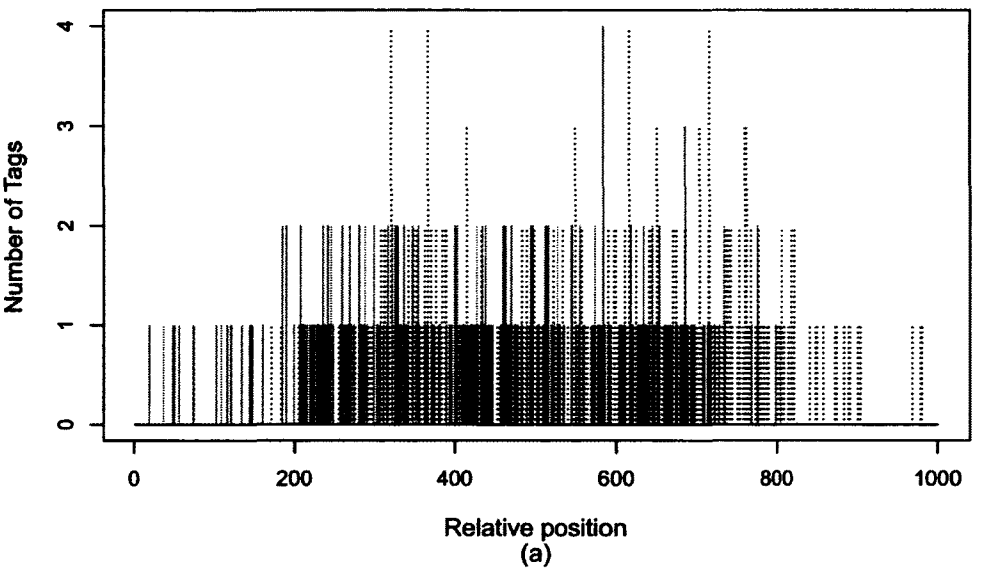


Figure 10. Examples of simulated data for two peaks with unequal intensities: (a) $(\mu_1, \mu_2) = (300, 500)$ and $(\nu_0, \nu_1, \nu_2) = (10, 50, 200)$. (b) $(\mu_1, \mu_2) = (300, 500)$ and $(\nu_0, \nu_1, \nu_2) = (10, 50, 125)$. (c) $(\mu_1, \mu_2) = (300, 400)$ and $(\nu_0, \nu_1, \nu_2) = (10, 200, 50)$. (d) $(\mu_1, \mu_2) = (300, 400)$ and $(\nu_0, \nu_1, \nu_2) = (10, 125, 50)$.



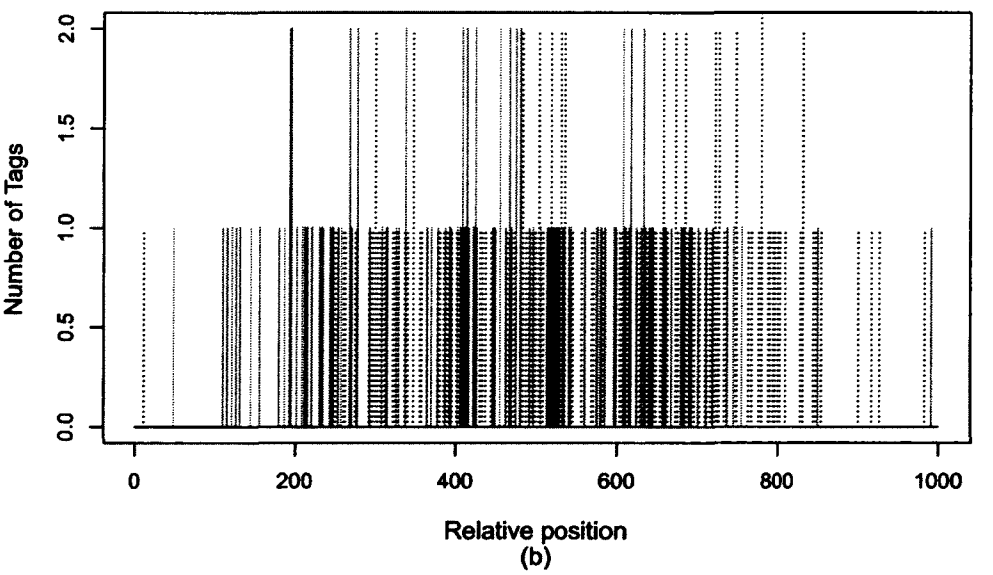


Figure 11. Examples of simulated data for three peaks: (a) $(\mu_1, \mu_2, \mu_3) = (300, 500, 700)$ and $(\nu_0, \nu_1, \nu_2, \nu_3) = (10, 100, 100, 100)$. (b) $(\mu_1, \mu_2, \mu_3) = (300, 500, 700)$ and $(\nu_0, \nu_1, \nu_2, \nu_3) = (10, 50, 50, 50)$.

CHAPTER 3

A BAYESIAN MODEL WITH RJMCMC SCHEME FOR ESTIMATING MULTIPLE BINDING SITES

In chapter 2, we introduced the basic model to determine multiple binding sites within a short region of the genome. The estimates of the parameters of this model provide the binding locations and their intensities as well as the number of binding events within the region. In statistical inference, estimation of these parameters can be viewed as a maximization of the likelihood or can be extended to a Bayesian model where Bayes estimates for the parameters can be obtained. Direct maximization of the likelihood of the observed data is challenging as the number of components itself is unknown. In Bayesian paradigm, the parameters of the model are treated as variables with prior distributions. Therefore, with the Bayesian approach, the number of components can be treated as a variable and can be estimated simultaneously with other parameters. However, variability of the number of components adds another complication by causing the dimension of the variable space to change as the number of components changes. Theory and methodologies have been developed in Bayesian framework to address this issue, especially in the setting of finite mixture models with unknown number of components (Green 1995 and Stephen 2000).

In this chapter we present details of estimating the parameters of the model for the multiple binding sites using a Bayesian method, with the reversible jump Markov chain Monte Carlo (RJMCMC) method for computation. We extend the model given in the previous chapter to a Bayesian model and its details are given in sections 3.1 and 3.2. A brief introduction to the theory of RJMCMC method is provided in section 3.3. Details of the RJMCMC scheme, especially its formulations, selection of RJMCMC proposals and priors are given in section 3.4. Some alternative choices in the implementation of the RJMCMC scheme, as well as detailed results from the simulation study is presented in section 3.5. The final section of the chapter shows the application of RJMCMC scheme to the STAT1 and ZNF143 ChIP-seq datasets.

3.1 BAYESIAN MODEL FOR THE MULTIPLE BINDING SITES

Bayesian inference about a parameter θ or unobserved data z are made in terms of probability statements conditional on the observed data x. A Bayesian statistical model consists of a parametric model, $f(x|\theta)$, and a priori information $\pi(\theta)$. Then the posterior distribution, the distribution of the parameters conditioned on the observed values, can be obtained by

$$f(\theta|x) = \frac{f(x|\theta)\,\pi(\theta)}{\int f(x|\theta)\,\pi(\theta)\,d\theta}.$$

Here the denominator $\int f(x|\theta)\,\pi(\theta)\,d\theta$, which is independent of θ, can be considered as a constant, usually referred to as the normalizing constant. Often in Bayesian modeling, the parametric model is the likelihood function $l(\theta|\mathbf{x})$. By omitting the normalizing constant we can obtain an equivalent form of the posterior distribution as

$$\begin{aligned} f(\theta|x) &\propto f(x|\theta)\,\pi(\theta). \\ &= \text{Likelihood} \times \text{Prior}. \end{aligned} \tag{9}$$

Let us consider estimation of the parameter $h(\theta)$ with $\delta(x)$ under the loss function $L(\delta,\, \theta)$. Then the Bayesian risk can be computed as

$$\begin{aligned} R(\pi, \delta) &= \int\int L(\theta, \delta)\, f(x|\theta(x))\, \pi(\theta)\, d\theta dx \\ &= \int\int L(\theta, \delta(x))\pi(\theta|x)\, d\theta f(x) dx \\ &= \int \left[\int L(\theta, \delta(x))\pi(\theta|x)\, d\theta\right] f(x) dx \\ &= \int E_\theta\left[L(\theta, \delta(x))|x\right] f(x)\, d\theta. \end{aligned}$$

Then the Bayes estimator, $\delta_\pi(x)$, is the value of $\delta(x)$ that minimizes $E_\theta\left[L(\theta, \delta(x))|x\right]$. When $L(\delta,\, \theta) = (h(\theta) - \delta(x))^2$ the Bayes estimator reduces to

$$\delta_\pi(x) = E_\theta\left[L(\theta, \delta(x))|x\right] = \int h(\theta)\, \pi(\theta|x)\, d\theta. \tag{10}$$

In depth theory and proofs in inferences of Bayesian models and applications are given by Robert and Casella (1999), Gelman et al. (2004), and Robert (2007).

The result given in (10) implies two main difficulties associated with Bayesian estimations.

- Often $\pi(\theta|x)$ is not available in closed form.
- Integration, i.e. $\int h(\theta)\,\pi(\theta|x)\,d\theta$, may not be done analytically.

Therefore, instead of seeking analytical solutions, the Bayes estimator is approximated by either numerical or simulation methods such as Markov chain Monte Carlo (MCMC) techniques.

The Bayesian approach to the estimation of multiple binding events requires us to determine the posterior distribution. As described in chapter 2, the model is for estimating the binding events for a short region, say i^{th}, of the genome. This model can be applied to all the partitioned regions of the genome. The likelihood function of the model for detecting multiple binding sites given in (8), referred to as $f(\mathbf{y}_i, \mathbf{z}_i|\boldsymbol{\theta})$, is the joint distribution of the observed data ($\mathbf{y}_i$'s) and the unobserved data ($\mathbf{z}_i$'s) of the i^{th} region. Since the prior distribution of the parameters depends on the number of components for a given region i, let us consider $\pi(\boldsymbol{\nu}_i|k) = \prod_{h=0}^{k} \pi(\nu_{ih})$ and $\pi(k)$ as the priors of $\boldsymbol{\nu}_i$ and k, respectively.

Generally in mixture models the identifiability of the components is important, as discussed by Green (1995), McLachlan and Krishnan (1997), and McLachlan and Peel (2000). The proposed model is also invariant to the permutation of the labels of the components, $h = (1, \ldots, k)$. Therefore, we propose a unique labeling for the components by imposing a natural restriction on the location parameter μ_{ih}'s such that they are of increasing order, i.e. $\mu_{i1} < \mu_{i2} < \ldots < \mu_{ik}$. Thus, the joint prior distribution is $\pi(\boldsymbol{\mu}_i|k) = k! \prod^{k} \pi(\mu_{ih})$. Then the posterior distribution can be computed by

$$\begin{aligned}
\pi(\boldsymbol{\theta}, k|\mathbf{y}_i, \mathbf{z}_i) &\propto f(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta})\,\pi(\boldsymbol{\mu}_i|k)\,\pi(\boldsymbol{\nu}_i|k)\,\pi(k) \\
&= \prod_{j \in mappable}^{w_i} \prod_{h=0}^{k} \frac{e^{-\lambda^L_{ijh}} (\lambda^L_{ijh})^{z^L_{ijh}}}{z^L_{ijh}!} \frac{e^{-\lambda^R_{ijh}} (\lambda^R_{ijh})^{z^R_{ijh}}}{z^R_{ijh}!} \\
&\quad \times \prod_{h=0}^{k} \pi(\nu_{ih}) \times k! \prod_{h=1}^{k} \pi(\mu_{ih}) \times \pi(k). \qquad (11)
\end{aligned}$$

Due to the complexity of the above distribution, it is difficult to obtain analytical estimates for the parameters. In such situations it is common to draw a large sample from the posterior distribution and compute the sample mean as the Bayes estimates. Here, the sample is generated using Markov chain Monte Carlo (MCMC) simulation techniques.

3.2 AN OVERVIEW OF THE RJMCMC SCHEME FOR MULTIPLE BINDING SITES MODEL

To generate a sample from the posterior distribution given in (11) we follow a scheme with four main steps.

- Update the intensity parameters ν_i.
- Update the location parameters μ_i.
- Update the number of tags (z_i^L, z_i^R).
- Update the number of components k.

This scheme will be referred to as the RJMCMC scheme. One sweep of these steps is considered to be one iteration. The first three steps do not change the dimension of the parameter space, therefore, the usual Metropolis-Hastings algorithm (Hastings 1970) and Gibbs sampler can be used. Complete description of these implementations will be provided in section 3.4. Intuitively, the number of components k can be updated by increasing the number of components by splitting an existing component into two or can be decreased by one by combining two adjacent components. These moves cause the dimension of the parameter space to change and requires the use of the generalized Metropolis-Hastings method as described by Green (1995).

3.3 REVERSIBLE JUMP MONTE CARLO MARKOV CHAIN (RJMCMC) METHOD

In general, Markov Chain Monte Carlo (MCMC) techniques, such as the Gibbs sampler and the Metropolis-Hastings algorithm, provide a feasible approach to approximate complex posterior distributions where analytical techniques are too complex or not applicable. The reversible jump MCMC introduced by Green (1995), can be considered as a generalization of the the Metropolis-Hastings algorithm allowing us to generate samples from target distributions with varying dimensions of the parameter space. Since its introduction, RJMCMC has been applied to mixture models (Richardson and Green 1997), change point estimations (Green 1995), clustering (Brooks 2001), and genomic studies (Tadesse, Naijun, and Vanucci 2005). Here we present a brief outline of the theory for RJMCMC. For a detailed description see Green (1995) and Waagepetersen and Sorensen (2001).

A Markov chain $(X_i)_{i\geq 1}$, with a stationary distribution π, is constructed similar to that of Metropolis-Hastings algorithm. Consider each state has two components, i.e. $X_i = (K_i, \boldsymbol{\Theta}_i)$, where K_i is the model indicator and $\boldsymbol{\Theta}_i$ is a stochastic vector in C_k. Let $(k, \boldsymbol{\theta})$ be the values of the current state X_n of the Markov chain. Let $Y_{n+1} = (K^*_{n+1}, \boldsymbol{\Theta}^*_{n+1})$ be a proposal for the next state X_{n+1} with K^*_{n+1} as the proposal of the model indicator K_{n+1} and $\boldsymbol{\Theta}^*_{n+1}$ as the proposal of the vector $\boldsymbol{\Theta}_{n+1}$. The new model indicator is set to the value k' with the probability $p_{kk'}$, where $\sum_{kk'=1}^{k_{max}} p_{kk'} = 1$. Given $K^*_{n+1} = k'$, the $\boldsymbol{\Theta}^*_{n+1}$ is generated in $C_{k'}$. Usually $\boldsymbol{\Theta}^*_{n+1}$ is obtained by applying a deterministic mapping to $\boldsymbol{\theta}$, the value of the current state and to a random component $\mathbf{U}$. This can be obtained by expressing $\boldsymbol{\Theta}^*_{n+1}$, as $\boldsymbol{\Theta}^*_{n+1} = g_{1kk'}(\boldsymbol{\theta}, \mathbf{U})$, where $g_{1kk'} : \mathbb{R}^{n_k + n_{kk'}} \to \mathbb{R}^{n_{k'}}$ is a deterministic mapping, and $\mathbf{U}$ is a random vector on $\mathbb{R}^{n_{kk'}}$ with density $q_{kk'}(\boldsymbol{\theta}, \cdot)$. When moving from state $(k, \boldsymbol{\theta})$ to $(k', \boldsymbol{\theta}')$, and for the reverse from $(k', \boldsymbol{\theta}')$ to $(k, \boldsymbol{\theta})$, the dimension of vectors of Markov chain states and proposal random variables, $(\boldsymbol{\theta}, \mathbf{u})$ and $(\boldsymbol{\theta}', \mathbf{u}')$ respectively, need to be equal. That is, the following dimension matching condition must be satisfied:

$$n_k + n_{kk'} = n_{k'} + n_{k'k}, \tag{12}$$

where $n_{kk'}$ is the dimension change of the parameter when making a move from k to k'. Similarly, $n_{k'k}$ is the dimension change of the parameter when making a move from k' to k. This ensures that $f_k(\boldsymbol{\theta}) q_{kk'}(\boldsymbol{\theta}, \mathbf{u})$ and $f_{k'}(\boldsymbol{\theta}') q_{k'k}(\boldsymbol{\theta}', \mathbf{u}')$ are joint densities on spaces of equal dimensions. Furthermore, assume there exist functions

$$g_{2kk'} : \mathbb{R}^{n_k + n_{kk'}} \to \mathbb{R}^{n_{k'k}} \quad \text{and}$$
$$g_{2k'k} : \mathbb{R}^{n_{k'} + n_{k'k}} \to \mathbb{R}^{n_{kk'}}$$

such that $g_{kk'}$ given by

$$(\boldsymbol{\theta}', \mathbf{u}') = g_{kk'}(\boldsymbol{\theta}, \mathbf{u}) = (g_{1kk'}(\boldsymbol{\theta}, \mathbf{u}), g_{2kk'}(\boldsymbol{\theta}, \mathbf{u})) \tag{13}$$

is one-to-one with

$$(\boldsymbol{\theta}, \mathbf{u}) = g^{-1}_{kk'}(\boldsymbol{\theta}', \mathbf{u}') = g_{k'k}(\boldsymbol{\theta}', \mathbf{u}') = (g_{1k'k}(\boldsymbol{\theta}', \mathbf{u}'), g_{2k'k}(\boldsymbol{\theta}', \mathbf{u}')) \tag{14}$$

and is differentiable.

In addition to the dimension matching, one must also ensure reversibility. Assume $X_n = (K_n, \boldsymbol{\Theta}_n) \sim \pi$, then the condition for reversibility is

$$P(K_n = k, \boldsymbol{\Theta} \in A_k, K_{n+1} = k', \boldsymbol{\Theta}_{n+1} \in B_{k'}) =$$
$$P(K_n = k', \boldsymbol{\Theta}_n \in B_{k'}, K_{n+1} = k, \boldsymbol{\theta}_{n+1} \in A_k), \tag{15}$$

for all k, $k' \in (1, \ldots, k_{max})$ and all subsets A_k and $B_{k'}$ in C_k and $C_{k'}$, respectively. Also, the left hand side of (15) can be written in terms of conditional distribution and $p_k = P(K = k)$ as

$$P(K_n = k, \boldsymbol{\Theta}_n \in A_k, K_{n+1} = k', \boldsymbol{\Theta}_{n+1} \in B_{k'}) = p_k \int_{A_k} f_k(\boldsymbol{\theta}) P(K_{n+1} = k', \boldsymbol{\theta}_{n+1} \in B_{k'} | X_n = (k, \boldsymbol{\theta})). \quad (16)$$

Let $Q^a_{kk'}(\boldsymbol{\theta}, B_{k'})$ be the joint probability of generating a proposal value with $K^*_{n+1} = k'$ and $\boldsymbol{\theta}^*_{n+1}$ in $B_{k'}$ and accepting the the proposal, given that $X_n = (k, \boldsymbol{\theta})$. That is,

$$Q^a_{kk'}(\boldsymbol{\theta}, B_{k'}) = P(K^*_{n+1} = k', \boldsymbol{\Theta}^*_{n+1} \in B_{k'} \text{ and } Y_{n+1} \text{ is accepted } | X_n = (k, \boldsymbol{\theta})).$$

Furthermore, let $s_k(\boldsymbol{\theta})$, the probability of rejecting the proposal, be

$$\begin{aligned} s_k(\boldsymbol{\theta}) &= P(Y_{n+1} \text{ is rejected } | X_n = (k, \boldsymbol{\theta})) \\ &= \sum_{k'=1}^{k_{max}} p_{kk'} \int q_{kk'}(\boldsymbol{\theta}, \mathbf{u})[1 - a_{kk'}(\boldsymbol{\theta}, g_{1kk'}(\boldsymbol{\theta}, \mathbf{u}))] d\mathbf{u}, \end{aligned}$$

where $a_{kk'}$ is the acceptance probability of the proposal. Then

$$P(K_{n+1} = k', \boldsymbol{\Theta}_{n+1} \in B_{k'} | X_n = (k, \boldsymbol{\theta})) = Q^a_{kk'}(\boldsymbol{\theta}, B_{k'}) + s_k(\boldsymbol{\theta}) I(k = k', \boldsymbol{\theta} \in B_{k'}).$$

The left hand side of (15) can be written as

$$\begin{aligned} &p_k \int_{A_k} f_k(\boldsymbol{\theta}) Q^a_{kk'}(\boldsymbol{\theta}, B_{k'}) d\boldsymbol{\theta} + p_k \int_{A_k} f_k(\boldsymbol{\theta}) s_k(\boldsymbol{\theta}) I(k = k', \boldsymbol{\theta} \in B_{k'}) d\boldsymbol{\theta} = \\ &\int_{A_k} p_k f_k(\boldsymbol{\theta}) Q^a_{kk'}(\boldsymbol{\theta}, B_{k'}) d\theta + \int p_k f_k(\boldsymbol{\theta}) s_k(\boldsymbol{\theta}) I(k = k', \boldsymbol{\theta} \in A_k \cap B_{k'}) d\boldsymbol{\theta}. \quad (17) \end{aligned}$$

By symmetry the right hand side of (15) is

$$\int_{B_{k'}} p_{k'} f_{k'}(\boldsymbol{\theta}') Q^a_{k'k}(\boldsymbol{\theta}, A_k) d\boldsymbol{\theta}' + \int p_{k'} f_{k'}(\boldsymbol{\theta}') s_{k'}(\boldsymbol{\theta}') I(k = k', \boldsymbol{\theta}' \in B_{k'} \cap A_k) d\boldsymbol{\theta}'. \quad (18)$$

When considering the equations (17) and (18) it can be observed that the second term is zero when $k \neq k'$ as the indicator function is zero and are equal when $k = k'$. Therefore, a sufficient condition for reversibility given in (15) to hold is

$$\int_{A_k} p_k f_k(\boldsymbol{\theta}) Q^a_{kk'}(\boldsymbol{\theta}, B_{k'}) d\boldsymbol{\theta} = \int_{B_{k'}} p_{k'} f_{k'}(\boldsymbol{\theta}') Q^a_{k'k}(\boldsymbol{\theta}, A_k) d\boldsymbol{\theta}' \quad \forall k, k'. \quad (19)$$

Consider the following assumptions and results obtained in previous steps:

(a) Y_{n+1} is generated in $C_{k'}$ with probability $p_{kk'}$.

(b) $Y_{n+1} \in B_{k'} \iff g_{1kk'}(\boldsymbol{\theta}, U) \in B_{k'}$.

(c) Y_{n+1} is accepted with probability $a_{kk'}(\boldsymbol{\theta}, g_{1kk'}(\boldsymbol{\theta}, \mathbf{U}))$.

(d) $\mathbf{U} \sim q_{kk'}(\boldsymbol{\theta}, \cdot)$.

It follows that

$$Q^a_{kk'}(\boldsymbol{\theta}, B_{k'}) = p_{kk'} \int I(g_{1kk'}(\boldsymbol{\theta}, u) \in B_{k'})\, a_{kk'}(\boldsymbol{\theta}, g_{1kk'}(\boldsymbol{\theta}, \mathbf{u}))\, q_{kk'}(\boldsymbol{\theta}, \mathbf{u})\, d\mathbf{u}. \tag{20}$$

Then the left hand side of (19) can be written as

$$\int_{A_k} p_k f_k(\boldsymbol{\theta})\, Q^a_{kk'}(\boldsymbol{\theta}, B_{k'})\, d\theta = \int\int I(\boldsymbol{\theta} \in A_k,\, g_{1kk'}(\boldsymbol{\theta}, u) \in B_{k'})\, p_k f_k(\boldsymbol{\theta})\, p_{kk'} \\ a_{kk'}(\boldsymbol{\theta}, g_{1kk'}(\boldsymbol{\theta}, \mathbf{u}))\, q_{kk'}(\boldsymbol{\theta}, \mathbf{u}) d\boldsymbol{\theta}\, d\mathbf{u}, \tag{21}$$

and the right hand of (19) can be written as

$$\int_{B_{k'}} p_{k'} f_{k'}(\boldsymbol{\theta}')\, Q^a_{k'k}(\boldsymbol{\theta}', A_k)\, d\boldsymbol{\theta}' = \int\int I(\boldsymbol{\theta}' \in B_{k'},\, g_{1k'k}(\boldsymbol{\theta}', u') \in A_{k'})\, p_k f_{k'}(\boldsymbol{\theta}')\, p_{k'k} \\ a_{k'k}(\boldsymbol{\theta}, g_{1k'k}(\boldsymbol{\theta}', \mathbf{u}))\, q_{k'k}(\boldsymbol{\theta}', \mathbf{u}')\, d\boldsymbol{\theta}\, d\mathbf{u}. \tag{22}$$

As stated in (13) and (14) we can consider $\boldsymbol{\theta} = g_{1k'k}(\boldsymbol{\theta}', \mathbf{u}')$, $\boldsymbol{\theta}' = g_{1kk'}(\boldsymbol{\theta}, \mathbf{u})$ and $\mathbf{u} = g_{2k'k}(\boldsymbol{\theta}', \mathbf{u}')$. Since $g_{kk'}$ is differentiable

$$g'_{kk'}(\boldsymbol{\theta}, \mathbf{u}) = \frac{dg_{kk'}(\boldsymbol{\theta}, \mathbf{u})}{d\boldsymbol{\theta}\, d\mathbf{u}}, \\ d\boldsymbol{\theta}' d\mathbf{u}' = |g'_{kk'}(\boldsymbol{\theta}, \mathbf{u})| d\boldsymbol{\theta} d\mathbf{u},$$

and

$$g'_{kk'}(\boldsymbol{\theta}, \mathbf{u}) = \frac{\partial g_{kk'}((\boldsymbol{\theta}, \mathbf{u}))}{\partial \boldsymbol{\theta} \partial \mathbf{u}} \\ = \begin{bmatrix} \frac{\partial g_{1kk'}(\boldsymbol{\theta}, \mathbf{u})}{\partial \boldsymbol{\theta}} & \frac{\partial g_{2kk'}(\boldsymbol{\theta}, \mathbf{u})}{\partial \boldsymbol{\theta}} \\ \frac{\partial g_{1kk}(\boldsymbol{\theta}, \mathbf{u})}{\partial \mathbf{u}} & \frac{\partial g_{2kk'}(\boldsymbol{\theta}, \mathbf{u})}{\partial \mathbf{u}} \end{bmatrix}.$$

With the above relations (22) can be expressed as

$$\int\int I(g_{1kk'}(\boldsymbol{\theta}, \mathbf{u}) \in B_{k'}, \boldsymbol{\theta} \in A_k)\, p_{k'}\, f_{k'}(g_{1kk'}(\boldsymbol{\theta}, \mathbf{u})) \\ p_{k'k}\, a_{k'k}(g_{1kk'}(\boldsymbol{\theta}, \mathbf{u}), \boldsymbol{\theta})\, q_{kk'}(g_{1kk'}(\boldsymbol{\theta}, \mathbf{u}), g_{2kk'}(\boldsymbol{\theta}, \mathbf{u}))|g'_{kk'}(\boldsymbol{\theta}, \mathbf{u})| d\boldsymbol{\theta} d\mathbf{u}. \tag{23}$$

From (21) and (23) it can be observed that the reversibility condition is satisfied by

$$\begin{aligned} p_k\, f_k(\boldsymbol{\theta})\, p_{kk'}\, q_{kk'}(\boldsymbol{\theta}, \mathbf{u})\, a_{kk'}(\boldsymbol{\theta}, g_{1kk'}(\boldsymbol{\theta}, \mathbf{u})) = \\ p_{k'}\, f_{k'}(g_{1kk'}(\boldsymbol{\theta}, \mathbf{u}))\, p_{k'k}\, q_{k'k}(g_{1kk'}(\boldsymbol{\theta}, \mathbf{u}), g_{2kk'}(\boldsymbol{\theta}, \mathbf{u})) \\ a_{k'k}(g_{1kk'}(\boldsymbol{\theta}, \mathbf{u}), \boldsymbol{\theta}) \left| \frac{\partial g_{kk'}(\boldsymbol{\theta}, \mathbf{u})}{\partial \mathbf{z}\boldsymbol{\theta}\partial \mathbf{u}} \right|. \end{aligned} \tag{24}$$

Choosing the acceptance probability to be as large as possible, while satisfying the reversibility condition, gives the following acceptance probability for the proposal Y_{n+1}:

$$a_{kk'}(\boldsymbol{\theta}, \boldsymbol{\theta}') = min \left\{ 1, \frac{p_{k'}\, f_{k'}(\boldsymbol{\theta}')\, p_{k'k}, q_{k'k}(\boldsymbol{\theta}', \mathbf{u}')}{p_k\, f_k(\boldsymbol{\theta})\, p_{kk'}, q_{kk'}(\boldsymbol{\theta}, \mathbf{u})} \left| \frac{\partial g_{kk'}(\boldsymbol{\theta}, \mathbf{u})}{\partial \boldsymbol{\theta}\partial \mathbf{u}} \right| \right\}. \tag{25}$$

3.4 IMPLEMENTATION OF THE RJMCMC SCHEME

In Bayesian models, the prior distributions on parameters reflect the prior information on the parameters. In the absence of such prior information, the priors are taken to be weakly informative. The priors chosen for the model are as follows:

$$\begin{aligned} \mu_{ih} &\sim U(1, w_i), \\ \nu_{ih} &\sim Gamma(a, b), \\ k &\sim Poisson(\lambda_c). \end{aligned}$$

These priors and values for hyper-parameters were chosen to be weakly informative. The prior for the μ is chosen to be a uniform distribution over the region length. The ν parameter can also be considered as the number of tags belonging to each component. In most regions, the number of tags are around 0-30, and in the presence of a binding event, it can increase to larger values such as 200-500. Therefore, we considered an exponential distribution with a mean of 25, which is skewed towards zero with smaller probability on larger values. The rate parameter λ_c for the number of additional components is chosen to be 0.5. Here we assume that the probability of observing large number of components is small. With these priors the posterior

distribution can be explicitly written as

$$f(\boldsymbol{\theta}, k|\mathbf{y}_i, \mathbf{z}_i) \propto \prod_{j \in mappable}^{w_i} \prod_{h=0}^{k} \frac{e^{-\lambda_{ijh}^L}(\lambda_{ijh}^L)^{z_{ijh}^L}}{z_{ijh}^L!} \frac{e^{-\lambda_{ijh}^R}(\lambda_{ijh}^R)^{z_{ijh}^R}}{z_{ijh}^R!}$$
$$\times k! \left(\frac{1}{w_i}\right) \left(\frac{1}{b^a \Gamma(a)}\right)^{k+1} (\nu_0 \ldots \nu_k)^{a-1} exp\left(\frac{1}{b}\sum_{h=0}^{k} \nu_{ih}\right)$$
$$\times \frac{e^{-\lambda_c}\lambda_c^k}{k_c!}. \tag{26}$$

In the following subsections we present the details of the four update steps introduced in section 3.2.

3.4.1 PROCEDURE FOR UPDATING LOCATION PARAMETERS $(\mu_I' S)$

New values of location parameters are generated by $\mu_{ih}^* \sim N(\mu_{ih}, \sigma_\mu^2)$, where μ_{ih} is the current value and $h = (1, \ldots, k)$. The new values are accepted with probability $\min\{1, \alpha_\mu\}$, where $\boldsymbol{\theta}^*$ is the vector of parameters with proposed values and

$$\alpha_\mu = \frac{f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}^*, k)}{f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, k)}.$$

3.4.2 PROCEDURE FOR UPDATING INTENSITY PARAMETERS (ν_{IH})'S

New values of intensity parameters, ν_{ih}^*, are generated by $\nu_{ih}^* \sim N(\nu_{ih}, \sigma_\nu^2)$, where $h = (0, \ldots, k)$ and ν_{ih}. The new values are accepted with probability $\min\{1, \alpha_\nu\}$:

$$\alpha_\nu = \frac{f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}^*, k)}{f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, k)} \frac{(\nu_0^* \ldots \nu_k^*)^{a-1}}{(\nu_0 \ldots \nu_k)^{a-1}} exp\left\{-\frac{1}{b}\left(\sum_{h=0}^{k} \nu_h^* - \sum_{h=0}^{k} \nu_h\right)\right\},$$

where $\boldsymbol{\theta}^*$ is the vector of parameters with proposed values.

3.4.3 PROCEDURE FOR UPDATING TAG COUNTS (Z_{IJ})

This update step can be performed using the Gibbs sampler. For the j^{th} *mappable* position in the region, new values of the unobserved tags $z^L_{ij1}, \ldots, z^L_{ijk}$ and $z^R_{ij1}, \ldots, z^R_{ijk}$ are generated from the the corresponding full conditional distributions given in (6) and (7).

3.4.4 PROCEDURE FOR UPDATING THE NUMBER OF COMPONENTS (K)

The number of components are updated by either decreasing the number of components by merging two adjacent components or by splitting an existing component into two. In the previous section we defined $p_{kk'}$ as the probability the new number of components is set to k' when the current number of component is k. Before invoking the step to update k, we set the values for these probabilities. In the implementation we only consider an increase or decrease by one component. Let us call a move type of k to $k+1$ as *split* and a move type of k to $k-1$ as *merge*. We set $p_{k(k+1)} = p_{k(k-1)} = 0.5$ when $k \in \{2, \ldots, k_{max} - 1\}$. Here k_{max} is the maximum number of components and we set $k_{max} = 4$, since we assume that the probability of observing more than four binding sites for any given region is very small in real data. When $k = k_{max}$, $p_{k_{max}(k_{max}+1)} = 0$ and $p_{k_{max}(k_{max}-1)} = 1$, as more components cannot be created, the only move allowed is a *merge*. Similarly, when $k = 1$, $p_{10} = 0$ and $p_{12} = 1$, as there are no other components to merge, the only move type allowed is *splitting* the current component. For all other cases $p_{kk'} = 0$. Next, we present the steps for updating the number of components.

First, a decision is made to split a component or merge a pair of adjacent components based on the probabilities $p_{kk'}$. Steps (a) and (b) give detailed description of the procedure for splitting or merging components.

(a) Merging two components

We choose two adjacent components randomly, say h_{*1} with parameters $(\mu_{h_{*1}}, \nu_{h_{*1}})$ and h_{*2} with parameters $(\mu_{h_{*2}}, \nu_{h_{*2}})$, to be merged into a single

component h_* with new parameters (μ_{h_*}, ν_{h_*}). We propose the following functions to set the values for the new component

$$\left.\begin{aligned} \mu_{h_*} &= \frac{\nu_{h_{*1}}\mu_{h_{*1}} + \nu_{h_{*2}}\mu_{h_2}}{\nu_{h*}} \\ \nu_{h_*} &= \nu_{h_{*1}} + \nu_{h_{*2}}. \end{aligned}\right\} \tag{27}$$

In addition the tags allocated to the two components is combined and re-allocated to the new component. The proposal will be accepted with probability $min\{1, \alpha_{split}^{-1}\}$, which will be described later in this subsection.

(b) Splitting a component

Consider choosing component h_* to split into two components h_{*1} and h_{*2} with parameters $(\mu_{h_{*1}}, \nu_{h_{*1}})$ and $(\mu_{h_{*2}}, \nu_{h_{*2}})$, respectively. This move increases the dimension of the parameter space by two. To match the dimension change, generate a random vector $\mathbf{U} = (u_1, u_2)$ where $u_1 \sim Exp(a)$ and $u_2 \sim Beta(1,1)$. The values for the new parameters are determined using the parameter values of h_* and $\mathbf{U}$ as follows,

$$\left.\begin{aligned} \mu_{h_{*1}} &= \mu_{h_*} - \frac{u_1}{u_2} \\ \mu_{h_{*2}} &= \mu_{h_*} + \frac{u_1}{(1-u_2)} \\ \nu_{h_{*1}} &= u_2\nu_{h_*} \\ \nu_{h_{*2}} &= (1-u_2)\nu_{h_*}. \end{aligned}\right\} \tag{28}$$

The tags from the left and right strands are allocated to h_* can be re-allocated to the new components with probabilities $(p_c^L, 1-p_c^L)$ and $(p_c^R, 1-p_c^R)$, respectively, where

$$p_c^L = \frac{\lambda_{ijh*1}^L}{(\lambda_{ijh_{*1}}^L + \lambda_{ijh_{*2}}^L)} \text{ and } p_c^R = \frac{\lambda_{ijh*1}^R}{(\lambda_{ijh_{*1}}^R + \lambda_{ijh_{*2}}^R)}.$$

The move will be accepted with probability $min\{1, \alpha_{split}\}$.

The sets of equations in (27) and (28) are the one-to-one mapping functions $g_{kk'}$ and $g_{k'k}$ described in the theory of RJMCMC in section 3.3. In many applications of RJMCMC, the mapping functions for combining components are obtained by matching the moments of the parameters that change in the two states. However,

it is not compulsory to follow this approach, especially when the moments of the parameters are complex, and any set of functions that is one-to-one, deterministic and differentiable will be sufficient. It can be observed that the set of equations in (27) and (28) satisfy these conditions. In addition, these functions are also intuitive. Specifically, when combining two components, we propose a weighted average based on the intensities of the components so the combined component will be located closer to the component with higher intensity. Simply adding the intensities to form the intensity of the new component ensures that the total intensity of the region will be preserved which also indirectly translates to the number of tags in the region. When splitting a component, locations of the new components are set by subtracting and adding a short length from the location of the original component. The advantage of this proposal is that the length that is subtracted or added can vary while preserving the order of the locations of the new components to satisfy the identifiability condition stated in section 2.3.

The acceptance probability, α_{split}, can be obtained from (25) given in section 3.3. Since the number of components are increased or decreased by one, we can set $k' = k + 1$. Also, we can consider $p_k f_k(\boldsymbol{\theta}) = c^{-1} \times$ *Posterior distribution*, where c^{-1} is the unknown normalizing constant. The ratio of $p_{kk'}/p_{k'k}$ is replaced by $p_{(k+1)k}/(p_{k(k+1)} \times P_{alloc})$. Here, P_{alloc} is the probability of the particular re-allocation of the tags in the original component to the new components created in the split move. For our scheme this can be explicitly expressed as

$$P_{alloc} = \frac{z^L_{ijh_*}!}{z^L_{ijh_{*1}}!\, z^L_{ijh_{*2}}!} \frac{z^R_{ijh_*}!}{z^R_{ijh_{*1}}!\, z^R_{ijh_{*2}}!} \frac{\left(\lambda^L_{ijh_{*1}}\right)^{z^L_{ijh_{*1}}} \left(\lambda^L_{ijh_{*2}}\right)^{z^L_{ijh_{*2}}} \left(\lambda^R_{ijh_{*1}}\right)^{z^R_{ijh_{*1}}} \left(\lambda^R_{ijh_{*2}}\right)^{z^R_{ijh_{*2}}}}{\left(\lambda^L_{ijh_{*1}} + \lambda^L_{ijh_{*2}}\right)^{z^L_{ijh_*}} \left(\lambda^R_{ijh_{*1}} + \lambda^R_{ijh_{*2}}\right)^{z^R_{ijh_*}}}.$$

Also, note that in the reverse move $P_{alloc} = 1$. Furthermore, since the random vector $\mathbf{U}$ is generated independent of the current state, the term $q_{k'k}(\boldsymbol{\theta}', \mathbf{u}')/q_{k'k}(\boldsymbol{\theta}', \mathbf{u}')$ reduces to $1/(\pi(u_1)\pi(u_2))$, where $\pi(u_1)$ and $\pi(u_2)$ are the densities of u_1 and u_2, respectively.

The Jacobian for the transformation of variables from $(\mu_{h_*}, \mu_{h_*}, u_1, u_2)$ to $(\mu_{h_{*1}}, \mu_{h_{*2}}, \nu_{h_{*1}}, \nu_{h_{*2}})$ can be computed as follows.

$$||J|| = \begin{Vmatrix} \frac{\partial \mu_{h_{*1}}}{\partial \mu_{h_*}} & \frac{\partial \mu_{h_{*2}}}{\partial \mu_{h_*}} & \frac{\partial \nu_{h_{*1}}}{\partial \mu_{h_*}} & \frac{\partial \nu_{h_{*2}}}{\partial \mu_{h_*}} \\ \frac{\partial \mu_{h_{*1}}}{\partial \nu_{h_*}} & \frac{\partial \mu_{h_{*2}}}{\partial \nu_{h_*}} & \frac{\partial \nu_{h_{*1}}}{\partial \nu_{h_*}} & \frac{\partial \nu_{h_{*2}}}{\partial \nu_{h_*}} \\ \frac{\partial \mu_{h_{*1}}}{\partial u_1} & \frac{\partial \mu_{h_{*2}}}{\partial u_1} & \frac{\partial \nu_{h_{*1}}}{\partial u_1} & \frac{\partial \nu_{h_{*2}}}{\partial u_1} \\ \frac{\partial \mu_{h_{*1}}}{\partial u_2} & \frac{\partial \mu_{h_{*2}}}{\partial u_2} & \frac{\partial \nu_{h_{*1}}}{\partial u_2} & \frac{\partial \nu_{h_{*2}}}{\partial u_2} \end{Vmatrix}$$

$$= \begin{Vmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & u_2 & 1-u_2 \\ -\frac{1}{u_2} & \frac{1}{1-u_2} & 0 & 0 \\ \frac{1}{u_2^2} & \frac{1}{(1-u_2)^2} & \nu_{h_*} & -\nu_{h_*} \end{Vmatrix} = \frac{\nu_{h_*}}{(1-u_2)u_2}.$$

Therefore, the acceptance probability can be expressed as,

$$\alpha_{split} = \text{likelihood ratio} \times (k+1)\left(\frac{1}{w_i}\right)\left(\frac{1}{b^a\Gamma(a)}\right)\left(\frac{\nu_{h_{*1}}\nu_{h_{*2}}}{\nu_{h_*}}\right)^{a-1}$$
$$\times \frac{\lambda_c}{m-1} \times \frac{d_{(k+1)k}}{(p_{k(k+1)k}P_{alloc})} \times \frac{1}{\pi(u_1)\,\pi(u_2)}||J||. \tag{29}$$

The acceptance ratio derived here is somewhat reminiscent to the the acceptance ratio used by Richardson and Green (1997) in their implementation of RJMCMC method for estimating parameters of a Gaussian mixture model.

The objective of the RJMCMC scheme is to generate a sample from the posterior distribution given in (26). Recall that one sweep of the four steps of the scheme is considered as an iteration. The parameters are updated in batches. To obtain the posterior sample, we ran 15000 iterations allowing a burn-in period of 5000 iterations to discard, where the chain may not converge to the true posterior distribution. The estimates for the number of components is calculated by taking the number of components with the highest frequency. Since the location and intensities are conditional on the number of components, their estimates were calculated with conditional sample averages.

Usually the simulated data from MCMC techniques tends to be correlated. Therefore, it is not accurate to consider the data to be independent and calculate the sample standard deviation as the the standard deviation of the estimates. Thus, to minimize this correlation among the sampled data, we sub-grouped the simulated draws into groups of 20 and calculated the sample mean (batch means) for the parameters $(\theta_m^1, \ldots, \theta_m^b$, b=number of subgroups). Then we approximate the standard deviation of the estimates by the standard deviation of the batch means (Albert 2009).

In addition to the priors and the one-to-one functions described in subsection 3.4.4, it is possible to propose alternative functions. We investigated a few alternatives, especially for priors for the ν parameter, and one-to-one functions for updating the k parameter. To compare the performance of the RJMCMC scheme with the alternatives, they were applied to simulated data, where the values of the parameters are known (see section 3.5).

3.5 COMPARING ALTERNATIVE CHOICES IN IMPLEMENTATION AND SIMULATION RESULTS

3.5.1 ALTERNATIVE ONE-TO-ONE FUNCTIONS FOR UPDATING NUMBER OF COMPONENTS

Among the parameters to be estimated in the model, the number of peaks, k, can be considered as the most crucial, as it directly affects the estimation of the intensity and the location parameters. The success of the estimation of k depends on the one-to-one functions described in 3.4.4. Here we compare the performance of the one-to-one functions given in section 3.4, referred to as *proposal 1*, with two alternative proposals, *proposal 2* and *proposal 3*.

- *Proposal 2*

  (a) For merging two components

$$\left.\begin{aligned} \nu_{h_*} &= \nu_{h_1} + \nu_{h_2} \\ \mu_{h_*} &= \frac{\mu_{h_1} + \mu_{h_2}}{2}. \end{aligned}\right\} \tag{30}$$

  (b) For splitting a component

$$\left.\begin{aligned} \mu_{h_{*1}} &= u_1 \mu_{h_*} \\ \mu_{h_{*2}} &= (1-u_1)\mu_{h_*} \\ \nu_{h_{*1}} &= u_2 \nu_{h_*} \\ \nu_{h_{*2}} &= (1-u_2)\nu_{h_*}. \end{aligned}\right\} \tag{31}$$

  Here $u_1 \sim Beta(2,2)$ and $u_2 \sim Beta(1,1)$.

- *Proposal 3*

  (a) For merging two components

$$\left.\begin{aligned} \mu_{h_*} &= \frac{\nu_{h_1}\mu_{h_1} + \nu_{h_2}\mu_{h_2}}{\nu_{h*}} \\ \nu_{h_*} &= \nu_{h_1} + \nu_{h_2}. \end{aligned}\right\} \tag{32}$$

  (b) For splitting a component

$$\left.\begin{aligned} \mu_{h_{*1}} &= u_1\mu_{h_*} \\ \mu_{h_{*2}} &= \frac{1-u_1u_2}{1-u_1}\mu_{h_*} \\ \nu_{h_{*1}} &= u_2\nu_{h_*} \\ \nu_{h_{*2}} &= (1-u_2)\nu_{h_*}. \end{aligned}\right\} \tag{33}$$

  Here $u_1 \sim Beta(2,2)$ and $u_2 \sim Beta(1,1)$.

The equations (30) and (31) in *proposal 2* and the equations (32) and (33) in *proposal 3* satisfy the conditions of being one-to-one, deterministic and differentiable. The *proposal 2* can be viewed as the simplest or naive proposal, where when combining two components, the location of the new peak is considered to be half way between the two components.

However, in this proposal when splitting a random component, it is possible to observe that the new location parameters are not in ascending order. In such cases the constraint on identifiability of the components is violated. Therefore, in the implementation, new proposal values that violate this condition are rejected.

In *proposal 1* and *proposal 3*, the functions by default preserves the order of values of the new locations. The combine steps in *proposal 1* and *proposal 3* are the same but the split steps are different. In *proposal 1*, a random quantity is subtracted and added to obtain the location parameters of the first and the second split components. In *proposal 3*, locations of the new components are positioned by a ratio of the distance between the start of the region and the location of the current component that is being split. In each case, if existing components falls between the new components, then we reject the proposed move.

The RJMCMC schemes with the separate proposals were applied to the simulated data described in section 2.4. The priors for the ν is considered as $Exponential(10)$. We consider the RJMCMC scheme to be a success if the estimate of the parameter

k is the number of peaks considered in simulating the data and the estimates for the location parameters μ_i is within 50 bp of the simulated values. Here we look at the number of successes the three RJMCMC schemes reports on the simulated data. Since each subgroup of data has 20 sample datasets, the maximum number of successes is twenty. In general, a larger number of sample datasets would improve the accuracy of the observation and conclusions derived. At the same time, 20 samples from each scenario provide a sufficiently large number of observations to make an informed decision on the performance of the computation schemes, especially, when considering the overall number of scenarios (96) and the time taken to run the RJMCMC scheme for each simulated dataset.

Table 1 summarizes the number of successes observed for the datasets simulated with two peaks with equal intensities (*group 1* to *group 4*), and Table 2 summarizes the number of successes observed for the datasets simulated with two peaks with unequal intensities (*group 5* to *group 8*). Finally, Table 3 gives the summary of the number of successes observed for the datasets simulated with three peaks, *group 9* to *group 12*.

When considering the results in Tables 1 and 2 for two peaks, it can be observed that the *proposal 1* and *proposal 3* performed comparably, but *proposal 2* performs poorly, especially when the intensities decreases. Therefore, *proposal 2* is disregarded in the analysis of datasets simulated with three peaks. The results for the three peaks (Table 3) also indicate that the performance of the *proposal 1* and *proposal 3* are comparable. However, when considering the results presented in Table 2, *proposal 1* performs slightly better than *proposal 3* when the distance between the peaks with unequal intensities decreases. The objective of the study is to detect binding events within short regions, where the sites are closer to each other. Therefore, *proposal 1* can be considered as the best choice among the three proposals for the RJMCMC scheme.

Table 1. Results from simulated data with two peaks and equal intensities using the RJMCMC schemes with the three proposals

Distance	Intensities		No. of successes (/20)		
between peaks	Peak1	Peak2	Proposal 1	Proposal 2	Proposal 3
200	150	150	5	14	8
	125	125	11	16	9
	100	100	14	15	12
	75	75	19	16	17
	50	50	19	19	18
	25	25	16	11	13
150	150	150	9	10	6
	125	125	14	16	9
	100	100	17	17	17
	75	75	19	17	19
	50	50	20	18	19
	25	25	17	4	17
100	150	150	13	14	15
	125	125	17	15	16
	100	100	18	11	16
	75	75	17	10	19
	50	50	13	2	11
	25	25	11	0	13
75	150	150	12	6	15
	125	125	17	4	17
	100	100	11	2	15
	75	75	11	0	14
	50	50	9	0	13
	25	25	0	0	8

Table 2. Results from simulated data with two peaks and unequal intensities using the RJMCMC schemes with the three proposals

Distance	Intensities		No. of successes (/20)		
between peaks	Peak1	Peak2	Proposal 1	Proposal 2	Proposal 3
200	200	50	9	17	10
	150	50	11	15	16
	150	75	12	11	18
	125	50	15	14	15
	100	25	10	10	7
	75	25	16	13	12
	50	25	18	17	15
150	50	200	19	19	19
	50	150	18	16	17
	75	150	16	17	18
	50	125	20	19	19
	25	100	16	14	13
	25	75	19	17	16
	25	50	20	15	17
100	200	50	7	14	10
	150	50	13	11	13
	150	75	13	13	16
	125	50	18	12	14
	100	25	7	4	6
	75	25	8	8	8
	50	25	9	8	7
75	200	50	10	6	12
	150	50	12	7	10
	150	75	19	7	10
	125	50	17	9	11
	100	25	6	0	7
	75	25	6	0	10
	50	25	12	2	9

Table 3. Results from simulated data with three peaks using the RJMCMC schemes with the three proposals

Distances between		Intensities	No. of successes (/20)	
First two peaks	Second two peaks	of the peaks	Proposal 1	Proposal 3
200	200	150	0	0
		125	1	2
		100	1	4
		75	8	11
		50	10	16
		25	9	10
150	150	150	0	1
		125	0	0
		100	4	2
		75	7	6
		50	16	18
		25	6	9
100	100	150	5	4
		125	9	2
		100	10	7
		75	4	6
		50	0	1
150	200	150	1	1
		125	1	0
		100	2	5
		75	11	10
		50	11	11
		25	0	2

3.5.2 UPDATING NUMBER OF COMPONENTS

In this section we present a comparison of the performance of the RJMCMC schemes with three different priors for the intensity parameter. The three priors we considered are as follows:

- Exponential prior
$$\pi(\nu_i) = \frac{1}{\lambda} e^{-\frac{\nu_i}{\lambda}} \qquad \text{with } \lambda = 25.$$
- Truncated Cauchy prior
$$\pi(\nu_i) = \frac{c}{(\lambda^2 + \nu_i^2)} \cdot I(\nu_i > 0) \qquad \text{with } \lambda = 25 \text{ and } c = \frac{2\pi}{\lambda}.$$
- Uniform(0, 2000)

Similar to the analysis in subsection 3.5.1, three RJMCMC schemes with these priors were implemented and applied to the simulated datasets. Their results are summarized in Tables 4-6. All three RJMCMC schemes employed *proposal 1* in the step for updating the number of components.

When compared to the results in subsection 3.5.1, the number of successes have increased significantly for all the three priors than with the prior of $Exp(10)$ on ν_i. Overall, when compared to the Cauchy as well as the uniform prior, the exponential prior with mean 25 performs better with more number of success, especially for shorter distances between the the peaks. Therefore, we will consider the exponential prior hereafter for the implementation of the RJMCMC scheme.

Table 4. Results from simulated data with two peaks and equal intensities for the RJMCMC schemes with the three priors

Distance	Intensities		No. of successes (/20)		
between peaks	peak1	peak2	Exponential	Cauchy	Uniform
200	150	150	20	20	20
	125	125	20	20	20
	100	100	19	20	20
	75	75	20	20	20
	50	50	20	20	20
	25	25	16	16	13
150	150	150	20	20	20
	125	125	20	20	20
	100	100	20	20	20
	75	75	19	20	20
	50	50	20	20	20
	25	25	16	8	6
100	150	150	20	20	20
	125	125	19	16	18
	100	100	18	18	17
	75	75	19	12	12
	50	50	14	7	5
	25	25	5	1	0
75	150	150	19	9	10
	125	125	16	6	7
	100	100	15	2	3
	75	75	10	3	3
	50	50	8	1	0
	25	25	0	0	0

Table 5. Results from simulated data with two peaks and unequal intensities for the RJMCMC schemes with the three priors

Distance	Intensities		No. of successes (/20)		
between peaks	peak1	peak2	Exponential	Cauchy	Uniform
200	200	50	19	20	20
	150	50	19	20	20
	150	75	18	20	20
	125	50	20	20	20
	100	25	13	15	13
	75	25	17	18	17
	50	25	18	18	17
150	50	200	20	20	20
	50	150	20	20	20
	75	150	20	20	20
	50	125	20	20	20
	25	100	17	16	16
	25	75	18	18	19
	25	50	20	20	20
100	200	50	17	20	19
	150	50	17	20	20
	150	75	19	20	20
	125	50	19	19	19
	100	25	10	12	12
	75	25	11	13	10
	50	25	12	14	12
75	200	50	10	8	8
	150	50	18	14	15
	150	75	19	16	17
	125	50	18	16	14
	100	25	13	1	0
	75	25	11	3	0
	50	25	11	1	0

Table 6. Results from simulated data with three peaks for the RJMCMC schemes with the three priors

Distances between		Intensities	No. of successes (/20)		
first two peaks	second two peaks	of the peaks	Exponential	Cauchy	Uniform
200	200	150	16	20	20
		125	15	20	20
		100	19	20	20
		75	20	19	19
		50	18	19	18
		25	10	3	3
150	150	150	18	20	20
		125	17	20	20
		100	19	20	20
		75	19	19	19
		50	20	18	14
		25	5	0	1
100	100	150	11	1	0
		125	7	0	0
		100	5	0	0
150	200	150	19	20	20
		125	19	17	18
		100	19	20	18
		75	15	17	17
		50	12	3	2
		25	1	0	0

3.5.3 SIMULATION RESULTS USING THE RJMCMC SCHEME

From the results in the previous section, we select Exponential(25) for the intensity parameter and *proposal 1* for updating parameter k. In this subsection, we present detailed results from the simulation study and discuss strengths and limitations of the RJMCMC scheme. Tables 7-8 summarizes the results for simulated data with two peaks and equal intensities, *groups 1-4*.

Table 7. Simulation results from the RJMCMC method for datasets in *group 1* and *group 2*

Peak 1		Peak 2		Background		Peak 1		Peak 2		No. of
μ_1	$\widehat{\mu_1}$ (sd)	μ_2	$\widehat{\mu_2}$ (sd)	ν_0	$\widehat{\nu_0}$ (sd)	ν_1	$\widehat{\nu_1}$ (sd)	ν_2	$\widehat{\nu_2}$ (sd)	successes (/20)
Group 1										
300	299.0 (6.4)	500	496.6 (6.2)	10	13.3 (5.9)	150	144.0 (12.0)	150	152.2 (11.9)	20
300	295.6 (6.9)	500	496.9 (7.0)	10	11.4 (5.5)	125	120.4 (10.7)	125	121.1 (10.9)	20
300	294.7 (7.9)	500	493.5 (7.9)	10	12.4 (5.5)	100	96.3 (9.7)	100	99.0 (9.9)	19
300	295.4 (9.6)	500	495.3 (9.3)	10	13.3 (5.5)	75	70.2 (8.4)	75	73.5 (8.6)	20
300	292.4 (13.5)	500	488.0 (11.9)	10	14.1 (5.5)	50	43.7 (7.2)	50	51.2 (7.3)	20
300	285.2 (21.7)	500	482.1 (20.0)	10	13.1 (5.1)	25	22.3 (5.5)	25	26.0 (5.6)	16
Group 2										
300	291.9 (8.2)	450	443.9 (7.6)	10	10.3 (5.2)	150	138.9 (15.4)	150	158.8 (15.5)	20
300	290.9 (9.6)	450	441.0 (8.5)	10	13.4 (5.8)	125	109.4 (13.9)	125	132.8 (14.4)	20
300	293.4 (9.7)	450	446.2 (9.3)	10	11.1 (5.1)	100	92.7 (12.3)	100	102.8 (12.4)	20
300	286.8 (13.2)	450	437.6 (10.8)	10	11.0 (5.0)	75	62.4 (10.6)	75	83.9 (11.3)	19
300	282.0 (16.4)	450	437.4 (14.0)	10	12.2 (5.2)	50	40.8 (9.0)	50	56.5 (9.3)	20
300	286.5 (42.2)	450	431.3 (30.9)	10	11.8 (5.0)	25	22.2 (8.5)	25	25.5 (8.5)	16

Table 8. Simulation results from the RJMCMC scheme for datasets in *group 3* and *group 4*

Peak 1		Peak 2		Background		Peak 1		Peak 2		No. of
μ_1	$\widehat{\mu_1}$ (sd)	μ_2	$\widehat{\mu_2}$ (sd)	ν_0	$\widehat{\nu_0}$ (sd)	ν_1	$\widehat{\nu_1}$ (sd)	ν_2	$\widehat{\nu_2}$ (sd)	successes (/20)
Group 3										
300	282 (14.6)	400	388.6 (10.2)	10	13.7 (5.7)	150	111.2 (27.2)	150	180.1 (27.6)	20
300	281.2 (16.4)	400	385.4 (11.6)	10	14.5 (5.8)	125	91 (25.7)	125	147.4 (26.2)	19
300	277.5 (22.7)	400	384 (13.2)	10	12 (5.3)	100	69 (24.2)	100	125 (24.6)	18
300	280.4 (30.4)	400	387 (17.5)	10	12 (5.2)	75	57 (22.9)	75	88 (23.2)	19
300	278.4 (39.2)	400	390.9 (24.9)	10	12.4 (5.5)	50	39.2 (18.1)	50	56.3 (18.2)	14
300	262.2 (62.0)	400	397.2 (46.5)	10	11.7 (5.6)	25	21.9 (12.4)	25	26.1 (11.7)	5
Group 4										
275	250.7 (30.9)	350	335.3 (14)	10	12.9 (6.1)	150	92.9 (50.2)	150	193.4 (50.7)	19
275	251.6 (40.8)	350	337.5 (18.6)	10	12.7 (6.1)	125	86.4 (49.8)	125	153.9 (50)	16
275	249.8 (43.4)	350	336.5 (20.8)	10	12.3 (6.2)	100	73.5 (43.1)	100	120.1 (43.4)	15
275	243.6 (53.3)	350	337.8 (23.9)	10	12.7 (6.3)	75	54.7 (35.2)	75	89.7 (35.6)	10
275	241.6 (58.3)	350	342.9 (29.3)	10	10.1 (5.6)	50	40.4 (27)	50	64.1 (27.2)	8

In addition to the number of successes as defined in subsection 3.5.1, we present a summary of the estimates and their standard deviations. As observed in previous subsections, the RJMCMC scheme performs successfully when the distance between the peaks are 100 bp or more and the intensities are higher (50 or more) by estimating the correct number of peaks and the correct locations and intensities. However, as the distance between the peaks and the intensities decrease, the performance starts to deteriorate. This limitation can be expected as it becomes increasingly difficult to distinguish peaks accurately as the distances decrease.

Tables 9-10 summarize results from the two peaks simulation with unequal intensities, *group 5* to *group 8*. When the intensities are unequal, it becomes increasingly difficult to detect the correct number of components and estimate their parameters. In the simulation data, we considered several scenarios of unequal intensities by varying the intensities from high to low as well as varying ratios of the intensities of the two peaks. The main goal of the simulated data in *group 5* and *group 6* is to investigate whether a significant difference in the performance of the RJMCMC can be observed if the order of the intensities given in *group 5* is reversed. From the results given in Table 9, we do not observe a significant difference between the estimates for the two groups. In addition, the intensities and the locations are estimated with a good accuracy even for the simulation data with low intensities.

Continuing with the simulation data on two peaks with unequal intensities, Table 10 presents the results for the simulated datasets in *group 7* and *group 8*, where the distances between the two peaks are decreased. Considering the results, again we observe a deteriorating trend in performances as the distance between the peaks decreases.

Table 9. Simulation results from the RJMCMC scheme for datasets in *group 5* and *group 6*

Peak 1		Peak 2		Background		Peak 1		Peak 2		No. of
μ_1	$\widehat{\mu_1}$ (sd)	μ_2	$\widehat{\mu_2}$ (sd)	ν_0	$\widehat{\nu_0}$ (sd)	ν_1	$\widehat{\nu_1}$ (sd)	ν_2	$\widehat{\nu_2}$ (sd)	successes (/20)
Group 5										
300	299.2 (4.9)	500	496.8 (14.0)	10	13.8 (6.0)	200	197.5 (12.8)	50	47.0 (8.3)	19
300	297.3 (6)	500	493 (12.7)	10	12.5 (5.6)	150	142.7 (11.4)	50	52.4 (8.3)	19
300	298.1 (6.1)	500	494.3 (9.9)	10	13.5 (5.9)	150	143.6 (11.2)	75	73.9 (9.2)	18
300	295.4 (6.7)	500	488.5 (13.3)	10	11.8 (5.5)	125	121 (10.8)	50	50.8 (8.3)	20
300	295 (8)	500	480.6 (22)	10	13.0 (5.7)	100	92 (10.2)	25	29.6 (7.6)	13
300	292.6 (9.8)	500	482.0 (22.1)	10	11.6 (5.5)	75	69.0 (9.0)	25	27.8 (6.8)	17
300	288.2 (11.0)	500	490.1 (17.8)	10	10.5 (4.8)	50	46.9 (6.9)	25	27.3 (5.7)	18
Group 6										
300	289.4 (13.9)	500	499.2 (4.8)	10	12.8 (5.6)	50	44.1 (7.8)	200	194.9 (12.1)	20
300	294.3 (13.1)	500	496.7 (6)	10	12.4 (5.4)	50	47.9 (7.9)	150	145.3 (10.8)	20
300	293.3 (9.5)	500	496.9 (5.9)	10	11.8 (5.4)	75	73.4 (8.9)	150	147 (11)	20
300	291.2 (12.4)	500	498.4 (6.3)	10	10.9 (5.2)	50	48.2 (7.5)	125	126.1 (10.1)	20
300	284.3 (25.4)	500	496.3 (7.6)	10	13.8 (5.3)	25	22.5 (6.3)	100	97.8 (9.2)	17
300	292.2 (20.1)	500	498.1 (8.5)	10	9.7 (4.5)	25	25.2 (6.0)	75	74.8 (7.9)	18
300	276.7 (20.2)	500	494.9 (10.7)	10	12.7 (5.1)	25	23.6 (5.6)	50	52.2 (6.8)	20

Table 10. Simulation results from the RJMCMC scheme for datasets in *group 7* and *group 8*

Peak 1		Peak 2		Background		Peak 1		Peak 2		No. of
μ_1	$\widehat{\mu_1}$ (sd)	μ_2	$\widehat{\mu_2}$ (sd)	ν_0	$\widehat{\nu_0}$ (sd)	ν_1	$\widehat{\nu_1}$ (sd)	ν_2	$\widehat{\nu_2}$ (sd)	successes (/20)
Group 7										
300	295.8 (7.2)	450	430.1 (20.2)	10	12.3 (5.8)	200	188 (20)	50	60.5 (17.1)	17
300	292.2 (8.7)	450	433.8 (18.8)	10	12.8 (5.7)	150	133.8 (16.8)	50	59.4 (14.6)	17
300	295.4 (8.1)	450	441.9 (12.9)	10	11.8 (5.5)	150	136.1 (14.9)	75	81.3 (13.5)	19
300	294.1 (8.5)	450	438.4 (17.3)	10	11.8 (5.5)	125	118 (14.1)	50	57.5 (12.2)	19
300	289.6 (19.6)	450	418.4 (32.2)	10	13.6 (5.8)	100	83 (17.9)	25	36.9 (15.8)	10
300	286.2 (23.3)	450	414.4 (30.2)	10	12.6 (5.6)	75	60.1 (15.2)	25	36.9 (14)	11
300	282.6 (24.4)	450	421.4 (29.3)	10	13.4 (5.8)	50	43.4 (11.5)	25	31.7 (10.2)	12
Group 8										
300	281.5 (31)	400	372 (25.8)	10	13 (5.7)	200	146.2 (48.2)	50	101.4 (47.7)	10
300	271.4 (20.5)	400	367.5 (18.3)	10	12.9 (5.4)	150	95 (30.8)	50	103.2 (30.6)	18
300	280 (18.8)	400	373.4 (16.5)	10	13.5 (5.4)	150	101.2 (32.4)	75	118.3 (32.1)	19
300	274.1 (23.5)	400	369.1 (19.8)	10	13 (5.4)	125	80.3 (27.7)	50	92.9 (27.6)	18
300	273.6 (44.2)	400	364.2 (39.3)	10	14.2 (5.9)	100	64.8 (33.7)	25	58.1 (33.1)	13
300	273.8 (54.2)	400	371.5 (48.1)	10	12.0 (5.5)	75	50.1 (29.1)	50	48.9 (28.5)	11
300	279.4 (63)	400	397.2 (27.9)	10	12.1 (5.1)	50	25.9 (17.1)	25	47.3 (17.4)	11

The estimates from the RJMCMC scheme for the simulated data with three peaks are given in Table 11 and Table 12. In Table 12, the three peaks are simulated to be equally distanced. The RJMCMC scheme performs successfully in estimating the parameters when distances between the peaks are 200 bp, *group 9*. It shows at least 50% of success rate even for a intensity level as low as 25. However, the RJMCMC scheme reaches its limitations as the distance between the peaks decreases below 150 bp, *groups 10-11*. Due to the absence of accurate estimates, some of the subgroups of simulation datasets are not presented here. The datasets in *group 12* were simulated to investigate the performance of the estimation process when the three peaks are at unequal distances. The success rate for many of the subgroups are comparable to those with equal distances (see Table 11).

Table 11. Simulation results from the RJMCMC scheme for datasets in *group 12*

Peak 1		Peak 2		Peak 3		Background		Peak 1		Peak 2		Peak 3		No. of
μ_1	$\widehat{\mu_1}$ (sd)	μ_2	$\widehat{\mu_2}$ (sd)	μ_3	$\widehat{\mu_3}$ (sd)	ν_0	$\widehat{\nu_0}$ (sd)	ν_1	$\widehat{\nu_1}$ (sd)	ν_2	$\widehat{\nu_2}$ (sd)	ν_3	$\widehat{\nu_3}$ (sd)	successes (/20)
Group	*12*													
300	291.6 (8.7)	450	439.5 (10.5)	650	644.8 (6.1)	5	8.7 (5.5)	150	132.7 (15.9)	150.0	150 (15.8)	150	154.6 (12.0)	18
300	297.2 (9.5)	450	445.6 (12.5)	650	648 (7.4)	5	8.1 (5.5)	125	118.6 (15.3)	125	120.3 (15.2)	125	120.2 (11.0)	17
300	296.4 (10.9)	450	442.9 (13.3)	650	647.5 (8.0)	5	7.6 (5.1)	100	93.8 (14.3)	100	103.6 (14.3)	100	97.5 (10.2)	19
300	291.5 (12.7)	450	440.5 (15.6)	650	642.6 (9.4)	5	7.8 (5.3)	75	66.8 (12.2)	75	77.8 (12.1)	75	75.4 (9.1)	19
300	290.9 (19.9)	450	428.3 (21.6)	650	636.6 (11.5)	5	8.8 (5.4)	50	43.0 (11)	50	48.4 (11.3)	50	52.6 (7.6)	20
300	283.7 (29.2)	450	432.3 (32.8)	650	640.9 (19.1)	5	8.2 (5.2)	25	23.5 (8.5)	25	27.7 (8.7)	25	27.4 (5.9)	5

Table 12. Simulation results from the RJMCMC scheme for datasets in *group 9-group 11*

Peak 1		Peak 2		Peak 3		Background		Peak 1		Peak 2		Peak 3		No. of
μ_1	$\widehat{\mu_1}$ (sd)	μ_2	$\widehat{\mu_2}$ (sd)	μ_3	$\widehat{\mu_3}$ (sd)	ν_0	$\widehat{\nu_0}$ (sd)	ν_1	$\widehat{\nu_1}$ (sd)	ν_2	$\widehat{\nu_2}$ (sd)	ν_3	$\widehat{\nu_3}$ (sd)	successes (/20)
Group 9														
300	296 (6.6)	500	493.9 (8.5)	700	696.2 (6.3)	5	7.6 (5.1)	150	142.6 (11.7)	150	144 (12.7)	150	147.8 (11.8)	16
300	297.3 (7.3)	500	490.7 (9.5)	700	693.1 (6.7)	5	6.4 (5)	125	119.7 (11.2)	125	121.4 (12)	125	130.6 (11.2)	15
300	297 (8.1)	500	490 (11.3)	700	692.6 (8)	5	10.7 (5.9)	100	98.2 (10.3)	100	95 (10.9)	100	98.1 (9.8)	19
300	294.5 (9.4)	500	489.4 (12.9)	700	693 (9.5)	5	8.4 (5.5)	75	70.2 (8.9)	75	72.4 (9.5)	75	72.9 (8.9)	20
300	295.3 (11.7)	500	488.6 (15.7)	700	690.6 (11.3)	5	8.3 (5.2)	50	47.6 (7.4)	50	47.0 (8)	50	51.6 (7.6)	18
300	291.5 (25.1)	500	473.2 (34.3)	700	673 (18.9)	5	8.5 (5.1)	25	23.4 (6.5)	25	21.9 (6.6)	25	28.8 (6.1)	10
Group 10														
300	292.7 (9.1)	450	438.9 (14.4)	600	593.9 (8.3)	5	6.8 (4.6)	150	136.9 (17.9)	150	147.5 (19.4)	150	153.4 (17.2)	19
300	293.9 (10.8)	450	433.3 (17.5)	600	589.3 (9.3)	5	9.3 (5.3)	125	113.2 (17.8)	125	117.5 (18)	125	133.9 (16)	19
300	290.3 (12)	450	432.5 (19.8)	600	591.5 (10.7)	5	7.9 (5.1)	100	92.1 (15.8)	100	93.2 (16.2)	100	106.5 (14.5)	19
300	282.6 (15.2)	450	428.6 (19.9)	600	593 (12.5)	5	7.8 (5)	75	60.4 (12.4)	75	78.1 (13.8)	75	78.7 (12.3)	15
300	288.3 (27.3)	450	417.6 (32.8)	600	578 (16.4)	5	11.5 (5.6)	50	42.0 (13.5)	50	47.1 (14.5)	50	58.8 (11)	12
300	271.8 (44.1)	450	400.7 (31.6)	600	575.5 (11.9)	5	6.5 (3.6)	25	14.7 (7.4)	25	25.4 (8.7)	25	37.6 (6.2)	1
Group 11														
300	284.7 (32.6)	400	364.7 (29.8)	500	488.7 (10)	5	8.2 (5.7)	150	108 (58.7)	150	156.7 (60.5)	150	185.2 (32)	11
300	275.2 (45.1)	400	362.4 (31.8)	500	486.0 (12.3)	5	10.4 (6)	125	87.1 (48.7)	125	124.8 (51.1)	125	152.2 (29.7)	7
300	275.3 (38.2)	400	362.5 (29.6)	500	486.7 (12.8)	5	6.7 (5.1)	100	63.2 (38)	100	105 (39.1)	100	122.5 (23.9)	5

Investigating the results from the simulation data revealed many of the strengths and weaknesses of the RJMCMC scheme. The scheme is capable of accurately estimating the parameters when the peaks are apart by 100 bp or more or when this distance is short but the intensities are of higher values.

It is also successful in detecting the correct number of components and estimating their parameters correctly when the intensities of the peaks are unequal. Again the limitation is reached as the distance between the peak decreases around 100 bp. This trend holds for the three peaks.

3.6 RESULTS FROM THE STAT1 AND ZNF143 CHIP-SEQ DATA

With its ability to estimate peaks separated by 100 bp or more, we applied the RJMCMC scheme to the two real ChIP-seq datasets introduced in section 1.5. As described in section 3.1, we first partitioned the genome into smaller regions of 150 bp to 1500 bp. Among these short regions many are expected to have only one binding sites. Therefore, only a subset of the regions were selected for further analysis. The single binding site model (section 2.1) was applied to all the partitioned regions and the p-value for the goodness of fit of the single binding site model was obtained (Kim, Jayatillake, and Spouge 2012). For STAT1 data, regions longer than 600 bp with a goodness of fit p-value less than 0.01 were chosen for further analysis. For ZNF143 data, regions longer than 650 bp with goodness of fit p-value less than 0.01 were chosen for further analysis. There were a total of 906 regions that matched the given criteria for STAT1 ChIP-seq data and 1245 regions that matched the given criteria for ZNF143 ChIP-seq data.

The RJMCMC scheme was applied to the selected regions. The binding sites estimated by the RJMCMC scheme were compared to the motif sites by computing the distance from a binding site to the nearest motif site. The number of binding sites within 50, 100, 200, and 250 bp of a motif site for STAT1 is given in Table 13. In addition, we also compare these results with those obtained from the single binding site model.

For the STAT1 data, the RJMCMC scheme detects an additional number of peaks than those detected by the single biding site model. The highest difference of 30 peaks between the two approaches is observed when the distance to the nearest motif site is 200 bp or less. The number of sites with a motif site in close proximity

estimated by both models do not increase significantly even when the distance to the nearest motif site is increased.

Table 13. Number of binding sites with motif site in close proximity for STAT1 data using the RJMCMC scheme

Distance to the motif site	Number of sites	
	Single binding site model	RJMCMC scheme
50	292	309
100	301	326
200	305	335
250	308	335

For further investigation of the performance of RJMCMC scheme in detecting multiple binding sites, we looked at the tag distribution of the regions with multiple binding sites. Four such regions for STAT1 data are presented in Figures 12-15. In these figures, locations with negative tag counts indicate *unmappable* locations. The chromosome of the region, the relative location of the nearest motif site and the relative location and intensity estimates for the binding sites for the four examples are presented in Tables 14-17. When considering the distribution of tags in these regions, we can clearly observe two peaks, but it is difficult to assert the location of the two peaks as well as the intensities. Among the two peaks in the regions, the RJMCMC scheme accurately detects peaks with high intensities (or those consist of a higher number of tags). As in example 3, illustrated in Figure 14, when both peaks have high intensities, the RJMCMC scheme estimates both peaks with a remarkable accuracy.

In addition to the performance of the RJMCMC scheme, the results of study on STAT1 indicate that there are at most two binding sites among these regions with multiple binding sites with a motif site nearby (200 bp).

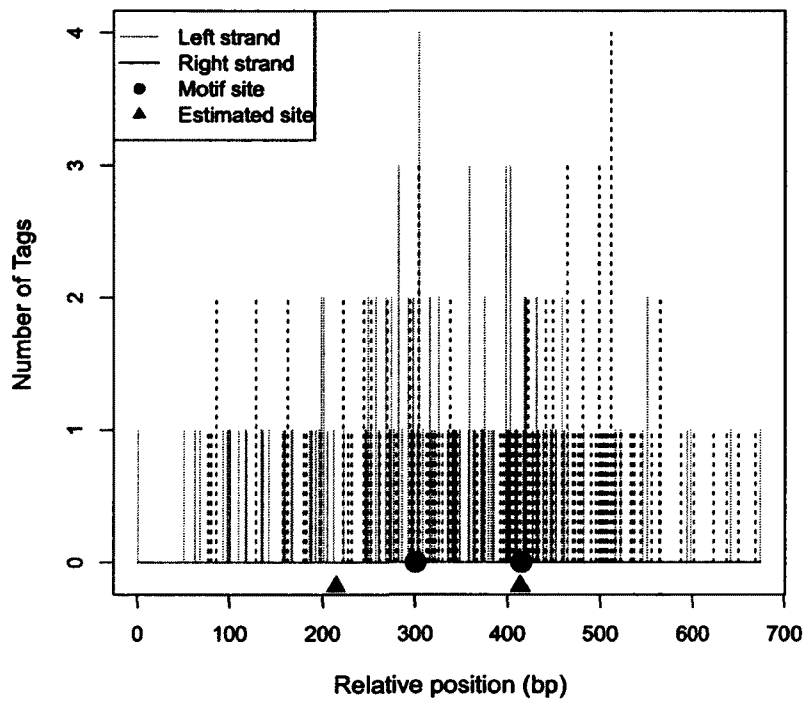


Figure 12. Estimated sites from the RJMCMC scheme and motif sites for STAT1 data for Example 1.

Table 14. Location of the motif site and estimates of the binding sites for STAT1 data using the RJMCMC scheme for Example 1

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
3	107332049	301	215 (25.4)	25 (7.0)
		415	414 (10.4)	70.2 (8.9)

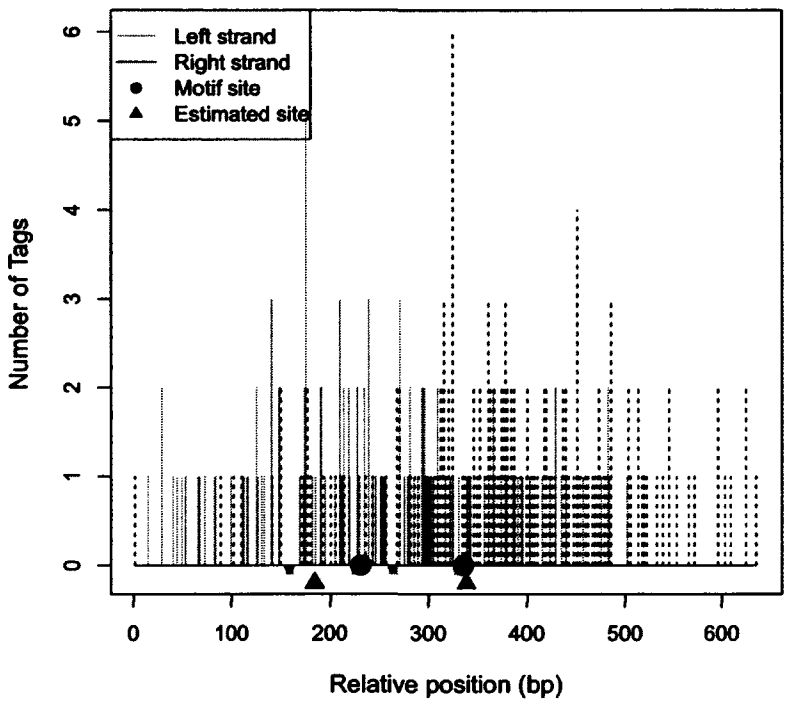


Figure 13. Estimated sites from the RJMCMC scheme and motif sites for STAT1 data for Example 2.

Table 15. Location of the motif site and estimates of the binding sites for STAT1 data using the RJMCMC scheme for Example 2

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
5	83349149	231	184 (53.0)	32.8 (19.6)
		335	338 (16.5)	103.6 (21.3)

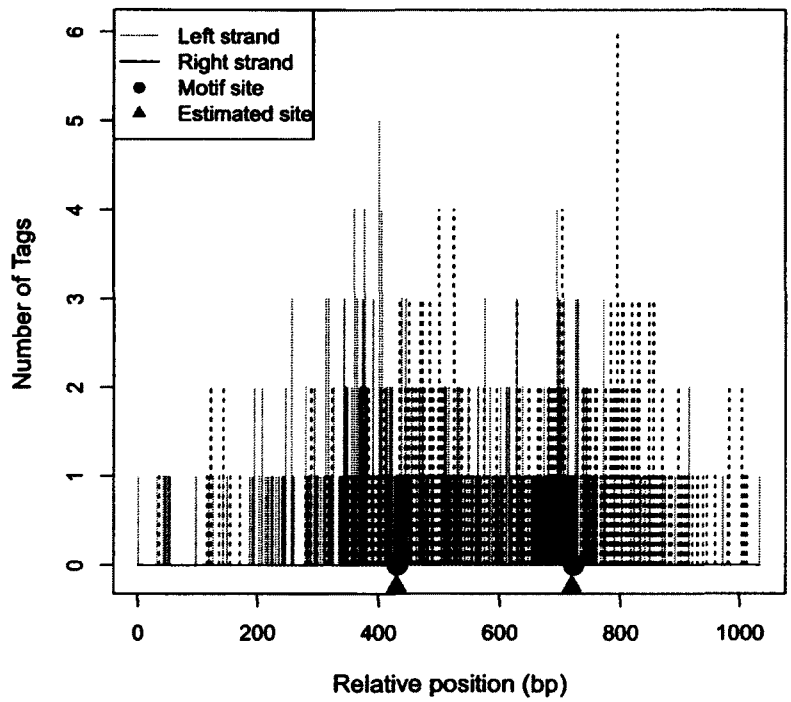


Figure 14. Estimated sites from the RJMCMC scheme and motif sites for STAT1 data for Example 3.

Table 16. Location of the motif site and estimates of the binding sites for STAT1 data using the RJMCMC scheme for Example 3

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
18	1906944	432	430 (5.8)	149.2 (10.8)
		722	719 (5.6)	149.2 (10.9)

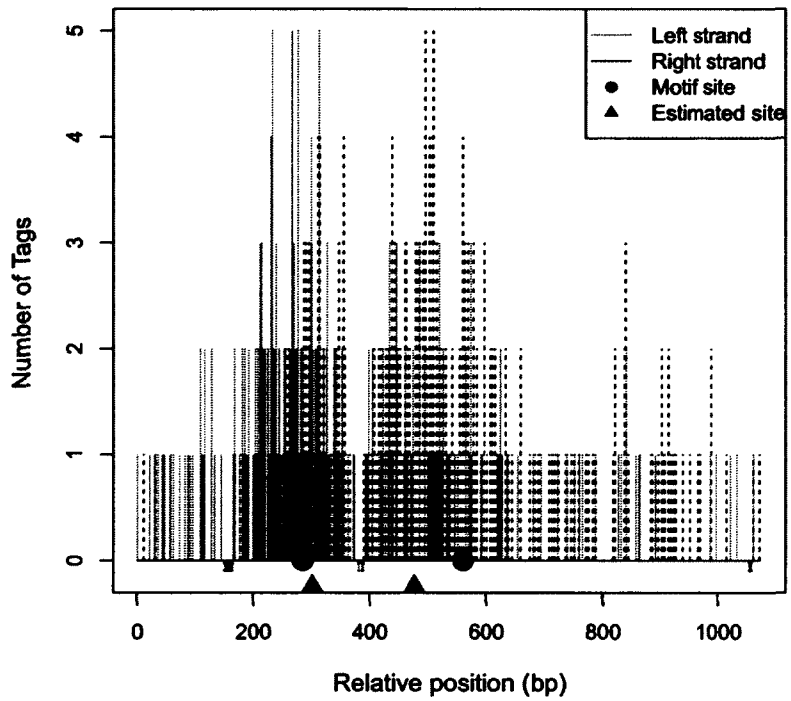


Figure 15. Estimated sites from the RJMCMC scheme and motif sites for STAT1 data for Example 4.

Table 17. Location of the motif site and estimates of the binding sites for STAT1 data using the RJMCMC scheme for Example 4

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
19	39859520	286	302 (7.2)	149.7 (14.5)
		560	477 (9.5)	121.2 (13.4)

Similar to the study of STAT1 data, the RJMCMC scheme was applied to ZNF143 ChIP-seq data. Here also, the number of binding sites estimated by the single site model was compared to the number of binding sites detected by the multiple binding sites model using the RJMCMC scheme, by counting the estimated sites that have motif sites in close proximity (see Table 18). The RJMCMC is able to detect at least 58 more peaks than those detected by the single binding site model. The additional 79 peaks with a motif site within 50 bp were detected by the RJMCMC method. In addition, it predicts the binding locations with a better accuracy than STAT1. Furthermore, for ZNF143 data, higher number of additional peaks were detected than STAT1 data. This can be attributed to the higher number of sequenced tags in ZNF143 ChIP-seq data, about 27 million sequenced tags, compared to 15.1 million sequenced tags in the STAT1 ChIP-seq data.

Table 18. Number of binding sites with motif site in proximity for ZNF143 data using the RJMCMC scheme

Distance to the motif site	Number of sites	
	Single binding site model	RJMCMC scheme
50	369	448
100	397	464
200	412	472
250	417	475

Distributions of the tag counts of some selected regions from ZNF143 with the motif sites and estimated binding sites are given in Figures 16-19. Due to the large number of sequenced tags in the ZNF143 dataset, most regions contain higher number of tags that results in high intensities and higher standard deviation for the estimates. These figures also illustrate that the binding sites predicted by the RJMCMC scheme compares with the overall distribution of the tags. For example, in Figure 17 two peaks of tag distributions with one peak smaller than the other can be clearly observed. The RJMCMC scheme successfully detects the two peaks with smaller intensities for the smaller peak and higher intensity for the more pronounced peak. Figure 19 illustrates the success of the RJMCMC scheme in detecting more than two binding sites in a region with different intensities; infact, it found four.

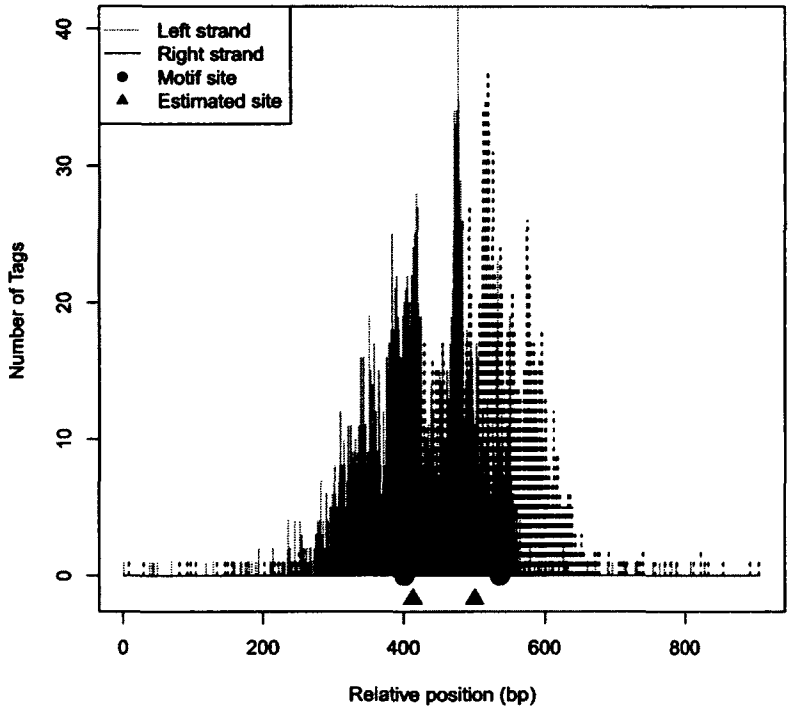


Figure 16. Estimated sites from the RJMCMC scheme and motif sites for ZNF143 data for Example 1.

Table 19. Location of the motif site and estimates of the binding sites for ZNF data using the RJMCMC scheme for Example 1

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
1	242882654	401	413 (26.2)	890.6 (309.8)
		537	501 (1.5)	1979.0 (177.2)

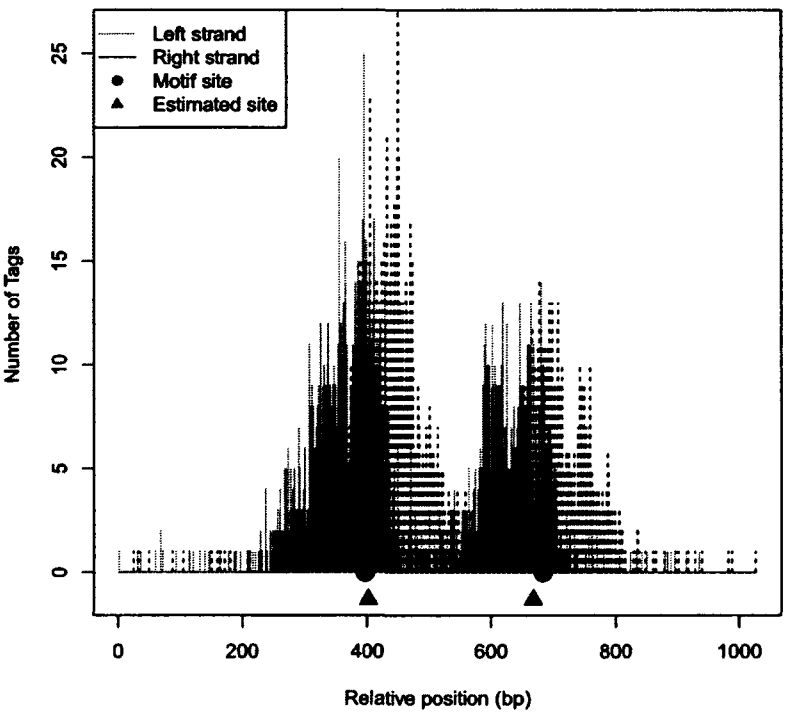


Figure 17. Estimated sites from the RJMCMC scheme and motif sites for ZNF143 data for Example 2.

Table 20. Location of the motif site and estimates of the binding sites for ZNF143 data using the RJMCMC scheme for Example 2

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
14	89867390	396	401 (1.1)	1426.1 (26.7)
		682	666 (1.4)	922.2 (23.1)

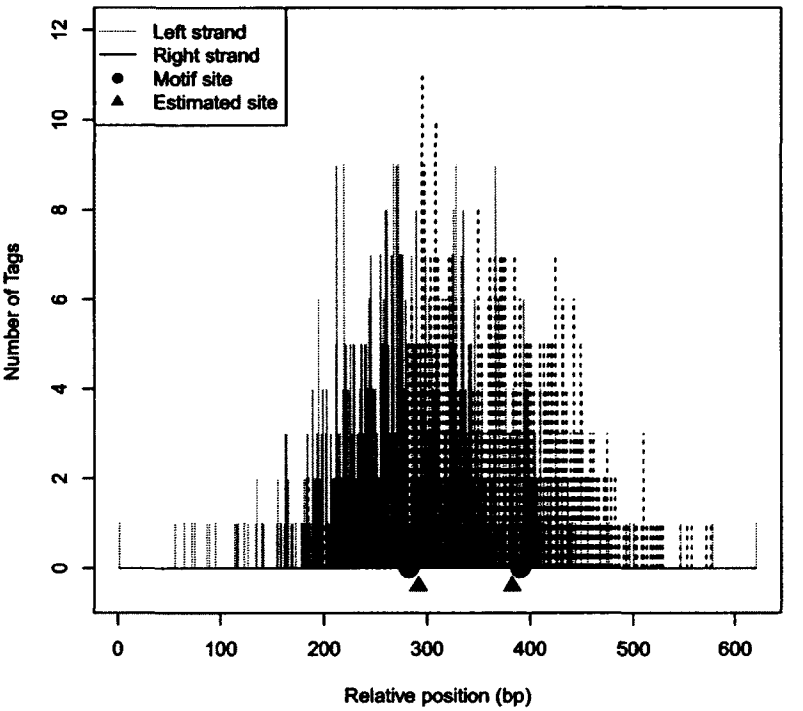


Figure 18. Estimated sites from the RJMCMC scheme and motif sites for ZNF143 data for Example 3.

Table 21. Location of the motif site and estimates of the binding sites for ZNF143 data using the RJMCMC scheme for Example 3

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
19	7522081	282	291 (3.2)	489.3 (29.5)
		390	382 (4.8)	278.9 (27.9)

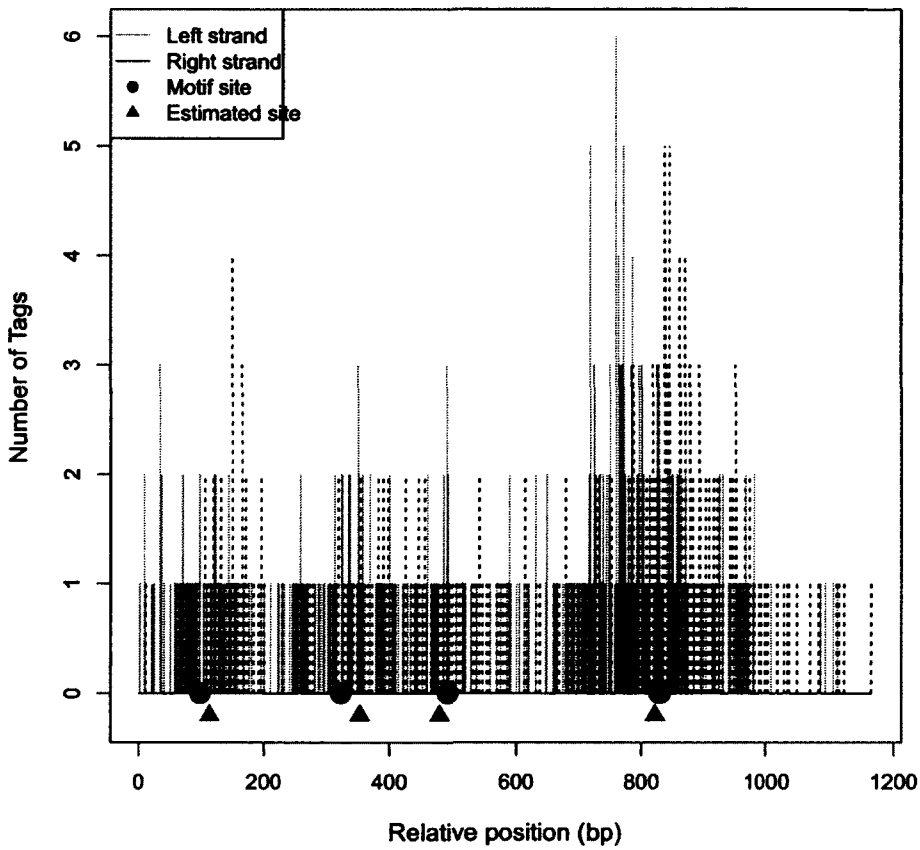


Figure 19. Estimated sites from the RJMCMC scheme and motif sites for ZNF143 data for Example 4.

Table 22. Location of the motif site and estimates of the binding sites for ZNF143 data using the RJMCMC scheme for Example 4

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
19	52627514	98	113 (8.2)	42.8 (26.2)
		323	353 (19.1)	46.2 (20.5)
		492	479 (42.2)	25.1 (22.2)
		828	821 (3.8)	170.7 (22.2)

In analyzing ChIP-seq experiment data for both STAT1 and ZNF143 transcription factors, the multiple binding sites model with the RJMCMC scheme successfully detected more binding sites than the number of binding sites detected by the single binding site model. As observed from the ZNF143 ChIP-seq data, the RJMCMC scheme detects even higher number of binding events in the presence of a large number of sequence tags in the experiment data. In the next chapter, we present the estimation of the parameters of the multiple binding sites using the EM algorithm.

CHAPTER 4

ESTIMATING MULTIPLE BINDING SITES USING EM-ALGORITHM

The basic model to estimate multiple binding events within a region was introduced in section 2.3. The likelihood function of the proposed model given in (8) consists of unobserved tag counts z_{ijh}'s belonging to putative multiple binding events. The maximum likelihood estimates (MLEs) for this particular type of likelihoods with unobserved data cannot be computed directly using the usual maximization methods. The expectation-maximization (EM) algorithm described by Dempster, Laird, and Rubin (1970) computes the MLEs from such likelihoods using two steps: expectation and maximization, iteratively. Since the number of binding sites itself is unknown, we propose estimating it by fitting several models with different numbers of components using the EM algorithm and choosing the best model based on a model selection criterion. The number of components in the chosen model will be considered as the estimated number of binding events within that region.

This chapter presents an application of the EM algorithm on ChIP-seq data to identify multiple binding events within a short region. Section 4.1 of the chapter gives a brief description of the EM algorithm. Its implementation to the current problem is given in section 4.2. Section 4.3 gives a description of the calculation of the asymptotic variance of the MLEs. Section 4.4 describes the model selection criteria considered in this study. In section 4.5, we present results from the simulation data and discuss the effectiveness and limitations of the EM algorithm in detecting and estimating multiple binding sites.

4.1 EXPECTATION-MAXIMIZATION ALGORITHM

This section presents a brief overview of the EM algorithm. More detailed derivation of the theory of the EM algorithm and its applications are given by Dempster, Laird, and Rubin (1970), McLachlan and Krishnan (1997) and McLachlan and Peel (2000).

Consider $LL(\boldsymbol{\theta})$, which consists only of observed data, to be the log likelihood of the observed data and $LL_C(\boldsymbol{\theta})$, which consists of unobserved data as well as observed data, to be the complete log likelihood. Then, in the EM algorithm, iterations begin by setting initial values $\boldsymbol{\theta}^{(0)}$ for the parameters. Then the expectation step is carried out by calculating $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}\right) = E_{\boldsymbol{\theta}^{(0)}}\{LL_C(\boldsymbol{\theta})\}$, where $E_{\boldsymbol{\theta}^{(t)}}\{\cdot\}$ is the expectation evaluated at the current value of the parameter vector. The M-step, or the maximization step, is followed by finding $\boldsymbol{\theta}^{(1)}$, the values of the parameters that maximize $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}\right)$. These two steps are repeated as:

- E-Step: Obtain $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) = E_{\boldsymbol{\theta}^{(t)}}\{LL_C(\boldsymbol{\theta})\}$.
- M-Step: Obtain $\boldsymbol{\theta}^{(t+1)}$ that maximizes $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)$.

Iteration continues until the observed likelihood converges, that is, $LL(\boldsymbol{\theta}^{(t+1)}) - LL(\boldsymbol{\theta}^{(t)}) < \epsilon$, where ϵ is an arbitrary small value.

4.2 EM ALGORITHM FOR THE MULTIPLE BINDING SITES MODEL

The observed data for ChIP-seq are the frequencies of the mapped tags at each location of the i^{th} region of the genome, y_{ij}^L's and y_{ij}^R's. The likelihood function given in (8) for the proposed model consists of unobserved tag counts $\mathbf{z}_{ijh}$'s that belongs to multiple putative binding events. Furthermore, the sum of these unobserved tags is equal to the observed tag count. Therefore, for a given region i, this likelihood function can be considered as the complete log likelihood function $LL_C(\boldsymbol{\theta})$, as in the context of the EM algorithm, where $\boldsymbol{\theta}$ is the vector of the parameters $(\mu_1, \ldots, \mu_k, \nu_0, \ldots, \nu_k)$. The number of components for the model is assumed to be

fixed and is set to be equal to k. Therefore,

$$LL_c(\boldsymbol{\theta}) = \sum_{j \in mappable}^{w_i} \sum_{h=0}^{k} \{-\lambda_{ijh}^L + z_{ijh}^L log(\lambda_{ijh}^L) - log(z_{ijh}^L!) \\ -\lambda_{ijh}^R + z_{ijh}^R log(\lambda_{ijh}^R) - log(z_{ijh}^R!)\}. \quad (34)$$

The EM algorithm is initiated by setting the initial values of the location and intensity parameters as follows:

$$\mu_{ih}^{(0)} = w_i \times \frac{h}{k},$$
$$\nu_{ih}^{(0)} = \frac{\left(\sum_{j \in mappable} y_{ij}^L + \sum_{j \in mappable} y_{ij}^R\right)}{2k},$$
$$\nu_{i0}^{(0)} = 0.05 \times n,$$

where w_i is the length of the i^{th} region, $h = (1, \ldots, k)$ and n is the total number of tags from the left and right strands. In setting these initial values for the location parameters, we considered the peaks to be equally spaced across the region. The initial values of the intensities of the binding events were set to be equal. Since the intensity parameters are similar to the total number of tags in the region, the intensity of the events were initialized by dividing the total number of tags by two times the number of components. We also assume that the background intensity would account for 5% of the total tag count.

4.2.1 E-STEP

In this step, we obtain $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, the expectation of the complete log likelihood function given the observed data y_{ij} and $\boldsymbol{\theta}^{(t)}$, the values of the parameters at the t^{th} iteration:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) =& E_{\boldsymbol{\theta}^{(t)}}(LL_c(\boldsymbol{\theta}|\mathbf{y})) \\ =& \sum_{j \in mappable}^{w_i} \sum_{h=0}^{k} \{-\lambda_{ijh}^L + E_{\boldsymbol{\theta}^t}(z_{ijh}^L|\mathbf{y})\, log(\lambda_{ijh}^L) - E_{\boldsymbol{\theta}^t}(log(z_{ijh}^L)|\mathbf{y}) \\ & -\lambda_{ijh}^R + E_{\boldsymbol{\theta}^t}(z_{ijh}^R|\mathbf{y})\, log(\lambda_{ijh}^R) - E_{\boldsymbol{\theta}^t}(log(z_{ijh}^R)|\mathbf{y})\}. \end{aligned} \quad (35)$$

In section 2.3, it was shown that the distribution of the unobserved tags $\mathbf{z}_{ij}^L$ and $\mathbf{z}_{ij}^R$ given the observed data y_{ij}^L and y_{ij}^R respectively, follows a multinomial distribution

with $p^L_{ijh} = \frac{\lambda^L_{ij0}}{\sum_{h=0}^{k} \lambda^L_{ijh}}$ and $p^R_{ijh} = \frac{\lambda^R_{ij0}}{\sum_{h=0}^{k} \lambda^R_{ijh}}$ respectively. Therefore,

$$E_{\theta^{(t)}}(z^L_{ijh}|\mathbf{y}) = y^L_{ij} \frac{\lambda^L_{ijk}}{\sum_{h=0}^{k} \lambda^L_{ijh}},$$
$$E_{\theta^{(t)}}(z^R_{ijh}|\mathbf{y}) = y^R_{ij} \frac{\lambda^R_{ijk}}{\sum_{h=0}^{k} \lambda^R_{ijh}}.$$

Let $c^L_{ijh} = E_{\theta^{(t)}}(z^L_{ijh}|\mathbf{y})$, $c^R_{ijh} = E_{\theta^{(t)}}(z^R_{ijh}|\mathbf{y})$, $b^L_{ijh} = E_{\theta^t}(log(z^L_{jh})|\mathbf{y})$ and $b^R_{ijh} = E_{\theta^t}(log(z^R_{jh})|\mathbf{y})$. Then,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{j \in mappable}^{w_i} \sum_{h=0}^{k} \left\{ -\lambda^L_{ijh} + c^L_{ijh} log(\lambda^L_{ijh}) - b^L_{ijh} - \lambda^R_{jh} + c^R_{ijh} log(\lambda^R_{ijh}) - b^R_{ijh} \right\}. \tag{36}$$

4.2.2 M-STEP

In this step, the values of the parameters that maximize (36) are determined by setting the score functions of each parameter to zero. Let us first consider the intensity parameter ν_{ih}, where $h = (0, \ldots, k)$. The score function of ν_{ih} is derived by taking the first derivative of $LL_C(\boldsymbol{\theta})$ with respect to ν_{ih} as follows:

$$\begin{aligned}
\frac{\partial}{\partial \nu_{ih}} \left(Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \right) &= \frac{\partial}{\partial \nu_{ih}} \left(\sum_{j \in mappable}^{w_i} \left\{ -\nu_h f_L(j|\mu_{ih}, \sigma^2, \beta) + c^L_{ijh} log(\nu_h f_L(j|\mu_{ih}, \sigma^2, \beta)) \right. \right. \\
&\quad \left. \left. - b^L_{ijh} - \nu_h f_R(j|\mu_{ih}, \sigma^2, \beta) + c^R_{ijh} log(\nu_{ih} f_R(j|\mu_{ih}, \sigma^2, \beta)) - b^R_{ijh} \right\} \right) \\
&= \sum_{j \in mappable}^{w_i} \left\{ -(f_L(j|\mu_{ih}, \sigma^2, \beta) + f_R(j|\mu_{ih}, \sigma^2, \beta)) + \frac{1}{\nu_{ijh}} (c^L_{ijh} + c^R_{ijh}) \right\}.
\end{aligned}$$

By setting the above function to zero, we obtain

$$\nu^{(t+1)}_{ih} = \frac{\sum_{j \in mappable}^{w_i} (c^L_{ijh} + c^R_{ijh})}{\sum_{j \in mappable}^{w_i} \left(f_L(j|\mu^{(t)}_{ih}, \sigma^2, \beta) + f_R(j|\mu^{(t)}_{ih}, \sigma^2, \beta) \right)}. \tag{37}$$

For the background component, since we assumed the tags to be distributed uniformly over the region, (37) simplifies to

$$\nu^{(t+1)}_{i0} = \frac{l}{2w_i} \sum_{j \in mappable}^{w_i} \left(c^L_{ijo} + c^R_{ijo} \right),$$

where l is the number of mappable locations in the i^{th} region. Similarly for binding location parameter,

$$\frac{\partial}{\partial \mu_{ih}}\left(Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right)\right) = \sum_{j \in mappable}^{w_i} \left\{ -\frac{\partial \lambda^L_{ijh}}{\partial \mu_{ih}} + c^L_{ijh}\frac{\partial \lambda^L_{ijh}}{\partial \mu_{ih}} \times \frac{1}{\partial \lambda^L_{ijh}} + \right.$$
$$\left. -\frac{\partial \lambda^R_{ijh}}{\partial \mu_{ih}} + c^R_{ijh}\frac{\partial \lambda^R_{ijh}}{\partial \mu_{ih}} \times \frac{1}{\partial \lambda^R_{ijh}} \right\}, \qquad (38)$$

where

$$\frac{\partial \lambda^L_{ijh}}{\partial \mu_{ih}} = \nu_{ih}\frac{\partial f_L\left(\mu_{ih}, \sigma^2, \beta\right)}{\partial \mu_{ih}} = \nu_{ih} f'_L(j|\mu_{ih}, \sigma^2, \beta) \qquad (39)$$

and

$$\frac{\partial \lambda^R_{ijh}}{\partial \mu_{ih}} = \nu_{ih}\frac{\partial f_R\left(\mu_{ih}, \sigma^2, \beta\right)}{\partial \mu_{ih}} = \nu_{ih} f'_R(j|\mu_{ih}, \sigma^2, \beta). \qquad (40)$$

Furthermore,

$$f'_L(j|\mu_{ih}, \sigma^2, \beta) = \frac{1}{\sigma}\phi\left(\frac{j - \mu_{ih} + \frac{\sigma^2}{\beta}}{\sigma}\right)\frac{1}{\beta}exp\left\{\frac{1}{\beta}\left(j - \mu_{ih} + \frac{\sigma^2}{2\beta}\right)\right\}$$
$$- \Phi\left(1 - \frac{j - \mu_{ih} + \frac{\sigma^2}{\beta}}{\sigma}\right)\frac{1}{\beta^2}exp\left\{\frac{1}{\beta}\left(j - \mu_{ih} + \frac{\sigma^2}{2\beta}\right)\right\} \qquad (41)$$

and

$$f'_R(j|\mu_{ih}, \sigma^2, \beta) = -\frac{1}{\sigma}\phi\left(\frac{j - \mu_{ih} - \frac{\sigma^2}{\beta}}{\sigma}\right)\frac{1}{\beta}exp\left\{-\frac{1}{\beta}\left(j - \mu_{ih} - \frac{\sigma^2}{2\beta}\right)\right\}$$
$$+ \Phi\left(\frac{j - \mu_{ih} - \frac{\sigma^2}{\beta}}{\sigma}\right)\frac{1}{\beta^2}exp\left\{-\frac{1}{\beta}\left(j - \mu_{ih} - \frac{\sigma^2}{2\beta}\right)\right\}. \qquad (42)$$

Using the results given in (39)-(42) and substituting $\nu^{(t+1)}_{ih}$ for ν_{ih}, we set (38) to zero and obtained the following nonlinear equation:

$$-\nu^{(t+1)}_{ih}\left\{\sum_{j \in mappable}^{w_i}\left\{f'_L(j|\mu_{ih}, \sigma^2, \beta) + f'_R(j|\mu_{ih}, \sigma^2, \beta)\right\}\right\}$$
$$+ \sum_{j \in mappable}^{w_i}\left\{c^L_{ijh} \times \frac{f'_L(j|\mu_{ih}, \sigma^2, \beta)}{f_L(j|\mu_{ih}, \sigma^2, \beta)} + c^R_{ijh} \times \frac{f'_R(j|\mu_{ih}, \sigma^2, \beta)}{f_R(j|\mu_{ih}, \sigma^2, \beta)}\right\} = 0. \qquad (43)$$

Value of $\mu^{(t+1)}_{ih}$ is obtained by solving (43) numerically for μ_{ih}. Here we use a simple root finding algorithm, the bi-section method (Press et al. 2002), to obtain $\mu^{(t+1)}_{ih}$.

These expectation and maximization steps are repeated until the value of the observed log likelihood is converged. Once all the MLEs of the parameters are determined, their asymptotic variances can be computed as described in the next section.

4.3 ASYMPTOTIC VARIANCE OF THE MLES

The asymptotic variance of the MLEs can be computed using the Fisher's information which is the negative of the expected value of Hessian of the log likelihood (McLachlan and Peel 2000). The Fisher's information for the parameters with k components is computed by

$$\mathcal{I}(\boldsymbol{\theta}) = -E\begin{bmatrix} \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{i0}^2} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{i0}\partial \mu_{i1}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{i0}\partial \nu_{i1}} & \cdots & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{i0}\partial \mu_{ik}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{i0}\partial \nu_{ik}} \\ \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \mu_{i1}\partial \nu_{i0}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial^2 \mu_{i1}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \mu_{i1}\partial \nu_{i1}} & \cdots & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \mu_{i1}\partial \mu_{ik}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \mu_{i1}\partial \nu_{ik}} \\ \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{i1}\partial \nu_{i0}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{i1}\partial \mu_{i1}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial^2 \nu_{i1}} & \cdots & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{i1}\partial \mu_{ik}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{1i}\partial \nu_{ik}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \mu_{ik}\partial \nu_{i0}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \mu_{ik}\partial \mu_{i1}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \mu_{ik}\partial \nu_{i1}} & \cdots & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial^2 \mu_{ik}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \mu_{ik}\partial \nu_{ik}} \\ \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{ik}\partial \nu_{i0}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{ik}\partial \mu_{i1}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{ik}\partial \nu_{i1}} & \cdots & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{ik}\partial \mu_{ik}} & \frac{\partial^2 LL(\boldsymbol{\theta})}{\partial^2 \nu_{ik}} \end{bmatrix}, \tag{44}$$

where $LL(\boldsymbol{\theta})$ is the log likelihood function of the observed data y_{ij}'s of the i^{th} region. The elements of the Information matrix are as follows:

$$-E\left(\frac{\partial^2 LL(\boldsymbol{\theta})}{\partial^2 \nu_{io}}\right) = \frac{1}{w_i^2}\sum_{j \in mappable}^{w_i}\left\{\frac{1}{\sum_{h=0}^{k}\lambda_{ijh}^{L}} + \frac{1}{\sum_{h=0}^{k}\lambda_{ijh}^{R}}\right\}.$$

$$\begin{aligned} -E\left(\frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{io}\partial \mu_{ih}}\right) &= -E\left(\frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \mu_{ih}\partial \nu_{i0}}\right) \\ &= \frac{\nu_{ih}}{w_i}\sum_{j \in mappable}^{w_i}\left\{\frac{f_L'(j|\mu_{ih},\sigma^2,\beta)}{\sum_{h=0}^{k}\lambda_{ijh}^{L}} + \frac{f_R'(j|\mu_{ih},\sigma^2,\beta)}{\sum_{h=0}^{k}\lambda_{ijh}^{R}}\right\}. \end{aligned}$$

$$-E\left(\frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{io}\partial \nu_{ih}}\right) = -E\left(\frac{\partial^2 LL(\boldsymbol{\theta})}{\partial \nu_{ih}\partial \nu_{i0}}\right)$$
$$= \frac{1}{w_i}\sum_{j\in mappable}^{w_i}\left\{\frac{f_L(j|\mu_{ih},\sigma^2,\beta)}{\sum_{h=0}^{k}\lambda_{ijh}^{L}} + \frac{f_R(j|\mu_{ih},\sigma^2,\beta)}{\sum_{h=0}^{k}\lambda_{ijh}^{R}}\right\}.$$

$$-E\left(\frac{\partial^2 LL(\boldsymbol{\theta})}{\partial^2 \mu_{ih}^2}\right) = \nu_{ih}\sum_{j\in mappable}^{w_i}\left\{\frac{(f_L'(j|\mu_{ih},\sigma^2,\beta))^2}{\sum_{h=0}^{k}\lambda_{ijh}^{L}} + \frac{(f_R'(j|\mu_{ih},\sigma^2,\beta))^2}{\sum_{h=0}^{k}\lambda_{ijh}^{R}}\right\}.$$

$$-E\left(\frac{\partial^2 LL_c(\boldsymbol{\theta})}{\partial \mu_{ih}\partial \mu_{ig}}\right) = -E\left(\frac{\partial^2 LL_c(\boldsymbol{\theta})}{\partial \mu_{ig}\partial \mu_{ih}}\right)$$
$$= \nu_h\nu_g\sum_{j\in mappable}^{w_i}\left\{\frac{f_L'(j|\mu_{ih},\sigma^2,\beta)f_L'(j|\mu_{ig},\sigma^2,\beta)}{\sum_{h=0}^{k}\lambda_{ijh}^{L}} + \frac{f_R'(j|\mu_{ih},\sigma^2,\beta)f_R'(j|\mu_{ig},\sigma^2,\beta)}{\sum_{h=0}^{k}\lambda_{ijh}^{R}}\right\}.$$

$$-E\left(\frac{\partial^2 LL_c(\boldsymbol{\theta})}{\partial \mu_{ih}\partial \nu_{ih}}\right) = -E\left(\frac{\partial^2 LL_c(\boldsymbol{\theta})}{\partial \nu_{ih}\partial \mu_{ih}}\right)$$
$$= \nu_h\sum_{j\in mappable}^{w_i}\left\{\frac{f_L'(j|\mu_{ih},\sigma^2,\beta)f_L(j|\mu_{ih},\sigma^2,\beta)}{\sum_{h=0}^{k}\lambda_{ijh}^{L}} + \frac{f_R'(j|\mu_{ih},\sigma^2,\beta)f_R(j|\mu_{ih},\sigma^2,\beta)}{\sum_{h=0}^{k}\lambda_{ijh}^{R}}\right\}.$$

$$-E\left(\frac{\partial^2 LL_c(\boldsymbol{\theta})}{\partial \mu_{ih}\partial \nu_{ig}}\right) = -E\left(\frac{\partial^2 LL_c(\boldsymbol{\theta})}{\partial \nu_{ig}\partial \mu_{ih}}\right)$$
$$= \nu_h\sum_{j\in mappable}^{w_i}\left\{\frac{f_L'(j|\mu_{ih},\sigma^2,\beta)f_L(j|\mu_{ig},\sigma^2,\beta)}{\sum_{h=0}^{k}\lambda_{ijh}^{L}} + \frac{f_R'(j|\mu_{ih},\sigma^2,\beta)f_R'(j|\mu_{ig},\sigma^2,\beta)}{\sum_{h=0}^{k}\lambda_{ijh}^{R}}\right\}.$$

$$-E\left(\frac{\partial^2 LL_c(\boldsymbol{\theta})}{\partial \nu_{ih}^2}\right) = \sum_{j\in mappable}^{w_i}\left\{\frac{(f_L(j|\mu_{ih},\sigma^2,\beta))^2}{\sum_{h=0}^{k}\lambda_{ijh}^{L}} + \frac{(f_R(j|\mu_{ih},\sigma^2,\beta))^2}{\sum_{h=0}^{k}\lambda_{ijh}^{R}}\right\}.$$

$$-E\left(\frac{\partial^2 LL_c(\boldsymbol{\theta})}{\partial \nu_{ih}\partial \nu_{ig}}\right) = -E\left(\frac{\partial^2 LL_c(\boldsymbol{\theta})}{\partial \nu_{ig}\partial \nu_{ih}}\right)$$
$$= \sum_{j\in mappable}^{w_i}\left\{\frac{f_L(j|\mu_{ih},\sigma^2,\beta)f_L(j|\mu_{ig},\sigma^2,\beta)}{\sum_{h=0}^{k}\lambda_{ijh}^{L}} + \frac{f_R(j|\mu_{ih},\sigma^2,\beta)f_R(j|\mu_{ig},\sigma^2,\beta)}{\sum_{h=0}^{k}\lambda_{ijh}^{R}}\right\}.$$

The asymptotic variance-covariance matrix is obtained by taking the inverse of

the Fisher's information matrix evaluated at the MLEs.

4.4 MODEL SELECTION CRITERIA

One of the key attributes of the EM algorithm is that the number of components of the mixture model needs to be fixed and known. However, in ChIP-seq experiment data the number of components or binding sites, within a region is unknown, and needs to be estimated. To overcome this limitation, we apply the EM algorithm on several models with different numbers of components ranging from one to k_{max}=3 and select the best model among them using a model selection criterion. The maximum number of components need not to be restricted to three as proposed here, but we assume that the possibility to exceed this upper limit is unlikely.

The usual approach of likelihood ratio test to select the number of components is unavailable as the regularity conditions do not hold for mixture models (McLachlan and Krishnan 1997 and McLachlan and Peel 2000). However, in statistical literature there are many other model selection criteria based on the likelihood. McLachlan and Peel (2000) presents a comprehensive summary of criteria for assessing the number of components of a mixture model. Most frequently used criteria are the Akaike information criterion (AIC), and Bayesian information criterion (BIC). Both of these criteria are based on penalized form of the likelihood. Usually the likelihood increases with the addition of components to a mixture model, which often leads to over-fitting. The AIC and BIC methods penalize the likelihood by subtracting a term that depends on the number of parameters in the model, thereby overcoming over-fitting. Following are the formulae to calculate AIC and BIC:

$$AIC = 2p - 2\,logL$$
$$BIC = p\,log(n) - 2\,logL,$$

where p is the number of parameters in the model, $log\,L$ is the maximized value of the log likelihood function ($LL_C(\hat{\boldsymbol{\theta}})$), n is the total number of observations. Since the BIC penalizes the likelihood more than the AIC, BIC tends to select models with fewer components. Due to the large number of regions that needs to be analyzed in a genome-wide study, efficiency and simplicity in calculating a selection criteria plays a very important role in overall efficiency of the estimation process. Many of the other penalized likelihood criterion would require additional computations in terms of the

information matrix or entropy. For this particular study we investigate only the AIC and BIC criteria. Once the best model is selected, the number of components (k) and the parameter estimates for the selected model are considered as the estimates for the number of binding events and the corresponding parameters of the binding events.

4.5 EM ALGORITHM ON SIMULATED DATA

The success of the above scheme using the EM algorithm mainly depends on determining the number of components accurately. Therefore the EM algorithm based scheme is applied on several simulated datasets, where the results can be compared with the true values of the number of components as well as the parameters. These simulated datasets are the same as the ones used in the study of the RJMCMC scheme.

4.5.1 ASSESSMENT OF SELECTION CRITERIA USING SIMULATED DATA

For each simulated dataset, the EM algorithm fits several models with the number of components ranging from one to three and selects the best model among them based on AIC or BIC. Table 23 gives the percentage of the number of times each selection criteria selected the correct number of components or peaks. From the table it can be observed that the success percentage is much higher for AIC than for BIC, indicating that AIC is able to determine the number of components more accurately. One of the known shortcoming of AIC in statistical literature, including McLachlan and Peel (2000), is that it tends to over-fit the number of components. Here the simulation *group* 1 to *group* 8 are simulated with two components and the EM scheme fits up to three component models. The AIC selects the two components model over the three components. Although BIC performs better in simulation *groups* of 1, 2, 5, 6, 9 and 10, where the peaks are well separated and the intensities of the peaks are high, it fails to select the correct model in other scenarios when the peaks are closer to each other and have lower intensities.

Even for AIC, the success percentage decreases dramatically for simulated datasets in *groups 4*, *8*, and *11*. The datasets in *groups 4* and *8* are simulated with two peaks separated by 75 bp and the datasets in *group 11*, for which the lowest

Table 23. Percentage of correct selections

Simulation	Success percentage	
	AIC	BIC
group 1	100.0	95.8
group 2	98.3	87.5
group 3	80.8	42.5
group 4	46.7	1.7
group 5	99.3	97.9
group 6	99.3	92.1
group 7	97.9	75.0
group 8	40.0	1.4
group 9	100.0	92.5
group 10	95.0	80.8
group 11	15.8	0.0
group 12	85.8	65.8

percentage is observed, is simulated with three peaks separated by 100 bp. It can be concluded that the AIC method outperforms BIC on average. Therefore, hereafter, we use only the AIC model selection criterion.

4.5.2 RESULTS FROM THE SIMULATION STUDY

Table 23 presents only percentages of selecting correct number of components for each of the simulation datasets. As described in section 1.9, each simulation *group* has 4 or 5 unique simulation scenarios with different intensities. Furthermore, each of those scenarios have 20 simulated datasets or samples. In this section accuracy of the estimates of the selected models is also presented. For ease of comparison, the results from the 20 samples of each scenario are accumulated by taking the average of the estimates of the *correctly estimated* models. Here models with the correct number of components and the estimates of the location parameter within 50 bp of the true values are considered as *correctly estimated* models.

Tables 24 and 25 present the summary for results from simulation datasets in *groups 1-4* that are generated by setting the number of peaks to be two and the intensity of the peaks to be equal. The last column of these tables gives the number of times the model is *correctly estimated* out of the 20 samples of each simulation

scenario. The EM algorithm scheme successfully identifies the correct model and estimates the parameters accurately almost for all the simulated datasets, where the peaks are far apart. This success rate decreases as the peaks get closer to each other and the intensities of the events decreases.

Table 24. Simulation results from the EM algorithm for *group 1* and *group 2*

Peak 1		Peak 2		Background		Peak 1		Peak 2		No. of
μ_1	$\widehat{\mu_1}$ (sd)	μ_2	$\widehat{\mu_2}$ (sd)	ν_0	$\widehat{\nu_0}$ (sd)	ν_1	$\widehat{\nu_1}$ (sd)	ν_2	$\widehat{\nu_2}$ (sd)	successes (/20)
Group 1										
300	302.2 (6.2)	500	499.7 (6.2)	10	11.9 (6.9)	150	149.8 (11)	150	152.3 (11)	20
300	299.0 (6.6)	500	500.3 (6.8)	10	9.8 (6.1)	125	124.9 (9.9)	125	121.7 (9.7)	20
300	298.7 (7.6)	500	497.7 (7.7)	10	10.4 (6)	100	100.5 (9)	100	98.9 (8.9)	20
300	300.6 (8.9)	500	501.5 (9.1)	10	11.1 (6.3)	75	74.2 (7.8)	75	73.1 (7.7)	20
300	302.2 (11.9)	500	497.5 (11.4)	10	11.7 (6.1)	50	47.5 (6.5)	50	50.3 (6.6)	20
300	300.3 (16.7)	500	499.6 (17.0)	10	10.5 (5.4)	25	25.4 (4.8)	25	24.9 (4.8)	20
Group 2										
300	297.3 (7.6)	450	448.9 (7.5)	10	8.5 (5.6)	150	149.1 (14.5)	150	154.7 (14.5)	20
300	299.0 (8.8)	450	448.0 (8.6)	10	11.7 (6.5)	125	120.4 (13.7)	125	126.1 (13.6)	19
300	299.3 (9.3)	450	452.6 (9.5)	10	9.3 (5.8)	100	100.0 (11.8)	100	99.7 (11.7)	20
300	298.7 (11.6)	450	448.1 (11)	10	8.9 (5.6)	75	72.1 (10.6)	75	78.0 (10.7)	20
300	297.2 (14)	450	452.6 (13.6)	10	9.7 (5.5)	50	48.8 (8.5)	50	51.5 (8.5)	20
300	301.9 (20)	450	454.6 (21.7)	10	8.8 (5)	25	26.8 (6.5)	25	24.4 (6.3)	19

Table 25. Simulation results from the EM algorithm for *group 3* and *group 4*

Peak 1		Peak 2		Background		Peak 1		Peak 2		No. of
μ_1	$\widehat{\mu_1}$ (sd)	μ_2	$\widehat{\mu_2}$ (sd)	ν_0	$\widehat{\nu_0}$ (sd)	ν_1	$\widehat{\nu_1}$ (sd)	ν_2	$\widehat{\nu_2}$ (sd)	successes (/20)
Group 3										
300	299.8 (11.9)	400	402.1 (12.3)	10	11.7 (6.4)	150	150.6 (32)	150	147.2 (31.7)	20
300	298.8 (13.7)	400	399.9 (14)	10	12.3 (6.4)	125	123 (30.5)	125	122.3 (30.2)	20
300	298.3 (15.2)	400	400.6 (15.4)	10	9.8 (5.8)	100	98.5 (27)	100	100.8 (26.8)	19
300	298.6 (16.6)	400	405.8 (17.2)	10	8.9 (5.5)	75	76.2 (21.5)	75	74.6 (21.2)	17
300	290.8 (19.7)	400	405.5 (19.3)	10	9.2 (5.3)	50	48.7 (14.9)	50	51.9 (14.8)	13
300	276.0 (23.2)	400	414.5 (22)	10	7.0 (4.5)	25	24.3 (7.6)	25	25.9 (7.6)	4
Group 4										
275	269.7 (19.1)	350	348.8 (16.5)	10	9.6 (6.3)	150	133.8 (61.4)	150	160.3 (61.1)	16
275	266.9 (20)	350	347.9 (16.8)	10	8.3 (6.1)	125	113.3 (52.9)	125	141.9 (52.7)	12
275	263.3 (21.5)	350	349.1 (17.8)	10	7.8 (5.6)	100	86.5 (41.2)	100	111.6 (41)	15
275	257.0 (21.2)	350	351.7 (17.7)	10	7.8 (5.6)	75	69.7 (29.1)	75	88.1 (28.9)	6
275	247.7 (24.3)	350	351.0 (18.5)	10	7.2 (4.8)	50	43.1 (18.6)	50	62.1 (18.7)	6

The same pattern can be observed in Tables 26 and 27 which present the summary for results from datasets simulated by setting the number of components to be two and the peaks to have unequal intensities (*groups 5-8*). However, the success rate is lower for the unequal intensities than for the equal intensities when the peaks are separated by 100 bp. The simulation datasets in *group 6* are generated to investigate whether a significant difference in the performance of the EM algorithm can be observed when the order of the intensities of the peaks are reversed such that the first peak has the smaller intensity as opposed to simulations, where the first peak has the higher intensity. From the comparison between the results from simulation datasets in *group 5* and *group 6*, it can be concluded that the order of magnitude of the intensities does not affect the estimation.

Results from the final sets of simulations (simulated with three peaks) are presented in Tables 28 and 29. Here we observe the same trend as for the two peaks simulations. Unlike in the two peaks simulations, the success rate is significantly lower when the peaks are separated by 100 bp even when the intensities are higher. This can be expected as it becomes increasingly difficult to distinguish all the peaks. Moreover, the estimates of the locations as well as the intensities are less accurate than those estimates when the peaks are far apart.

The simulation *group 12* given in Table 29 presents the results from the EM algorithm when the peaks are separated by unequal distances. Since the peaks are separated with sufficient distance the success rate is almost 100% except when the intensity of the peaks are as low as 25.

Table 26. Simulation results from the EM algorithm for *group 5* and *group 6*

Peak 1		Peak 2		Background		Peak 1		Peak 2		No. of
μ_1	$\widehat{\mu_1}$ (sd)	μ_2	$\widehat{\mu_2}$ (sd)	ν_0	$\widehat{\nu_0}$ (sd)	ν_1	$\widehat{\nu_1}$ (sd)	ν_2	$\widehat{\nu_2}$ (sd)	successes (/20)
Group 5										
300	300.9 (4.7)	500	506.5 (13.3)	10	11.4 (6.6)	200	204.4 (11.5)	50	46.2 (7)	20
300	300.0 (5.8)	500	500.0 (12.4)	10	10.5 (5.9)	150	147.7 (10.1)	50	50.8 (7.2)	20
300	300.5 (5.9)	500	499.1 (9.6)	10	11.3 (6.4)	150	149.5 (10.4)	75	73.9 (8.3)	19
300	298.6 (6.3)	500	497.6 (12.2)	10	9.5 (5.7)	125	126.1 (9.5)	50	49.8 (7)	20
300	299.4 (7.3)	500	489.8 (18)	10	10.1 (5.6)	100	97.8 (8.6)	25	28.9 (6)	20
300	297.8 (8.3)	500	499.7 (17.5)	10	9.5 (5.3)	75	73.8 (7.3)	25	26.5 (5.3)	20
300	295.9 (10.3)	500	503.5 (16.9)	10	8.1 (5.2)	50	50.2 (6.1)	25	25.9 (4.9)	20
Group 6										
300	299.8 (13)	500	502.2 (4.8)	10	11.2 (6.5)	50	47.8 (7.3)	200	196 (11.3)	20
300	304 (12.1)	500	500.2 (5.9)	10	10.1 (6.2)	50	51.5 (7.4)	150	146.2 (10.1)	19
300	298.7 (9.2)	500	500.2 (5.9)	10	9.9 (6.1)	75	77.2 (8.3)	150	148 (10.3)	20
300	299.4 (11.7)	500	502 (6.2)	10	9 (5.8)	50	51.4 (7)	125	126.7 (9.4)	20
300	301.8 (19.7)	500	501.1 (7.1)	10	11.4 (5.9)	25	25 (5.5)	100	98.6 (8.3)	19
300	304.5 (17.4)	500	503.2 (8.3)	10	8.1 (4.9)	25	27.2 (5.4)	75	74.4 (7.3)	20
300	294 (16.6)	500	503.5 (10.3)	10	10.3 (5.5)	25	26.5 (5)	50	51.6 (6.1)	20

Table 27. Simulation results from the EM algorithm for *group 7* and *group 8*

Peak 1		Peak 2		Background		Peak 1		Peak 2		No. of
μ_1	$\widehat{\mu_1}$ (sd)	μ_2	$\widehat{\mu_2}$ (sd)	ν_0	$\widehat{\nu_0}$ (sd)	ν_1	$\widehat{\nu_1}$ (sd)	ν_2	$\widehat{\nu_2}$ (sd)	successes (/20)
Group 7										
300	299.8 (6)	450	445.6 (17.4)	10	10.1 (5.9)	200	200.5 (15.2)	50	52.7 (12.2)	20
300	297.9 (7.3)	450	447.1 (16.2)	10	10.8 (6)	150	145.2 (13.3)	50	52.5 (11.1)	20
300	300 (7.3)	450	450.6 (12.1)	10	9.6 (5.9)	150	146.3 (13.4)	75	76.4 (11.8)	20
300	298.3 (7.6)	450	452.1 (15.1)	10	8.6 (5.5)	125	126.7 (12)	50	53.3 (10.1)	20
300	298.4 (9.1)	450	442.8 (24.4)	10	9.2 (5.4)	100	97.7 (11.5)	25	28.9 (9.5)	20
300	295.2 (11)	450	439.8 (23.5)	10	10.6 (5.6)	75	71.5 (9.9)	25	28.7 (8.5)	19
300	292.7 (13.1)	450	446.6 (21.4)	10	9.2 (5.3)	50	50.5 (8.1)	25	27.4 (7.2)	18
Group 8										
300	295.9 (9.9)	400	395.4 (25.1)	10	9.4 (5.2)	200	186.1 (32.8)	50	62.1 (31.5)	14
300	294.7 (10.7)	400	400.6 (23.3)	10	10.1 (5.2)	150	145.9 (25.6)	50	58.3 (24.3)	19
300	300.4 (11)	400	401.3 (20.7)	10	11.1 (5.7)	150	151.3 (29.1)	75	73.7 (28.1)	19
300	296 (12.2)	400	398.7 (23.2)	10	10.5 (5.5)	125	120.4 (25)	50	57.2 (24.1)	20
300	293.3 (12.9)	400	400.7 (31.3)	10	9.9 (5.2)	100	94.8 (19.5)	25	34.3 (18.4)	9
300	285.8 (15.3)	400	399.9 (28.4)	10	7.7 (4.6)	75	65.8 (14.8)	25	33.8 (14)	9
300	307.7 (25)	400	420.0 (22.1)	10	9.2 (5)	50	34.0 (13.3)	25	39.9 (13.3)	7

Table 28. Simulation results from the EM algorithm for *group 9* to *group 11*

Peak 1		Peak 2		Peak 3		Background		Peak 1		Peak 2		Peak 3		No. of
μ_1	$\widehat{\mu_1}$ (sd)	μ_2	$\widehat{\mu_2}$ (sd)	μ_3	$\widehat{\mu_3}$ (sd)	ν_0	$\widehat{\nu_0}$ (sd)	ν_1	$\widehat{\nu_1}$ (sd)	ν_2	$\widehat{\nu_2}$ (sd)	ν_3	$\widehat{\nu_3}$ (sd)	successes (/20)
Group 9														
300	299.1 (6.1)	500	499.9 (8.3)	700	700 (6.2)	5	5.2 (5)	150	149.6 (10.9)	150	149.3 (11.8)	150	148.7 (10.9)	20
300	300.6 (6.9)	500	497.3 (9.2)	700	697.6 (6.6)	5	4.6 (4.8)	125	122.8 (10.1)	125	125.6 (11)	125	129 (10.1)	20
300	300.7 (7.5)	500	499.2 (10.7)	700	697.8 (7.8)	5	7.9 (6.1)	100	103 (9.2)	100	97.4 (9.8)	100	96.9 (9)	20
300	299.4 (8.8)	500	500.3 (12)	700	700 (9.2)	5	5.3 (4.9)	75	74.2 (7.8)	75	75 (8.5)	75	71.4 (7.7)	20
300	303.3 (10.8)	500	502.9 (15.5)	700	699.9 (11.2)	5	5.5 (4.9)	50	51.6 (6.5)	50	48.3 (7)	50	49.5 (6.4)	20
300	300.2 (16.2)	500	497.6 (22.9)	700	692.7 (16.4)	5	5.5 (4.6)	25	25.2 (4.7)	25	24.2 (5.1)	25	25.8 (4.8)	20
Group 10														
300	297.8 8	450	449.8 (10.7)	650	649.8 (6.2)	5	6.3 (5.7)	150	145 (15.3)	150	149.3 (15.1)	150	153.3 (11.3)	20
300	301.1 (8.5)	450	453.8 (11.9)	650	651.5 (7.1)	5	5.4 (4.9)	125	127.1 (14.1)	125	122.9 (13.7)	125	119.2 (10.1)	20
300	302.1 (9.7)	450	453.2 (12.9)	650	652.9 (7.9)	5	5 (4.8)	100	101.4 (13)	100	103.8 (12.7)	100	96.2 (9.1)	20
300	298.2 (11.5)	450	452.6 (14.7)	650	651 (9.2)	5	4.7 (4.5)	75	72.5 (10.8)	75	78.3 (10.8)	75	73 (8)	20
300	300.8 (14.3)	450	451.5 (19.8)	650	649.2 (11.3)	5	5.2 (4.7)	50	49.4 (9.3)	50	48.9 (9.1)	50	49.8 (6.6)	20
300	298 (18.1)	450	460.1 (25.0)	650	657.9 (17.8)	5	4.3 (4.0)	25	28.3 (6.4)	25	27.3 (6.4)	25	22.5 (4.6)	14
Group 11														
300	286 (15.6)	400	396.5 (19.1)	500	517 (13.4)	5	3 (3.6)	150	120.4 (37.2)	150	221.5 (32.5)	150	127.4 (30.7)	6
300	284.8 (16)	400	399.5 (21.5)	500	517 (16.5)	5	5.1 (4.6)	125	105.4 (31.5)	125	173.7 (28.7)	125	96.7 (28.6)	6
300	281.3 (19.1)	400	399.3 (20.1)	500	520.3 (18.4)	5	4.5 (4.6)	100	70.1 (24.1)	100	148.3 (22.6)	100	72 (23.3)	3
300	277.0 (18.8)	400	396.0 (22.5)	500	515.0 (22.2)	5	2.3 (3.5)	75	63.3 (21.4)	75	117.8 (20)	75	51.5 (20.5)	1
300	282.0 (16.5)	400	425.0 (27.8)	500	530.0 (27.8)	5	0.02 (0.5)	50	46.6 (11.5)	50	72.2 (18.4)	50	37.5 (21.4)	1

Table 29. Simulation results from the EM algorithm for *group 12*

Peak 1		Peak 2		Peak 3		Background		Peak 1		Peak 2		Peak 3		No. of
μ_1	$\widehat{\mu_1}$ (sd)	μ_2	$\widehat{\mu_2}$ (sd)	μ_3	$\widehat{\mu_3}$ (sd)	ν_0	$\widehat{\nu_0}$ (sd)	ν_1	$\widehat{\nu_1}$ (sd)	ν_2	$\widehat{\nu_2}$ (sd)	ν_3	$\widehat{\nu_3}$ (sd)	successes (/20)
Group	*12*													
300	300.6 (7.5)	450	461.4 (13.2)	650	605.9 (9.2)	5	4.5 (4.1)	150	154.3 (15.1)	150	157.9 (17.6)	150	134.3 (17.3)	20
300	304.4 (8.2)	450	463.9 (15.6)	650	604.7 (10.6)	5	6.5 (4.9)	125	132.8 (14.3)	125	127.2 (16.7)	125	112.1 (16.6)	19
300	300.8 (9.0)	450	463.2 (16.5)	650	608.5 (11.4)	5	4.7 (4.2)	100	107 (12.3)	100	102 (14.2)	100	89.8 (14.0)	20
300	298.2 (11.1)	450	458.7 (18.1)	650	609.7 (12.9)	5	4.9 (4.3)	75	75.3 (10.8)	75	80.2 (12.2)	75	66.6 (11.3)	20
300	301.1 (14.4)	450	458.1 (23.1)	650	603.5 (17.0)	5	0.02 (5.0)	50	50.3 (9.3)	50	55.4 (10.6)	50	45.3 (10.3)	15
300	293.2 (24.0)	450	447.0 (26.1)	650	616.0 (21.7)	5	4.2 (3.4)	25	22.4 (7.1)	25	33 (7.7)	25	21.4 (5.6)	6

The results from the simulation study also reveal the limitations of the EM algorithm based estimation scheme. It is capable of estimating the locations and intensities accurately in the presence of two binding events when they are apart by at least 100 bp. The accuracy and the the ability to detect the correct number of binding events decrease as the intensities of the binding events decrease. For the case of three binding events, estimation accuracy of the scheme is further reduced. Therefore, for more accurate estimation, the binding events need to be separated by at least 100 bp.

4.6 RESULTS FROM THE STAT1 AND ZNF143 CHIP-SEQ DATA

The EM algorithm described in the previous section was applied to the ChIP-seq datasets of STAT1 and ZNF143 transcription factors. As described in section 3.6, rather than applying the EM algorithm to all regions, it was applied to a selected set of regions. These selected regions are same as those analyzed using the RJMCMC scheme. The number of binding sites detected by the EM algorithm and the single binding site model with a motif site within 50 bp, 100 bp, 200 bp, and 250 bp are given in Table 30. Compared to the single binding site model, the multiple binding sites model with the EM algorithm detects more binding sites that have a motif site

in close proximity.

Table 30. Number of binding sites with motif site in proximity for STAT1 data using the EM algorithm

Distance to the motif site	Number of sites	
	Single binding site model	RJMCMC scheme
50	292	305
100	301	320
200	305	326
250	308	327

We also present a few examples of the tag distribution of the regions where the EM algorithm detected multiple binding sites. These are given in Figures 20-23. In addition to detecting two binding sites, as in the majority of the regions, the EM algorithm was also able to detect three binding events that have a motif site within 200 bp (see Figures 22 and 23). In each of these examples, it can be clearly observed that the locations of the binding sites as well as the intensities are comparable with the tag distribution in the regions.

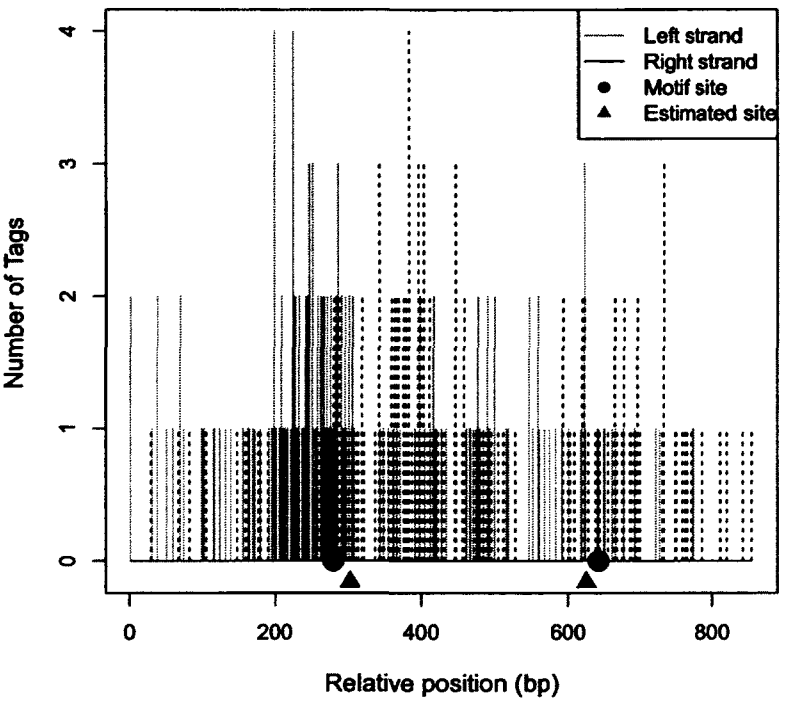


Figure 20. Estimated sites from the EM algorithm and motif sites for STAT1 data for Example 1.

Table 31. Location of the motif site and estimates of the binding sites for STAT1 data using the EM Algorithm for Example 1

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
12	88308402	280	303 (6.2)	103.1 (8.1)
		642	325 (11.3)	42.1 (5.8)

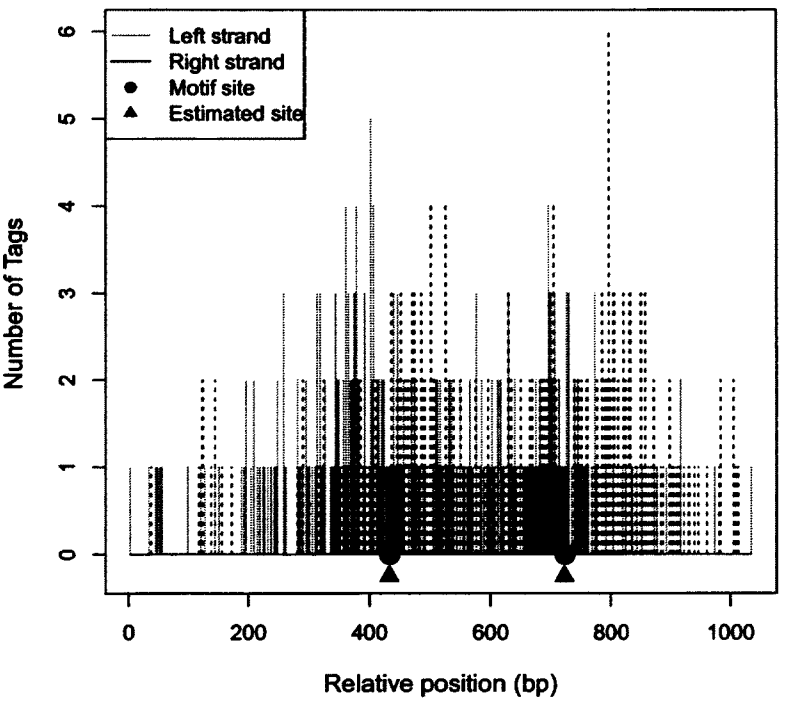


Figure 21. Estimated sites from the EM algorithm and motif sites for STAT1 data for Example 2.

Table 32. Location of the motif site and estimates of the binding sites for STAT1 data using the EM Algorithm for Example 2

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
18	1906944	432	431 (5.3)	152.7 (9.9)
		722	721 (5.4)	152.2 (9.9)

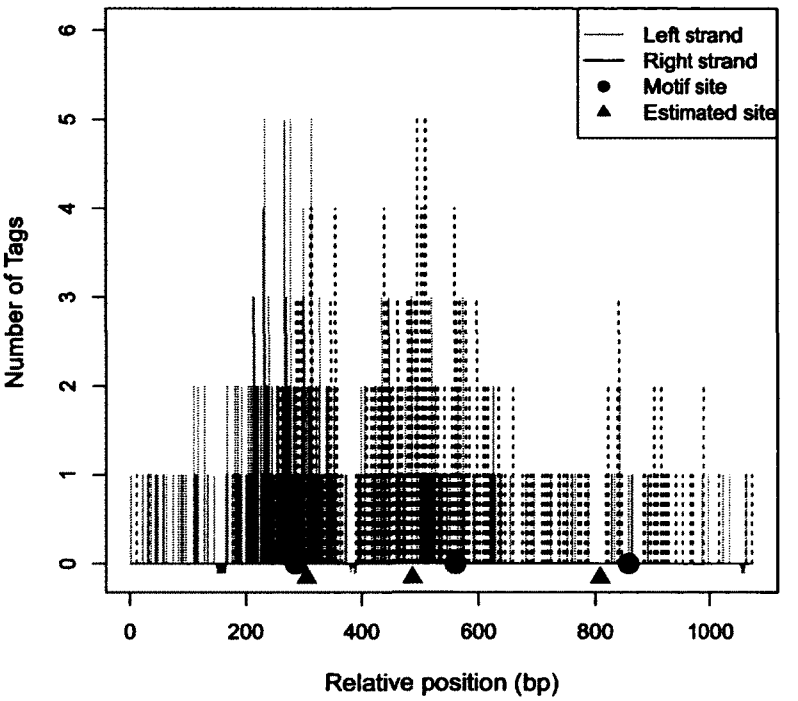


Figure 22. Estimated sites from the EM algorithm and motif sites for STAT1 data for Example 3.

Table 33. Location of the motif site and estimates of the binding sites for STAT1 data using the EM Algorithm for Example 3

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
19	39859520	286	305 (6.5)	171.6 (12.5)
		560	487 (7.7)	125.5 (11.3)
		857	808 (17.6)	22.4 (4.6)

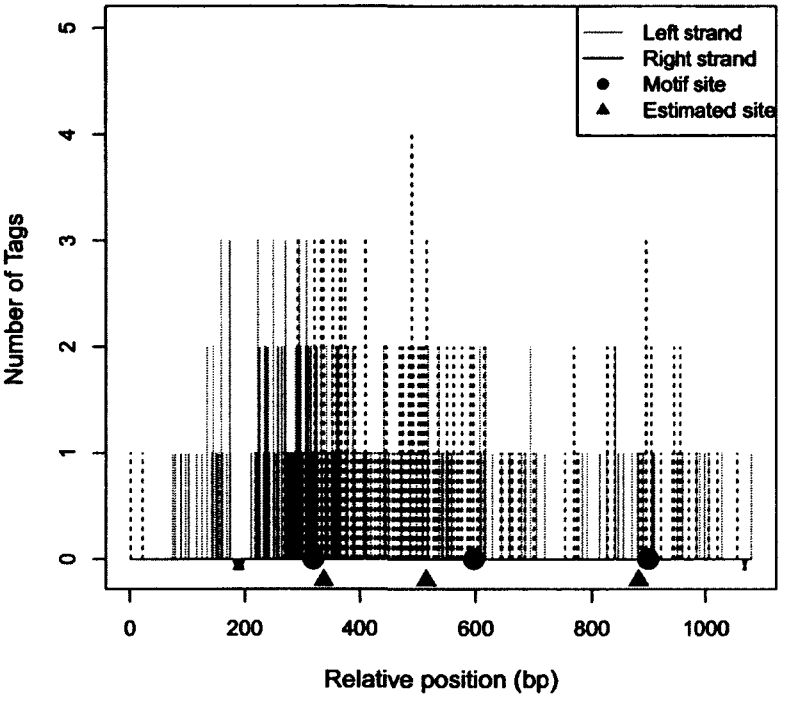


Figure 23. Estimated sites from the EM algorithm and motif sites for STAT1 data for Example 4.

Table 34. Location of the motif site and estimates of the binding sites for STAT1 data using the EM Algorithm for Example 4

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
19	39916036	319	337 (6.7)	132.3 (10.3)
		597	514 (20.5)	30.9 (7.2)
		896	880 (17.2)	21.2 (4.4)

Similarly to the analysis of STAT1 ChIP-seq data, the EM algorithm was applied to the ZNF143 ChIP-seq data. Here also, we considered the number of binding sites detected with a motif site in close proximity. The number of binding events detected by the multiple binding sites model using the EM algorithm, exceeds the number of binding events detected by the single binding event model.

Table 35. Number of binding sites with motif site in proximity for ZNF143 data using the EM algorithm

Distance to the motif site	Number of sites	
	Single binding site model	RJMCMC scheme
50	369	422
100	397	444
200	412	450
250	417	456

The tag distribution for the selected regions given in Figures 24-27 illustrates the precision of the model and the effectiveness of the EM algorithm in detecting multiple binding sites in a short region of the genome.

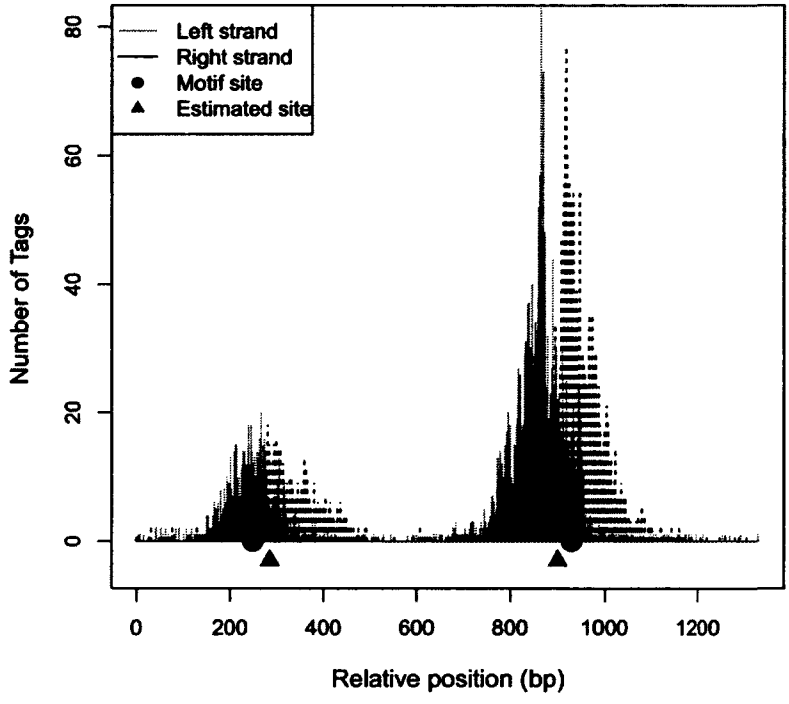


Figure 24. Estimated sites from the EM algorithm and Motif sites for ZNF143 data for Example 1.

Table 36. Location of the motif site and estimates of the binding sites for ZNF143 data using the EM Algorithm for Example 1

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
17	111877276	250	287 (1.6)	1232.9 (25.5)
		931	901 (0.9)	3947.8 (44.8)

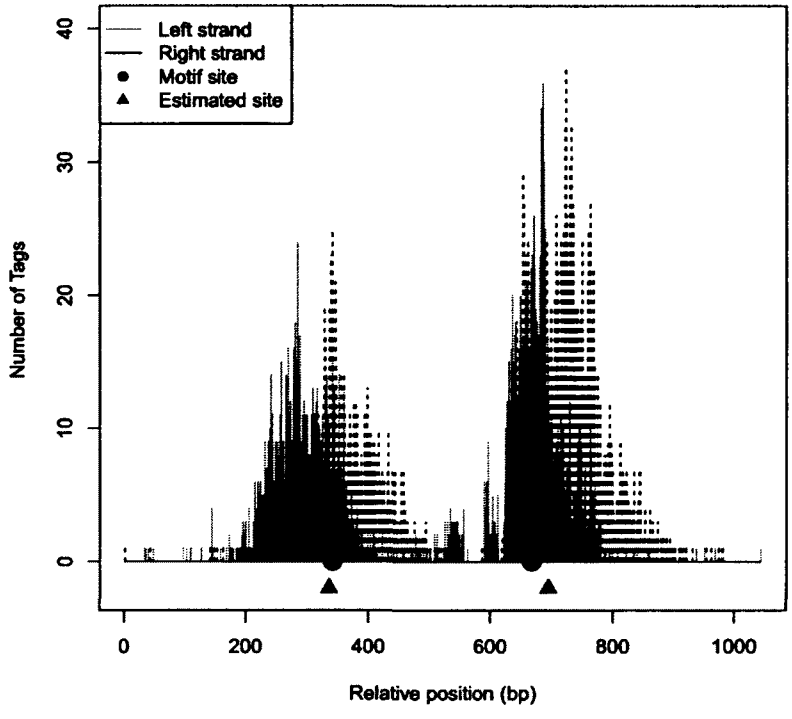


Figure 25. Estimated sites from the EM algorithm and Motif sites for ZNF143 data for Example 2.

Table 37. Location of the motif site and estimates of the binding sites for ZNF143 data using the EM Algorithm for Example 2

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
11	72986597	341	335 (1.6)	1295.1 (26.4)
		666	693 (1.2)	2182.2 (33.8)

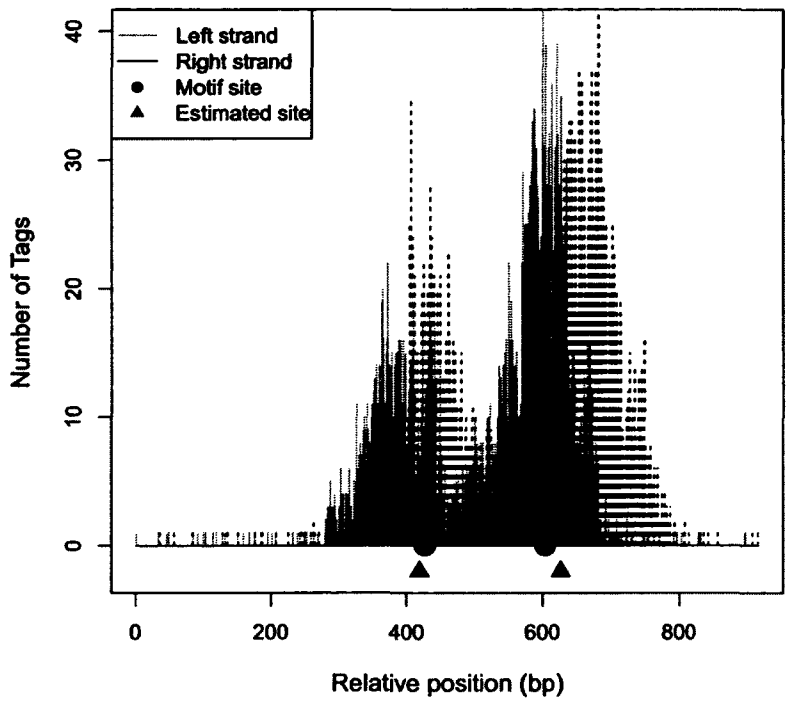


Figure 26. Estimated sites from the EM algorithm and Motif sites for ZNF143 data for Example 3.

Table 38. Location of the motif site and estimates of the binding sites for ZNF143 data using the EM Algorithm for Example 3

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
16	517026	427	419 (2.0)	1472.2 (34.8)
		604	627 (1.2)	3086.7 (45.0)

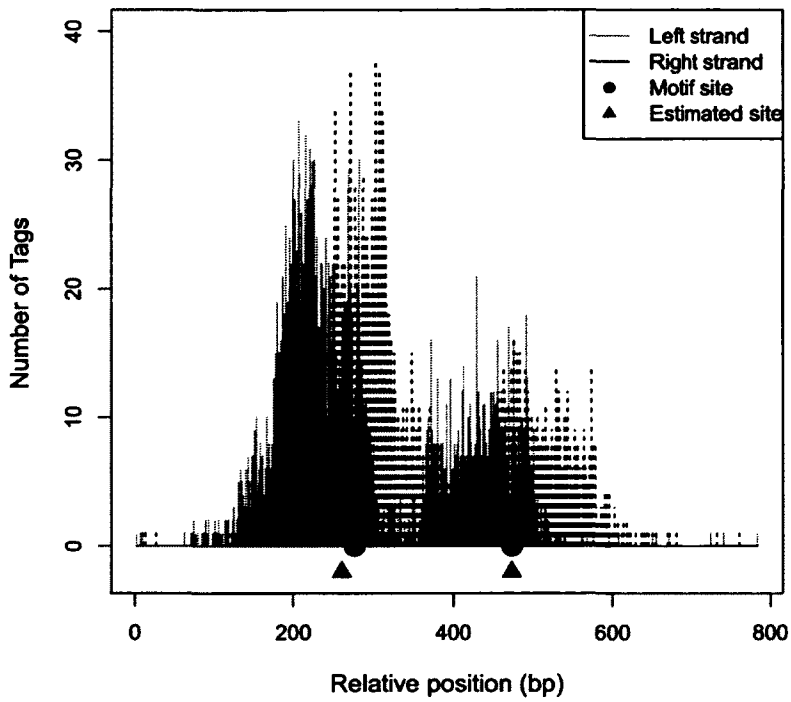


Figure 27. Estimated sites from the EM algorithm and Motif sites for ZNF143 data for Example 4.

Table 39. Location of the motif site and estimates of the binding sites for ZNF143 data using the EM Algorithm for Example 4

Chrom.	Region start position	Motif location	Binding site (sd)	Intensity (sd)
19	19292070	276	260 (1.3)	2533.2 (40.1)
		473	473 (2.3)	1053.2 (29.2)

4.7 COMPARISON OF THE MULTIPLE BINDING SITES MODEL WITH EXISTING PEAK CALLING METHODS

In the previous chapter as well as in this chapter, the multiple binding sites model using the RJMCMC scheme and the EM algorithm were evaluated using simulated datasets as well as the real ChIP-seq data. From the simulation datasets, it was observed that both implementation methods performs successfully when the binding events are separated by at least 100 bp or have higher intensity. Results from STAT1 and ZNF143 ChIP-seq data were mainly evaluated by counting the number of estimated binding sites that have a motif site within different cutoff distances. Table 40 summarizes these counts from both the RJMCMC scheme and the EM algorithm for the two ChIP-seq data.

Table 40. Number of binding sites with motif sites

Distance to	STAT 1		ZNF 143	
the nearest	RJMCMC	EM	RJMCMC	EM
motif site	scheme	algorithm	scheme	algorithm
50	309	305	448	422
100	326	320	464	444
200	335	326	472	450
250	335	327	475	456

For STAT1 data, the number of binding sites detected (that has a motif site in close proximity) by the RJMCMC scheme and the EM algorithm differ by a few binding sites. However, for ZNF143 data the RJMCMC scheme detects at least twenty additional binding events than the EM approach. The RJMCMC scheme has the flexibility to estimate the number of components and other parameters simultaneously. However, the RJMCMC scheme has following challenges:

1. It is computationally more expensive.
2. It requires an efficient one-to-one mapping function in updating the number of components.
3. It requires judicious selection of priors.

Due to the intensive computations, the RJMCMC scheme presented in chapter 3 was implemented in C++ for fast execution.

The main disadvantage of the EM algorithm is its inability to estimate the number of components. This requires running the EM algorithm on several models with different numbers of components and selecting the best model based on a model selection criterion. However, it is simpler to implement and computationally less intensive than the RJMCMC. Moreover, the performance of the EM method is comparable to the RJMCMC scheme.

We also compare the results from the multiple binding sites model obtained by the RJMCMC scheme and the EM algorithm with several existing peak-calling programs. Currently there are at least 60 peak-calling algorithms and more are been introduced every year. The selected methods are among the better performers in recent evaluation studies (Laajala et al. 2009, Wilbanks and Facciotti 2010). A brief description of these selected peak-calling algorithms was provided in section 1.4. Genome wide data of STAT1 and ZNF143 ChIP-seq data were analyzed using these programs. Whenever required, default values of the parameters for the models were used and the binding site estimates were collected.

The multiple binding sites model was applied only to a selected set of regions of the genome, where we expect to have a higher probability of observing multiple binding sites. Therefore, we selected the estimated binding sites given by the programs that fall in the selected regions. These selected binding sites were then compared with the motif sites. Summary of the number of binding sites that have motif sites within short distance are given in Tables 41 and 42 for STAT1 and ZNF143 ChIP-seq data, respectively.

Compared to the multiple binding sites model, the other programs detect very small number of binding sites that have a motif site in close proximity. This is mainly due to the fact that these algorithms lack the ability to detect peaks that are separated by small distances. Another reason is that these programs also screen out false positive binding sites based on false discovery rates or p-values. Many of the less prominent peaks were screened out in this process, decreasing the total number of binding sites that fall in the selected regions.

Table 41. Comparison of the peak calling methods using STAT1 data

Peak calling	Total binding	No. of binding sites with motif site within			
method	sites	50 bp	100 bp	200 bp	250 bp
RJMCMC	1514	309	326	335	335
EM	1323	305	320	326	327
MACS	67	22	25	25	25
sppMTC	88	25	26	27	27
sppWTD	87	25	26	27	27
QuEST	66	21	25	25	25
SISSRS	199	28	33	35	35
CisGenome	68	22	25	25	25
Hpeak	68	23	25	25	25

Table 42. Comparison of the peak calling methods using ZNF143 data

Peak calling	Total binding	No. of binding sites with motif site within			
method	sites	50 bp	100 bp	200 bp	250 bp
RJMCMC	2590	448	464	472	475
EM	1656	422	444	450	456
MACS	69	23	26	27	29
sppMTC	75	24	27	28	29
sppWTD	77	22	27	28	29
QuEST	68	24	26	27	28
SISSRS	164	25	30	35	39
CisGenome	7	0	0	1	1
Hpeak	7	0	0	0	0

Among the existing programs, for both transcription factors, SISSRs found the most number of binding sites with a motif site in close proximity. For ZNF143 ChIP-seq data, peak calling programs CisGenome and HPeak did not detect any binding sites with a motif site (within 100 bp) that fall in the selected regions. When considering the total count of binding sites predicted by the multiple binding sites model with either the RJMCMC or the EM algorithm, only 18%-22% of them are validated by the motif sites. Usually, for any transcription factor there are several motifs that is known to be associated with the binding sites. In the evaluation of the predicted binding sites we only considered the most dominant motif. Instead of directly binding to the DNA, in some instances a transcription factor may interact with other DNA bound proteins. In the ChIP-seq process, these indirect binding sites can also be precipitated and sequenced. All of these can contribute to the large number of seemingly false positive sites from the multiple binding sites. Overall, the multiple binding site model is successful in detecting a larger number of binding sites that are not reported by the other methods.

CHAPTER 5

DISCUSSION

In this thesis we introduced a statistical model to identify multiple binding sites of a transcription factor within a short region of the genome using ChIP-seq data. In our model, we propose that the number of tags y_{ij}^L at the *mappable* location is the sum of unobserved tag counts z_{ijh}^L, $h = 1, \ldots, k$, belonging to the k number of binding events. Therefore, the mapped sequence reads from ChIP-seq experiments were modeled as the sum of observations from unknown number of Poisson distributions. The rate parameters of these Poisson distributions are considered as a function of the underlying distribution of the tags that depends on the location of the binding site and an intensity parameter. The underlying distribution of the tags takes into account some features of the ChIP-seq data such as *mappability* and strand specific information. The background noise that is common in ChIP-seq data is modeled as one of the components following the Poisson distribution whose underlying distribution, is considered as uniform for a given region.

One of the main challenges in estimating the parameters of the proposed model arise from the fact that the number of components itself is unknown and needs to be estimated. Therefore, the estimation of the parameters were conducted using two different approaches: a Bayesian method and the EM algorithm.

In Bayesian paradigm, parameters of the model are considered as variables with prior distributions. This provides direct capability for estimating the number of components as well as other parameters, simultaneously. Sampling of the posterior distribution of the Bayes model was carried out by using the reversible jump Markov chain Monte Carlo (RJMCMC) method that is capable of handling the change of dimension of the parameter space. The simulation study on the RJMCMC scheme allowed us to investigate several alternatives for the RJMCMC proposals and different priors for the intensity parameters. The RJMCMC scheme with the exponential prior with mean 25 for the intensity parameter and one-to-one functions given in (27) and (28) were observed to perform the best in terms of estimating the correct number of components and location parameters. The results of the simulation study

also indicate that the the estimation process is limited to detecting binding sites separated by at least 100 bp.

In the EM algorithm approach, to determine the number of components for each region, we fitted several models with different numbers of components and selected the best model based on a model selection criteria. Here we considered the AIC and BIC methods, where when applied to simulated data, the AIC selected the correct model more frequently than the BIC method.

When the results from the multiple binding site model on real ChIP-seq data for transcription factors STAT1 and ZNF143 were compared with those from the single binding event model, it was observed that the multiple binding model successfully detected two or more binding sites within a short region. These binding sites were confirmed by the presence of motif sites in close proximity. Comparing the number of binding sites detected for the two ChIP-seq data, we observed that more sites were identified for the ZNF143 data than for the STAT1 data due to the large number of sequence reads present in the ZNF143 ChIP-seq data. This also indicates that the RJMCMC scheme may be capable of detecting a larger number of peaks when more sequence reads are available.

When considering results from both the simulation data and real ChIP-seq data, performances of the RJMCMC scheme and the EM algorithm are comparable. However, the RJMCMC method is more computationally intensive and time consuming than the EM algorithm. Therefore, for a genome wide analysis, the EM algorithm method may be more preferable.

The results from the multiple binding sites compared to those from existing peak calling methods revealed that the multiple binding sites model is successful in detecting significantly higher number of binding sites that are verified by the presence of the motif sites at short distance. At the same time, a large number of predicted binding sites were not validated by motif sites. By introducing other motifs that are known to be associated with the transcription factor, we may decrease the false positive sites and improve the results from other existing peak-calling programs.

BIBLIOGRAPHY

Babu, M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004) "Structure and evolution of transcriptional regulatory networks," *Current Opinion in Structural Biology*, 24, 283-291.

Brooks, S. P. (2001), "On Bayesian analyses and finite mixtures for proportions," *Statistics and Computing*, 11, 179-190.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1970), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1-38.

Fields, S. (2007), "Site-Seeing by Sequencing," *Science*, 316, 1441-1442.

Gelman, A., Carlin, J. B, Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Boca Raton: Chapman & Hall/CRC.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711-732.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97-109.

Ho, J. WK., Bishop, E., Kharchenko, P., Nègre, N., White, K. P., and Park, P. J. (2011), "ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis," *BMC Genomic Research*, 12:134.

Izumi, H., Wakasugi, T., Shimajiri, S., Tanimoto, A., Sasaguri, Y., Kashiwagi, E., Yasuniwa, Y., Akiyama, M., Han, B., Wu, Y., Uchiumi, T., Arao, T., Nishio, K., Yamazaki, R. and Kohno, K. (2010), "Role of ZNF143 in tumor growth through transcriptional regulation of DNA replication and cell-cycle-associated genes," *Cancer Science*, 101, 25382545.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wald, B. (2007), "Genome-Wide Mapping of in Vivo Protein-DNA Interaction," *Science*, 316, 1497-1502.

Jothi, R., Cuddapath, S., Barski, A., Cui, K., and Zhao, K. (2008), "Genome-wide identification of in-vivo protein-DNA binding sites from ChIP-seq data," *Nucleic Acids Research*, 36, 5221-5231.

Ji, H., Jian, H., Johnson, D. S., Myers, M., and Wong, W. H. (2008), "An integrated system CisGenome for analyzing ChIP-chip and ChIP-seq data," *Nature Biotechnology*, 26, 1293-1300.

Kharchenko, P. V., Tolstorukov, M. Y., and Park P. J. (2008), "Design and analysis of ChIP-seq experiments for DNA-binding proteins," *Nature Biotechnology*, 26, 1351-1359.

Kim, N. -K., Jayatillake, R., and Spouge, J. L.(2012), "A Statistical Model for Peak calling in ChIP-seq Peak Identification," Unpublished manuscript.

Kuznetsov, V. A., Singh, O., and Jenjaroenpun, P. (2010), "Statistics of Protein-DNA binding and the total number of binding sites for a transcription factor in the mammalian genome," *BMC Genomics*, 11:S12.

Laajala, T. D., Raghav, S., Tuomela, S., Lahesmaa, R., Aittokallio T., and Elo, L. L. (2009), "A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments," *BMC Genomics*, 10:618.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009), "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, 10:R25.

Li, H., Ruan, J., and Durbin, R. (2008), "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome Research*, 18, 1851-1858.

MacLachlan, G. J., Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: John Wiley & Sons Inc.

MacLachlan, G. J., Peel, D. (2000), *Finite Mixture Models*, New York: John Wiley & Sons Inc.

Park, P. J. (2009), "ChIP-seq: advantages and challenges of a maturing technology," *Nature Reveiws Genetics*, 10, 669-680.

Pepke, S., Wold, B., Mortazavi, A. (2009), "Computation for ChIP-seq and RNA-seq studies," *Nature Methods*, 6:S22-32.

Press, H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2002), *Numerical Recipes in C++: The Art of Scientific Computing*, Cambridge: Cambridge University Press.

Qin, Z. S., Yu, Z., Shen, J., Maher, C. A., Hu, M., Kalyana-Sundaram, S.,Yu, J., and Chinnayan, A. M. (2010), "HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data," *BMC Bioinformatics*, 11:369.

Reid, J. E., Evans, K. J., Dyer, N., Werinsch, L., and Ott, S. (2010), "Variable structure motifs for transcription factor binding sites," *BMC Genomics*, 11:30.

Richardson, S. and Green, P. J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society. Series B (Methodological)*, 59, 731-792.

Robert, C. P. (2007), *The Bayesian Choice: From Decision-Theoretic foundations to Computational Implementation*, New York: Springer.

Robert, C. P. and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer.

Robertson, G., Hirst, M., Bainbridge, M., Bilnkey, M., Zhao, Y., Zeng, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007), "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing," *Nature Methods*, 4, 651-657.

Semenza, G. L. (1998), *Transcription Factors and Human Disease*, New York: Oxford University Press.

Staden, R. (1989), "Methods for calculating the probabilities of finding patterns in sequences," *Computational Applied Biosciences*, 5, 89-96.

Stephen, M. (2000), "Bayesian analysis of mixture models with an unknown number of components: an alternative to reversible jump methods," *The Annals of Statistics*, 28, 40-74.

Tadesse, M. G., Naijun, S., and Vannucci M. (2005), "Bayesian Variable Selection in Clustering High-Dimensional Data," *Journal of the American Statistical Association*, 100, 602-617.

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., and Sidow, A. (2008), "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data," *Nature Methods*, 5, 829-834.

Waagepetersen, R. and Sorensen D. (2001), "A Tutorial on Reversible Jump MCMC with a View toward Applications in QTL-mapping," *International Statistical Review*, 69, 49-61.

Wanga, H., Zoub, J., Zhaoe, B., Johannsenc, E., Ashwortha, T., Wonga, H., Peard, W. S., Schuge, J., Blacklowf, S. C., Arnettf, K. E., Bernsteing, B. E., Kieffc, E., and Astera, J. C. (2011), "Genome-wide analysis reveals conserved and divergent features of Notch1/RBPJ binding in human and murine T-lymphoblastic leukemia cells," *Proceedings of the National Academy of Sciences of the United States of America*, 108, 14908-14913.

Wilbanks, E. G. and Facciotti, M. T. (2010) "Evaluation of Algorithm Performance in ChIP-seq Peak," *PLoS one*, 7, e11471.

Yilmaz, A., and Grotewold, E. (2010), "Components and Mechanisms of Regulation of Gene Expression," in *Computational Biology of Transcription Factor Binding*, ed. I. Ladunga, New York: Springer, pp. 23-32.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R., Brown, M., Li, W., and Liu, X. S. (2008), "Model-based Analysis of ChIP-Seq (MACS)," *Genome Biology*, 9:R137.

APPENDIX A

DETAILS OF SIMULATION STUDY DATA

Table 43. Simulation data in *group 1* to *group 4*

Group	Distance between the two peaks	Intensity of the first peak	Intensity of the second peak
1	200	150	150
		125	125
		100	100
		75	75
		50	50
		25	25
2	150	150	150
		125	125
		100	100
		75	75
		50	50
		25	25
3	100	150	150
		125	125
		100	100
		75	75
		50	50
		25	25
4	75	150	150
		125	125
		100	100
		75	75
		50	50
		25	25

Table 44. Simulation data in *group 5* to *group 8*

Group	Distance between the two peaks	Intensity of the first peak	Intensity of the second peak
5	200	200	50
		150	50
		150	75
		125	50
		100	25
		75	25
		50	25
6	150	50	200
		50	150
		75	150
		50	125
		25	100
		25	75
		25	50
7	100	200	50
		150	50
		150	75
		125	50
		100	25
		75	25
		50	25
8	75	200	50
		150	50
		150	75
		125	50
		100	25
		75	25
		50	25

Table 45. Simulation data in *group 9* to *group 12*

Group	Distance between first and second peaks	Distance between second and third peaks	Intensity of the first peak	Intensity of the second peak	Intensity of the third peak
9	200	200	150	150	150
			125	125	125
			100	100	100
			75	75	75
			50	50	50
			25	25	25
10	150	150	150	150	150
			125	125	125
			100	100	100
			75	75	75
			50	50	50
			25	25	25
11	100	100	150	150	150
			125	125	125
			100	100	100
			75	75	75
			50	50	50
			25	25	25
12	150	200	150	150	150
			125	125	125
			100	100	100
			75	75	75
			50	50	50
			25	25	25

VITA

Rasika Jayatillake
Department of Mathematics and Statistics
Old Dominion University
Norfolk, VA 23529

Education

- Ph.D. in Applied and Computational Mathematics (Statistics), Old Dominion University, Norfolk, VA (August 2012)
- M.S. in Statistics, Old Dominion University, Norfolk, VA (May 2010)
- B.Sc. (Statistics), University of Colombo, Sri Lanka (September 2005)

Experience

- Teaching Assistant, Old Dominion University, Norfolk, VA (since Aug 2007)
- Bio-medical Statistician, South Asian Clinical Toxicology Research Collaboration, Sri Lanka (Oct 2005 to Apr 2007)

Typeset using LaTeX.