Bayesian Integration and Modeling for Next-generation Sequencing Data Analysis

Xi Chen

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical and Computer Engineering

Jianhua J. Xuan
Yue Wang
Aloysius Beex
Michael S. Hsiao
Chang-Tien Lu

April 28th, 2016
Arlington, Virginia

Keywords: NGS Data Analysis, Transcriptional Regulatory Network, Genomic Mutation, Bayesian Modeling, Expectation-Maximization, Gibbs Sampling

© Copyright by Xi Chen 2016

Bayesian Integration and Modeling for Next-generation Sequencing Data Analysis

Xi Chen

ABSTRACT

Computational biology currently faces challenges in a 'big data' world with thousands of data samples across multiple disease types including cancer. The challenging problem is how to extract biologically meaningful information from large-scale genomic data. Next-generation Sequencing (NGS) can now produce high quality data at DNA and RNA levels. However, in cells there exist a lot of non-specific (background) signals that affect the detection accuracy of true (foreground) signals. In this dissertation work, under Bayesian framework, we aim to develop and apply approaches to learn the distribution of genomic signals in each type of NGS data for reliable identification of specific foreground signals.

We propose a novel Bayesian approach (ChIP-BIT) to reliably detect transcription factor (TF) binding sites (TFBSs) within promoter or enhancer regions by jointly analyzing the sample and input ChIP-seq data for one specific TF. Specifically, a Gaussian mixture model is used to capture both binding and background signals in the sample data; and background signals are modeled by a local Gaussian distribution that is accurately estimated from the input data. An Expectation-Maximization algorithm is used to learn the model parameters according to the distributions on binding signal intensity and binding locations. Extensive simulation studies and experimental validation both demonstrate that ChIP-BIT has a significantly improved performance on TFBS detection over conventional methods, particularly on weak binding signal detection.

To infer cis-regulatory modules (CRMs) of multiple TFs, we propose to develop a Bayesian integration approach, namely BICORN, to integrate ChIP-seq and RNA-seq data of the same tissue. Each TFBS identified from ChIP-seq data can be either a functional binding event mediating target gene transcription or a non-functional binding. The functional bindings of a set of TFs usually work together as a CRM to regulate the transcription processes of a group of genes. We develop a Gibbs sampling approach to learn the distribution of CRMs (a joint distribution of multiple TFs) based on their functional bindings and target gene expression. The robustness of BICORN has been validated on simulated regulatory network and gene expression

data with respect to different noise settings. BICORN is further applied to breast cancer MCF-7 ChIP-seq and RNA-seq data to identify CRMs functional in promoter or enhancer regions.

In tumor cells, the normal regulatory mechanism may be interrupted by genome mutations, especially those somatic mutations that uniquely occur in tumor cells. Focused on a specific type of genome mutation, structural variation (SV), we develop a novel pattern-based probabilistic approach, namely PSSV, to identify somatic SVs from whole genome sequencing (WGS) data. PSSV features a mixture model with hidden states representing different mutation patterns; PSSV can thus differentiate heterozygous and homozygous SVs in each sample, enabling the identification of those somatic SVs with a heterozygous status in the normal sample and a homozygous status in the tumor sample. Simulation studies demonstrate that PSSV outperforms existing tools. PSSV has been successfully applied to breast cancer patient WGS data for identifying somatic SVs of key factors associated with breast cancer development.

In this dissertation research, we demonstrate the advantage of the proposed distributional learning-based approaches over conventional methods for NGS data analysis. Distributional learning is a very powerful approach to gain biological insights from high quality NGS data. Successful applications of the proposed Bayesian methods to breast cancer NGS data shed light on underlying molecular mechanisms of breast cancer, enabling biologists or clinicians to identify major cancer drivers and develop new therapeutics for cancer treatment.

Bayesian Integration and Modeling for Next-generation Sequencing Data Analysis

Xi Chen

GENERAL AUDIENCE ABSTRACT

Next-generation sequencing data is currently widely used in biomedical research. How to extract biological meaningful signals from noisy 'big' data is the main challenge for today's computational biology. Transcription factor-gene regulation is a classical topic in this filed. Recently using Next-generation sequencing technology biologists can measure signals related to gene regulation with a greatly improved accuracy. In this dissertation research, to study cell type specific regulatory mechanism, under a Bayesian framework, we propose three novel distribution learning-based approaches to respectively detect transcription factor binding sites, infer cis-regulatory modules, and identify somatic mutations. We demonstrate the advantage of the proposed distributional learning-based approaches over conventional methods for Next-generation sequencing data analysis. Distributional learning is a very powerful approach to gain biological insights from high quality and large scale biomedical data. Successful applications of the proposed Bayesian methods to breast cancer Next-generation sequencing data shed light on underlying molecular mechanisms of breast cancer, enabling biologists or clinicians to identify major cancer drivers and develop new therapeutics for cancer treatment.

Acknowledgement

I would like to express my special appreciation and thanks to my advisor Dr. Jason Xuan. I would like to thank him for his time, energy, patience, ideas and funding for my Ph.D. study and allowing me to grow as a biomedical researcher. He is always encouraging me to explore novel ideas and develop unique solutions to the problems in the field of computational biology. Under his direction, I have developed several computational tools for next-generation sequencing data analysis and completed this dissertation work. I greatly appreciate his effort to maximize my personal capability. All in all, his advice on both my research and my career has been priceless.

I would also like to thank my committee members, Dr. Yue Wang, Dr. A. A. (Louis) Beex, Dr. Chang-Tien Lu and Dr. Michael Hsiao, for serving in my dissertation advisory committee. In particular, their suggestions and comments during my preliminary exam helped me reshape my ideas. Guided by their advice, I came to stay focused on those fundamental yet challenging points with a refreshed mind and finally found novel solutions to my dissertation research problems.

I am grateful to our major collaborators, Dr. Robert Clarke at Georgetown University and Dr. Tian-Li Wang at Johns Hopkins University, who are leaders in breast or ovarian cancer research. They have provided me with their best support in biological validation experiments. With their help, I have successfully published my first journal paper in Nucleic Acids Research. I also want to thank Dr. Ayesha N. Shajahan-Haq at Georgetown University. I am not a native English speaker. I still remember that in the beginning of my Ph.D. study she encouraged me to speak English more and more; in each group meeting she was patient enough to understand my questions and provide her professional explanations. She has also provided me with a lot of help on paper writing and been a co-author in my paper publications.

I would like to express my gratitude to people in CBIL, especially Dr. Guoqiang Yu for his suggestion on my career choice and Xu Shi for his help in big data analysis. I also want to thank Dr. Li Chen, Dr. Chen Wang and Dr. Jinghua Gu for their help on my initial exploration of the field of computational biology.

A special thanks to my family. Words cannot express how grateful I am to my wife, my parents-in-law and my parents. With my mother and father's support, five years ago I decided to join Virginia Tech for my Ph.D. study. My mother and father-in-law are also proud of my research and encourage me to continue my academic career in this field, although they know that this path is challenging and their daughter will sacrifice a lot. Finally I would like to express my deepest appreciation to my wife, Xiaoyi Wang. Her faith motivates me to go further in the moments when I could not find light on my research. With her love, support, encouragement and patience, I have devoted all my energy to research and completed this dissertation work. Thank You!

Table of contents

List of Figures

List of Tables

List of Abbreviations

ARD	Average Read Depth
AUC	Area Under the Curve
BICORN	Bayesian Integration of ChIP-seq and RNA-seq data for cis-regulatory module inference
ChIP-BIT	Bayesian Inference of transcription factor binding site using ChIP-seq data
ChIP-seq	Chromatin immunoprecipitation (ChIP) with massively parallel sequencing
CRM	Cis-Regulatory Module
DNA-seq	DNA sequencing
EM	Expectation-Maximization
ER-	Estrogen Receptor Negative
eRNA	enhancer RNA
FDR	False Discovery Rate
HM	Histone Markers
LASSO	Least Absolute Shrinkage and Selection Operator
mRNA	massager RNA
NGS	Next Generation Sequencing
PSSV	Patten based Somatic Structural Variation identification
RNA-seq	RNA sequencing
ROC	Receiver Operating Characteristic
SNR	Signal-to-Noise Ratio
SV	Structural Variation
TF	Transcription Factor
TFA	Transcription Factor Activity
TFBS	Transcription Factor Binding Site
TSS	Transcription Starting Site
WGS	Whole Genome DNA Sequencing

1. Introduction

1.1 Motivations

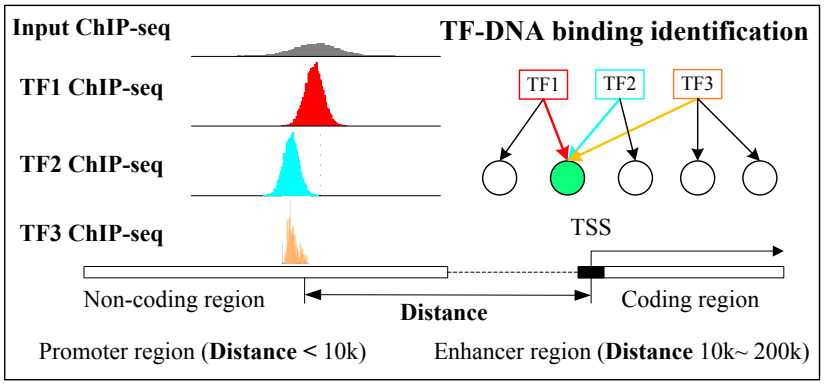
Computational biology is a cross-subject area that mainly focuses on the development and application of mathematical modeling techniques to the study of biological systems [1]. Under biological hypotheses proposed by biologists and clinicians, computational scientists aim to develop specialized algorithms to assess the complexity of biological systems. Currently, most computational efforts have been devoted to the translational cancer research [2]. Cancer is a complex disease, involving multiple and specific changes at the DNA, RNA and protein levels, which can be inherited from parents and/or induced by environmental factors [3]. There are many different types and subtypes of cancer. The main factor driving cancer development is still unclear. The translational cancer research seeks to identify and understand the causes and effects of cancer-specific variations at different levels and to translate this knowledge to the clinical application to improve cancer prevention and therapy [4]. Over the last ten years, in order to identify new drug targets, many biological hypotheses have been proposed and multiple types of omics data have been generated [5, 6]. Novel computational methods are needed for the high throughput biomedical data analysis.

Next-generation sequencing (NGS) technology has greatly improved the accuracy of biological signal measurements and sped our progress on fighting against cancer [7]. Using only a small number of deeply sequenced biological samples we are able to identify specific molecule markers of a specific type of cancer cells under a certain condition. The current NGS platform can provide different data types including exome or whole genome DNA [8, 9], coding or non-coding RNA [10, 11], protein-DNA interaction [12], protein-RNA interaction [13], DNA 3D structure [14], etc. The major objective of modelling NGS data is to differentiate foreground signals (highly active signals under a certain condition) from background signals. However, the difficulty is that background signals cannot be simply treated as random noise. Although they are much less active, they may be transmitted from some fundamental elements [12] and can form a certain distribution pattern. Therefore, using conventional methods based on the ‘foreground signal + noise’ model we may obtain a lot of false positive predictions. Alternatively, using a foreground-background mixture model can greatly improve the prediction accuracy. In this

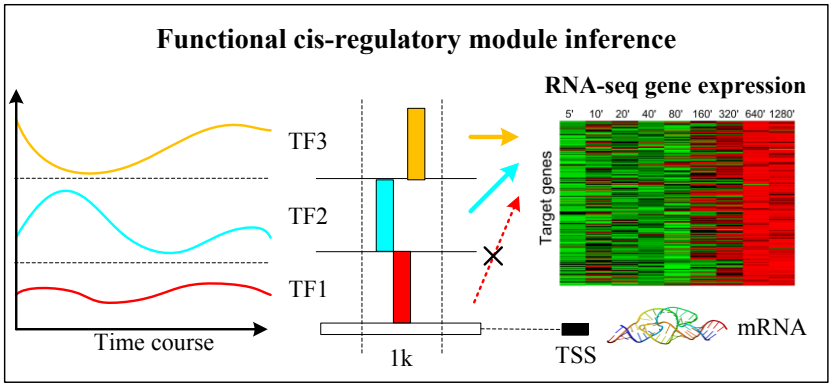
mixture model, a reasonable hypothesis of the distribution of foreground or background signals is a prerequisite for accurately extracting foreground information from each NGS data sample. Due to the diversity of biological systems, in most cases we only know the basic distribution shape of the measured signals. The distribution parameters need to be learned from the data directly. For some particular biological questions, a joint distribution of multiple factors needs to be learned. However, such a joint distribution is difficult to model using a particular probabilistic distribution. Hence, novel distributional learning algorithms need to be developed.

1.2 Background and data sources

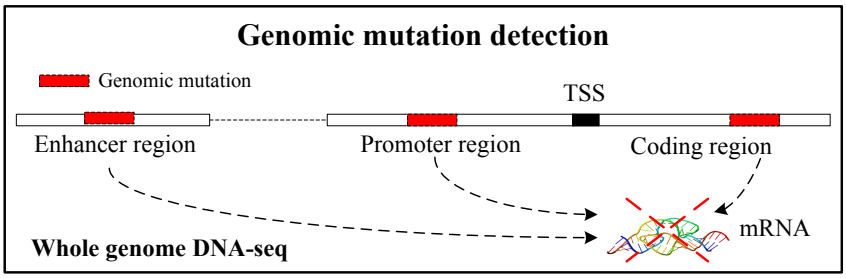
NGS data have multiple types including three major ones including ChIP-seq, RNA-seq and DNA-seq. Transcription factor (TF) is a protein that binds to specific DNA sequences, thereby controlling the rate of target gene mRNA transcription. Chromatin immunoprecipitation (ChIP) with massively parallel sequencing (seq) technology is used to investigate the mechanism of TF-DNA interactions [15]. ChIP-seq data provides a detailed TF-DNA binding signal map along the whole genome. As illustrated in Fig. 1.1(a), a region with high binding signals and low input signals can be called transcription factor binding site (TFBS). If a TFBS is located within a certain distance (e.g., 10k base pairs (bps); promoter region) from the nearest transcription starting site (TSS), we will use the proximal binding theory that the TF will directly regulate the expression of the nearest target gene [16]. If the binding location of a TFBS is quite far away from any TSS (e.g., 200k bps), we will use the distal enhancer loop theory that the TF will activate an enhancer region and then regulate the expression of a distal target gene [17]. An enhancer region may be distally located either upstream or downstream of the gene it regulates, as some enhancer regions have been found located a few hundred thousand base pairs away from TSS [18]. The enhancer region may ‘jump’ over the nearest target gene to regulate one or multiple distal genes [19]. Obviously, the key step in ChIP-seq data analysis is TFBS identification. Only after that can we further study TF-gene regulation through promoter or enhancer mechanism.



(a)



(b)



(c)

Figure 1.1 Transcriptional regulation of gene expression. (a) Physical binding signals in the ChIP-seq data can be used to identify TFBSs at promoter and enhancer regions. (b) TFs bind together as cis-regulatory modules to regulate the mRNA transcription of target genes. RNA-seq data provides gene expression measurement. (c) Whole genome DNA-seq data can be used to detect genomic mutations occurring at promoter, enhancer, coding regions. Functional mutations may interrupt the normal gene transcriptional process.

ChIP-seq data has been widely used to identify TFBSs in several major types of cancer cells [20]. General TFs usually bind at the promoter region to directly regulate the expression of the target gene. Some specific TFs bind at distal enhancer regions, recruit RNA polymerase II

and the general TFs, and then begin transcribing the genes (3). Modelling multiple TFs simultaneously at enhancer or promoter regions helps to investigate the collaborative roles of multiple TFs on target gene regulation. A cis-regulatory module (CRM) is defined is a stretch of DNA, e.g. 1k base pairs in length, with binding sites of multiple TFs. TFs in each CRM regulate a group of target genes whose expression is either significantly activated or repressed [21]. With TF-DNA binding information obtained from ChIP-seq data, researchers can measure the target gene expression and then use reverse engineering approaches to infer how TFs collaboratively work together (function as CRMs) [22]. RNA-seq is a recently developed technique for transcriptome profiling [23]. Using RNA-seq data, functional bindings or CRMs can be inferred from the initial TF-DNA binding network obtained from ChIP-seq data, as shown in Fig. 1.1(b). In this integrative analysis, high quality RNA-seq expression data plays an important role to accurately predict CRMs.

The above mentioned studies about TFBS identification and their functional effects on target gene expression are based on the hypothesis that both encoding regions (promoter or enhancer regions) and coding regions (exons of gene body) are 'normal'. However, in cancer research, genomic mutations occur frequently at places in the whole genome. If a genomic mutation happens to an encoding and coding region, TF may not be able to bind at that location and/or the gene transcription process will be affected [24], as shown in Fig. 1.1(c). The whole genome DNA-seq (WGS) provides a full map of genomic mutations including single nucleotide variation, small insertion or deletion, structural variation (SV) and copy number variation [25]. Most computational efforts are focused on SVs because they can only be detected using paired-end reads in WGS data [26, 27], while other types of mutations can also be detected using SNP array data [28]. Functional effects of SVs on gene expression or cell phenotype have been widely discussed [29]. Currently, one major task in WGS data analysis is to identify cancer specific SVs [30, 31], also called somatic SVs. Somatic mutations are likely to be critical factors determining how tumors progress and respond to treatments. Besides somatic SVs, there are also a lot of SVs which are observed at the same genomic regions in both tumor and normal samples. Such SVs are called germline SVs and are mainly inherited from parents. Germline mutations are less possible to be related with cancer development and are usually treated as background signals in

the study of somatic mutations. Then, how to accurately differentiate somatic SVs from germline SVs is another major problem we will address in this dissertation research.

1.3 Objective and statement of problem

In this dissertation work, we propose three methods to model or integrate multiple types of NGS data for a better understanding of transcriptional regulation. The major focuses of this dissertation research are summarized as follows: (1) to propose a Gaussian mixture model for TFBS identification; (2) to develop a Gibbs sampling approach to infer functional CRMs; (3) to propose a mutation pattern based Poisson mixture model to predict somatic SVs. Briefly, for each type of NGS data, we propose a unique probabilistic mixture model to denote the co-existence of foreground and background signals. Under the Bayesian framework, based on the prior distribution hypothesis made for foreground or background signals, we learn the posterior distributions of foreground and background signals from the high quality NGS data.

1.3.1 Identifying transcription factor binding sites using ChIP-seq data

To identify TFBSs of a specific TF, biologists run a sample ChIP-seq experiment to capture both foreground and background signals and another input ChIP-seq experiment to capture background signals only. Genomic regions with distinct signal pattern between the sample experiment and the input experiment should contain TFBSs or foreground bindings. The regulation strength of different TFs may be also different. Some TFs serve as major regulators and bind at encoding regions frequently with strong binding signals; while some co-factors bind at diverse binding loci with relatively weak binding signals. It is a challenging task to identify those weak but foreground bindings since they are mixed with background signals in the sample experiment. Recent research shows that functional effects of weak bindings on gene transcription can be very significant [32].

A number of TFBS detection tools were developed using both sample and input ChIP-seq data. MACS [33] and PeakSeq [34] are the two most widely used TFBS detection methods; these tools use local Poisson or Binomial statistics on ChIP-seq read counts to identify TFBSs. BCP [35] uses a Bayesian change point approach to model read coverage change along the genome and locate peak boundaries. DFilter [36] trains an optimal detection filter using read count

observations from the input ChIP-seq data and then applies the detection filter to the sample data for peak prediction. MOSAiCS [37] uses a mixture model of negative binomial distributions of read counts to identify foreground bindings. ChIP-seq read count is the most widely used binding signal format. Here, we present the distributions of read count in a pair of input and sample data sets in Fig. 1.2(a) and (b), respectively. Strong, weak and background bindings are illustrated in Fig. 1.2(d). Strong bindings can be easily predicted because they are located at the right tail of the read count distribution in Fig. 1.2(b); also 'far away' from the read count distribution of background bindings in Fig. 1.2(a). Weak bindings with medium read counts are possibly falsely classified as background because their 'distance' to distribution of background bindings is shorter than their 'distance' to those strong bindings located at the far end of the distribution in Fig. 1.2(b). Therefore, if we still use conventional TFBS detection tools (e.g. PeakSeq) to predict weak bindings by decreasing the threshold, many background bindings will be included in the final TFBS list, as shown in Fig. 1.2(f).

In this dissertation, we propose to develop a Gaussian mixture model-based approach, namely ChIP-BIT, to efficiently identify both strong and weak TFBSs at encoding regions. As shown in Fig. 1.2(c), the read count can be converted to read intensity that follows a Gaussian distribution. Using read intensity as a new measurement of binding signals, we can shrink the 'distance' between weak and strong bindings and enlarge the 'distance' between weak and background bindings. Therefore, we have a better chance to predict weak and strong bindings together as foreground bindings and differentiate them from background ones. In ChIP-BIT, as illustrated in Fig. 1.2(e), background signals in the sample experiment are assumed to follow a local Gaussian distribution that is quite close to the distribution of input signals (all background); foreground binding signals are modelled using a global Gaussian distribution that is 'far away' from the distribution of input signals. The distribution parameters of foreground and background components of the proposed Gaussian mixture model are directly learned from the sample and input ChIP-seq data using an Expectation-Maximization algorithm. Comparing TFBS prediction results of ChIP-BIT (Fig. 1.2(g)) with those of PeakSeq (Fig. 1.2(f)), the contamination of background signals is much lower.

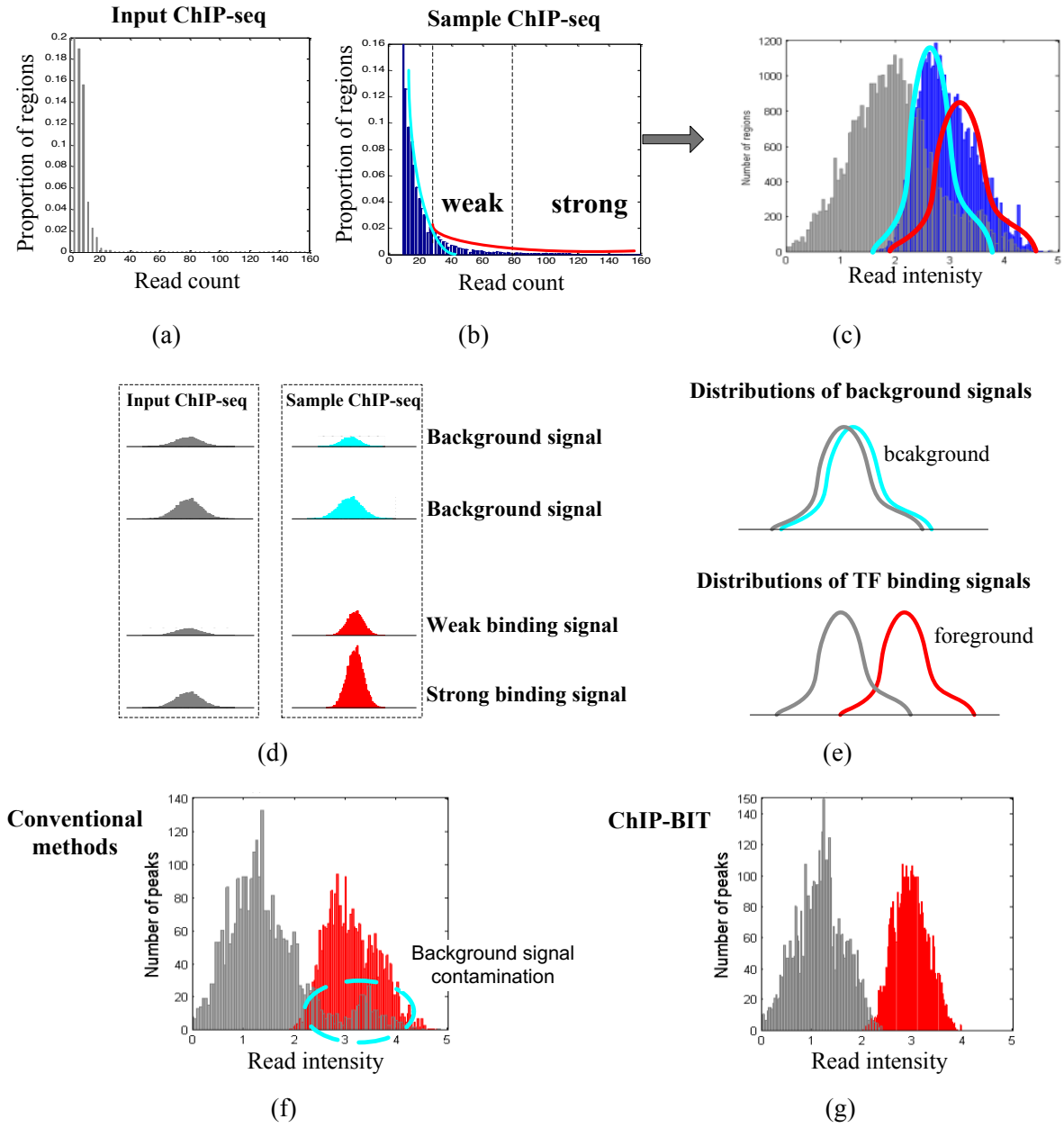


Figure 1.2 Foreground TFBS identification by modelling read intensity of sample and input ChIP-seq data using a Gaussian mixture model. (a) Read count distribution in the input ChIP-seq data; (b) read count distribution in the sample ChIP-seq data; (c) read intensity distributions in sample (blue) and input (grey) ChIP-seq data; (d) illustration of strong, weak and background bindings; (e) two Gaussian components for background and foreground bindings; (f) peak calling results using the conventional tool PeakSeq; (g) peak calling results using the new ChIP-BIT.

1.3.2 Inferring cis-regulatory modules by integrating ChIP-seq and RNA-seq data

To study the cooperation of multiple TFs on gene expression or infer functional cis-regulatory modules (CRMs), RNA-seq gene expression data should be jointly modelled with the

binding information of TFs, as shown in Fig. 1.3. Conventional studies of the regulatory mechanism were focused on promoter regions. In recent years the machinery of gene regulation from the distant enhancer region is more and more evident [38]. An enhancer region is first activated by a specific set of TFs and then interacts with general TFs at the promoter region to achieve distal gene regulation. Although several existing approaches like BNCA [39] and COGRIM [40] can also be used to integrate ChIP-seq and RNA-seq data for regulatory network inference, they model each binding event individually and can only infer functional TF bindings at promoter regions. Novel computational tools need to be developed for TF module inference at both promoter and enhancer regions. After a review of traditional integrative approaches on statistical integration of ChIP-seq and RNA-seq data [41], Bayesian integration methods would be better approaches for causal inference of genome-wide CRM inference.

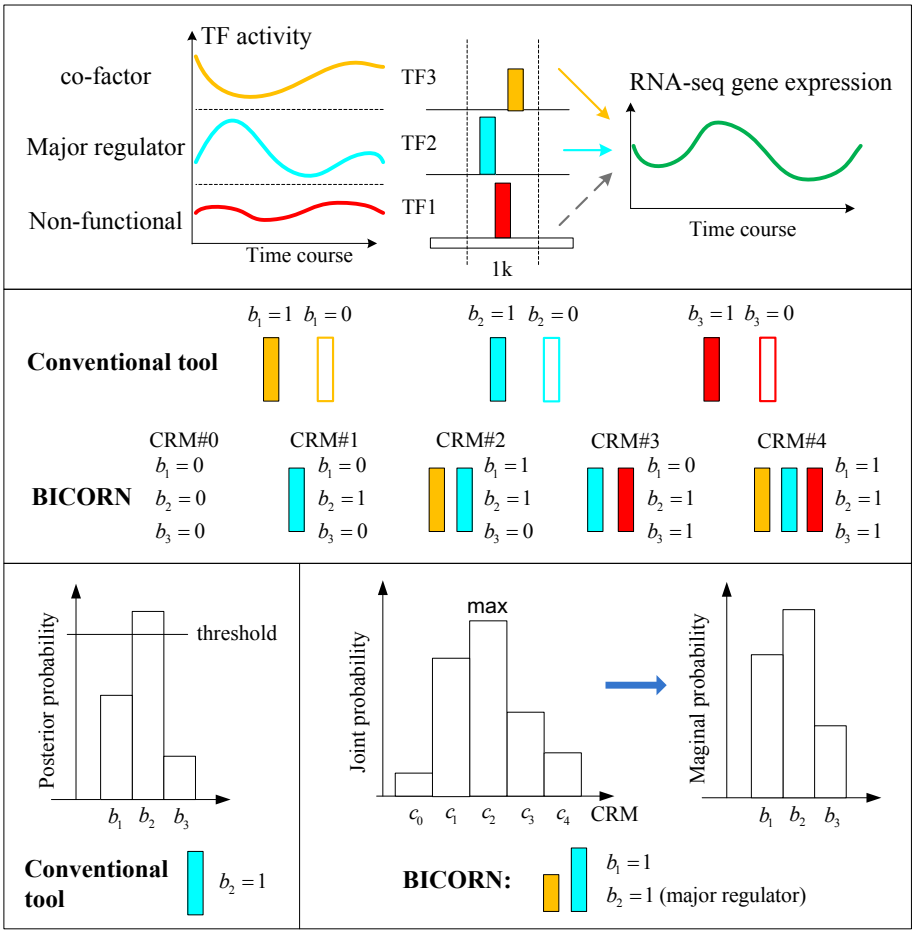


Figure 1.3 Inferring cis-regulatory modules by integrating ChIP-seq and RNA-seq data.

As shown in Fig. 1.3, conventional Bayesian integration tools like BNCA and COGRIM evaluate each individual binding event by comparing the two posterior probabilities between binding occurrence and non-occurrence. They cannot be directly used to infer CRMs because there is no variable to model the cooperation of multiple TFs. Here, we propose a Bayesian integration approach, namely BICORN, to model CRMs directly. Each CRM represents a unique combination of at least two TFs. We use a Gibbs sampling algorithm to learn the distribution CRMs (a joint distribution of multiple TFs) because the joint distribution is difficult to model using a particular distribution. For each gene, among multiple candidate CRMs, we directly evaluate which CRM is more functional than the others. As shown in Fig. 1.3, the learned distribution of CRMs shows that CRM#2 (a combination of TF1 and TF2) is more possible to regulate the target gene under investigation if compared to the other combinations. The marginal distribution of each TF can then be calculated and used to predict which TF is the major regulator.

1.3.3 Identifying somatic structural variation using WGS data

Genomic mutation analysis has been accelerated by the accumulation of DNA-seq data acquired from exome regions to the whole genome. Structural variation (SV), a major type of genomic mutation [42], has been characterized in several studies [43, 44]. Somatic SVs detection can be achieved by comparing the sequence data of a tumor sample with that of its matched normal sample [45, 46].

BreakDancer [47] and GASVPro [48] are two widely used SV detection tools. For somatic SV prediction, these tools will identify SVs in a tumor sample and its matched normal sample independently, and then retain the SVs unique in the tumor sample. As shown in Fig. 1.4, BreakDancer can predict two types of somatic SVs where the normal chromosome has to be ‘healthy’. In the real cellular system, there is another very interesting type of somatic SVs with one copy of chromosome mutated (heterozygous mutation) in the normal sample and both copies mutated (homozygous mutation) in the tumor sample. However, this type of somatic SVs will be missed if BreakDancer is used. GASVPro modelled two copies of chromosome in a probabilistic framework so that each SV can be either predicted as a heterozygous or a homozygous mutation. Then, three types of somatic events can be predicted. However, in cancer research, tumor sample

impurity cannot be overlooked. Although the contamination of normal cells can be greatly reduced by jointly analyzing sequencing data of tumor and normal samples [49], tumor sequencing data is still noisy. Using exiting SV detection tools, SV detection accuracy in the tumor sample is much lower than that in the normal sample. In addition, for the normal sample, usually read coverage is not as high as that of tumor sample, which also degrades SV detection accuracy in normal sample as well.

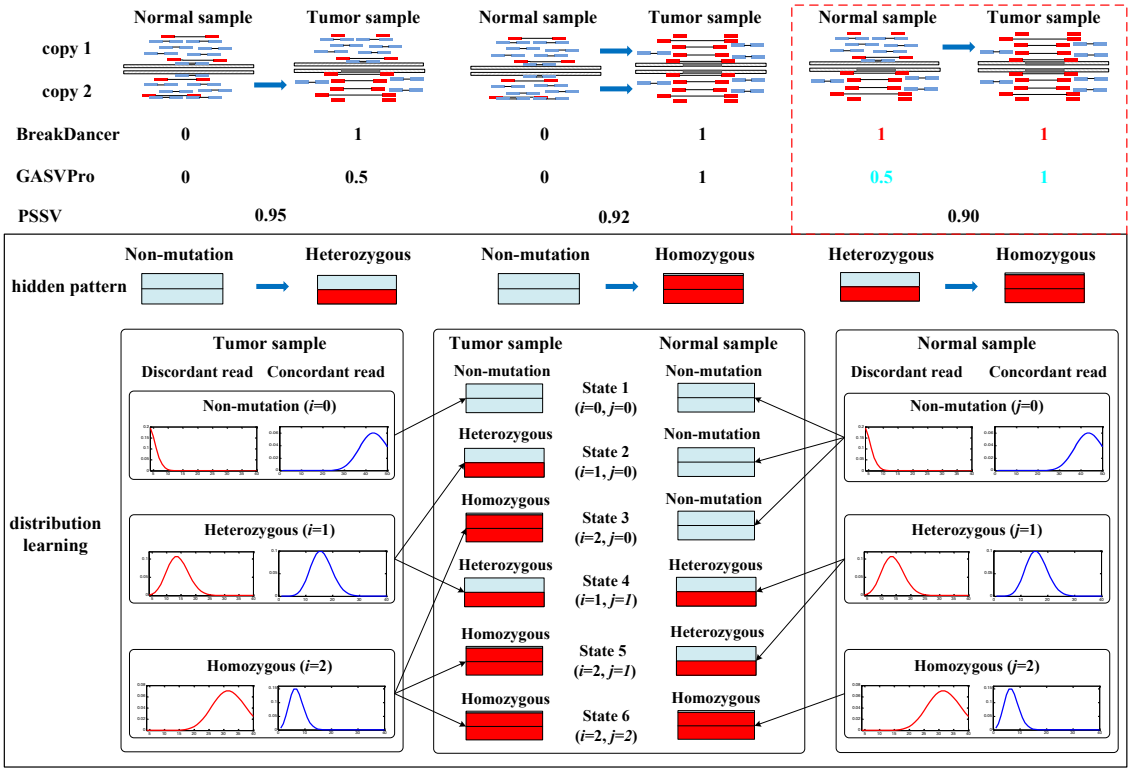


Figure 1.4 Somatic SV detection using conventional approaches and the proposed PSSV.

We propose a pattern-based probabilistic approach (PSSV) for somatic SV prediction by jointly analyzing WGS data from a paired tumor and normal samples. A statistical approach is used to find SVs with a significant mutation pattern between a tumor sample and its matched normal sample. In detail, as shown in Fig. 1.4, we define the ideal mutation pattern of a SV in a pair of pure tumor and normal samples as one of six states, including three 'somatic', two 'germline', and one 'none' (as a special case of 'germline'). In each state, read counts from the tumor and normal samples are assumed to follow specific Poisson distributions. The real sequencing data is noisy so the mutation pattern of a SV is not as clear as the ideal mutation

status defined in Fig. 1.4. Hence, each SV is modeled as a mixture of six hidden states under a probabilistic framework. Through an Expectation-Maximization algorithm, we iteratively estimate the prior probability for each state as well as its Poisson distribution parameters. Finally, for each SV, we can predict a posterior probability for each state to denote at what confidence level this SV can be represented by that state. We use the most reliable state to represent the major mutation status and further classify this SV to either the somatic or germline category.

1.4 Summary of contributions

In this dissertation, we focus on developing computational methods to differentiate foreground (biologically meaningful) signals from background ones in each type of NGS data. We summarize the major contributions as follows:

(1) We have developed a novel Bayesian approach (ChIP-BIT) to reliably detect TFBSs at promoter or encoding regions by jointly modeling read intensities from the sample and input ChIP-seq data. ChIP-BIT uses a Gaussian mixture model (consisting of global and local Gaussian components) to denote foreground binding and background signals in the sample data. A unique feature is that the Gaussian component modeling background signals is specially designed as a local Gaussian distribution that can be estimated accurately from the input data. An Expectation-Maximization algorithm is used to efficiently learn the Gaussian distribution parameters. ChIP-BIT aims to detect weak binding signals together with strong ones such that a more complete regulatory network can be generated.

(2) We have developed a Bayesian integration approach (BICORN) to directly identify CRMs by integrating ChIP-seq and RNA-seq data. The novelty of BICORN lies in the direct modelling of TF combinations or TF modules. The distribution of CRMs (also a joint distribution of multiple TFs) is directly learned from ChIP-seq and RNA-seq data. Each target gene expression is modeled as a log-linear combination of the activities of TFs in a CRM. In each CRM, there is a unique combination of TFs. Using a Gibbs sampler, the proposed algorithm iterates between the estimations of hidden TF activities and the learning of the posterior distribution of CRMs. After generating enough samples, for each target gene, the most reliable CRM is selected. In addition, major regulators, co-factors and background TFs among all candidate TFs will be predicted as well.

(3) We have developed a novel pattern-based probabilistic approach (PSSV) to identify somatic SVs from WGS data of a pair of tumor and normal samples. Somatic or germline SVs are jointly modelled under a probabilistic framework. Ideal mutation patterns of somatic or germline SVs are used as a unique feature to define hidden states. PSSV features a Poisson mixture model with hidden states representing different mutation patterns; PSSV can thus differentiate heterozygous and homozygous SVs in each sample, enabling the identification of those somatic SVs with a heterozygous mutation status in the normal sample and a homozygous mutation status in the tumor sample. An Expectation-Maximization algorithm is used to learn the Poisson distribution parameters. Compared to conventional approaches, PSSV can capture more somatic SVs from the noisy tumor-normal WGS data.

1.5 List of relevant publications

Peer-reviewed Journal Publication

X. Chen, J. Jung, A. N. Shajahan-Haq, R. Clarke, I. M. Shih, Y. Wang1, L. Magnani, T.L. Wang, and J. Xuan, "**ChIP-BIT**: Bayesian Inference of Target genes using a novel joint probabilistic model of ChIP-seq profiles," ***Nucleic Acids Research***, 2015**.** doi: 10.1093/nar/gkv1491.

X. Shi, J. Gu, **X. Chen**, A. N. Shajahan-Haq, L. Hilakivi-Clarke, R. Clarke, and J. Xuan, "**mAPC-GibbsOS**: An integrated approach for robust identification of gene regulatory networks," ***BMC Systems Biology***, 7 (Suppl 5):S4, 2013.

X. Chen, J. Xuan, C. Wang, A. N. Shajahan-Haq, R. B. Riggins, and R. Clarke, "Reconstruction of transcriptional regulatory networks by stability-based network component analysis," ***IEEE/ACM Trans. Computational Biology and Bioinformatics***, vol. 10, issue (6), pp. 1347-1358, 2013.

Manuscript Submitted/In Preparation

X. Chen, X. Shi, A. N. Shajahan-Haq, L. Hilakivi-Clarke, R. Clarke, Y. Wang, and J. Xuan, "**PSSV**: A novel Pattern-based probabilistic approach for Somatic Structure Variation identification," ***Bioinformatics***, 2016 (under R1 revision).

X. Chen, J. Xuan et al., "**BICORN**: Bayesian integration of ChIP-Seq and RNA-Seq data for functional cis-regulatory module identification," in preparation.

Other manuscripts in preparation and conference publications can be found in **Appendix A**.

1.6 Organization of the dissertation

The major objective of this dissertation work is to develop computational methods for a better understanding of the machinery of transcriptional regulation. The remainder of this dissertation is organized as follows:

In Chapter 2, ChIP-BIT is proposed to detect TFBSs. We first demonstrate several distribution hypotheses used in ChIP-BIT by examining real sample and input ChIP-seq data. Then, the methodology development of ChIP-BIT is in detail explained. To demonstate the advantage of ChIP-BIT on TFBS detection over conventional methods, especially on weak bindings, we have applied ChIP-BIT and comparable methods to several simulated ChIP-seq data sets with known binding sites at encoding regions. Finally, we have applied ChIP-BIT to breast cancer ChIP-seq data to help understand the functional roles of NOTCH3 and PBX1 at promoter regions and that of ER-α at enhancer regions.

In Chapter 3, to infer CRMs, BICORN is developed to integrate bindings predicted by ChIP-BIT and RNA-seq gene expression data. We first highlight the importance of module analysis on gene regulation, especially in enhancer regions. Then, a Gibbs sampling algorithm is in detail explained to estimate hidden variables in the proposed BICORN model. We apply BICORN to simulated or benchmark data and demonstrate its advantage on module inference over conventional integrative approaches. BICORN is finally applied to breast cancer MCF-7 ChIP-seq and RNA-seq data for module inference. We infer CRMs respectively at promoter and enhancer regions since the mechanism of gene regulation at these two types of regions are different.

In Chapter 4, we address the problem of somatic SV prediction by proposing a pattern based probabilistic model, PSSV. For each SV type, we examine the read count distributions in the real tumor and normal samples to validate the hypothesis of the proposed mixture Poisson distribution. Using an Expectation-Maximization algorithm, we estimate the model parameters as well as the posterior probability of each SV. We test the robustness of PSSV on simulated DNA-seq data with different noise and sequencing depth settings and further compare the performance of PSSV with those of several existing SV detection tools. We finally apply PSSV to breast cancer patient WGS data and do functional studies on detected somatic SVs.

In Chapter 5, we summarize the major contributions of this dissertation research, lay out some future tasks to further improve the methods, and finally draw conclusions of this dissertation.

2. ChIP-BIT: Bayesian inference of transcription factor binding sites using ChIP-seq data

2.1 Introduction

The advent of chromatin immunoprecipitation with massively parallel DNA sequencing (ChIP-seq) has dramatically accelerated the field of genomic research in gaining an in-depth understanding of complex functions of regulatory elements in the finest scale [5]. ChIP-seq profiling of eukaryote cells has been used successfully to identify histone modifications [50], distal-acting enhancers [51] and proximal transcription factor binding sites (TFBSs) at promoter regions [52]. With the TFBSs identified from ChIP-seq data, it is now possible to reliably define target genes for specific transcription factors (TFs) [53]. If multiple ChIP-seq data sets are available, researchers can investigate the extent of co-association among multiple TFs based on TF-gene binding patterns [54]. Hence, it is important to develop accurate computational approaches for identifying binding sites and target genes from ChIP-seq data [55].

A number of peak-calling methods have been developed to detect peaks using TF ChIP-seq and the matched input data. MACS [33] and PeakSeq [34] are the two most widely used peak calling methods; these tools use local Poisson or Binomial statistics to identify enriched peaks. BCP [35] uses a Bayesian change point approach to model read coverage change along the genome and identifies peak boundaries. Dfilter [36] calculates the form of an optimal detection filter with reads from input data and then applies the filter to reads from sample data for peak prediction. MOSAiCS [37] uses a negative binomial mixture model to identify foreground and background bindings based on read counts from sample and input data. Only TIP and SignalSpider model the weak binding signals existing in the sample ChIP-seq data. However, reliable identification of weak binding signals from background signals (i.e., non-specific binding signals) is a challenging task itself, since it requires a high sequencing depth of both sample and input ChIP-seq data sets [56]. If the sequencing depth is not sufficient or the input data set is not well modelled, existing peak detection methods return a high rate of false positives in the so-called weak binding signals.

In this dissertation research, to reduce the false positive rate in weak binding signal detection, we have proposed a novel probabilistic approach for TFBS identification. Sample and input ChIP-seq data sets are jointly analyzed to reliably identify weak binding signals. Our proposed approach takes into account three major factors that determine the possibility of a encoding region (a promoter or enhancer region) containing a TFBS: sample read intensity, input read intensity and specific for the promoter region, the relative distance to the nearest transcription starting site (TSS) [57]. The basic idea of the ChIP-BIT can be briefly described to highlight its novelty and uniqueness. ChIP-BIT uses a Gaussian mixture model (consisting of global and local Gaussian components) to capture both binding and background signals in the sample data. A unique feature is that the component modeling background signals is specially designed as a local Gaussian distribution that can be estimated accurately from the input data. Specific for the promoter region analysis, an exponential distribution is used to model the relative distance of TFBS to the nearest TSS. Estimated by an Expectation-Maximization (EM) algorithm, a posterior probability is assigned to each TFBS under consideration, indicating the likelihood of a foreground binding event.

To demonstrate the capability of ChIP-BIT on TFBS detection, we have applied it to several simulated or real ChIP-seq data sets with known or validated binding peaks, and compared its performance with several existing peak detection methods. We further applied ChIP-BIT to an in-house ChIP-seq profiling study focused on NOTCH3 and PBX1 to help understand their functional role in breast cancer cells. The effect of NOTCH3 and PBX1 co-regulation on target genes was investigated in conjunction with TF knockdown gene expression data. Our analysis of ChIP-seq data showed that NOTCH3 and PBX1 are also involved in the regulation of Wnt signaling pathway, indicating that crosstalk may exist between the Notch and Wnt signaling pathways. Finally, we applied ChIP-BIT to an enhancer study by examining ER-α activation effect on enhancer regions of breast cancer cells. GRO-seq data measuring enhancer RNA (eRNA) is used to locate active enhancer regions. Our analysis showed that ER-α is a major activator of enhancer regions in breast cancer cells.

2.2 Methods

2.2.1 ChIP-BIT model description

Sample ChIP-seq data and its matched input are jointly analyzed in this framework such that a majority of mappability and GC content biases could be resolved. We first search for genomic regions with read coverage in the sample ChIP-seq data by using uniquely aligned ChIP-seq reads. After accommodating genome mappability variation, filtering out regions containing low read coverage and normalizing the input ChIP-seq data against the sample data, we identify an initial set of candidate genomic regions. We define the scale of a promoter region as ±10k bps of each gene's TSS using gene annotation file [53]; the scale of enhancer region as ±10k bps of the midpoint of an H3K4me1 peak region with enrichment of H3K27ac and depletion of H3K4me3.[58]. We then perform region region partition, map candidate regions to small windows each with a size of 200 base pairs (bps), and calculate read intensities of sample and input data respectively in each window.

The proposed ChIP-BIT method is then used to compute a posterior probability for each window in order to call whether or not a window contains TFBS. As shown in Fig.2.1, ChIP-BIT jointly analyzes sample read intensity and input read intensity using a Bayesian framework. Under the Bayesian framework, ChIP-BIT models the sample read intensity with a Gaussian mixture model, consisting of a global Gaussian component for binding signals and a local Gaussian component for background signals. Importantly, the local Gaussian component can be accurately estimated from the input read intensity, making it possible to detect weak binding signals reliably because of the lower false positive rate. For TFBSs that lay in promoter regions, the relative distance of each TFBS to nearest TSS is modeled by an exponential distribution function; while for those TFBSs at enhancer regions, the distribution of binding locations is assumed to be uniform since they are quite far away from target genes. This distance based distribution is also incorporated into the Bayesian approach of ChIP-BIT. An algorithm of EM estimates a posterior probability that each window contains a true TFBS. Windows with posterior probabilities over a predefined probability threshold are merged together if they are continuous at their genomic locations. The merged windows are finally reported as TFBSs for potential binding events to occur.

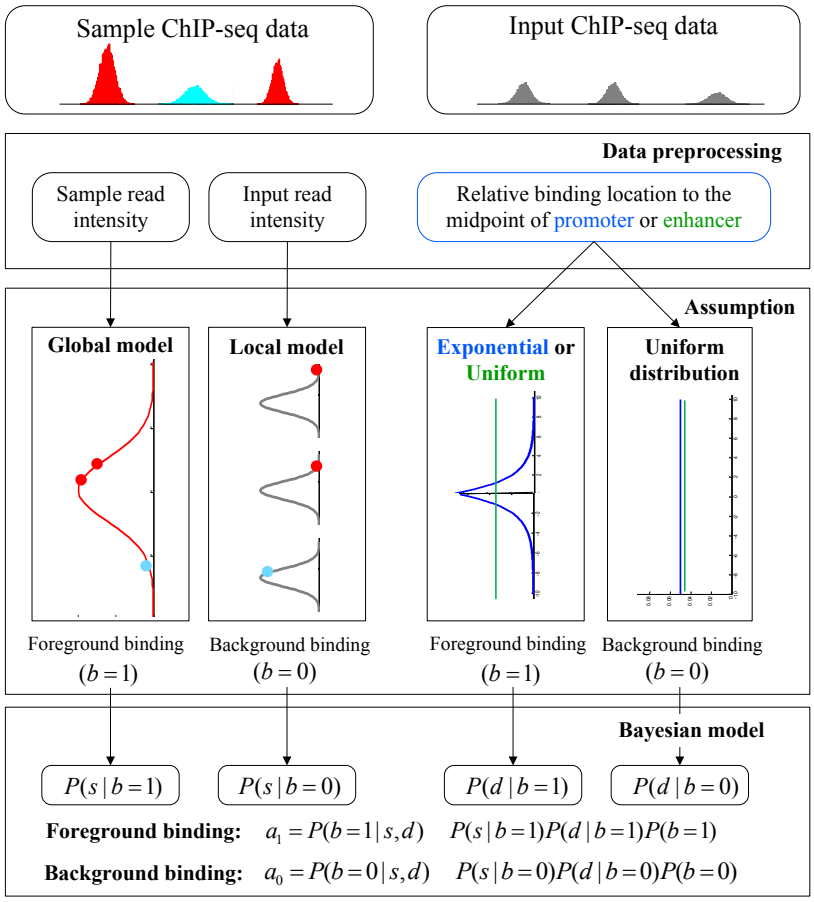


Figure 2.1 Flowchart of the proposed ChIP-BIT approach. ChIP-BIT features (1) a joint analysis of sample and input ChIP-seq data with a unique Gaussian mixture model, and (2) a Bayesian framework to incorporate the location information of TFBS.

The novelty of ChIP-BIT mainly is main reflected by modelling read intensity s, the natural log of the read coverage, as shown in Fig. 2.2. ChIP-seq data can be treated as a mixture of TFBSs and background signals. Background signals are fully represented in the input data, but TFBSs need to be distinguished from background signals in the sample data. Hence, each enriched region has two hidden states: binding occurrence $b=1$ and non-occurrence $b=0$, with probabilities a_1 and a_0, respectively. The sum of these two probabilities equals to 1. If $a_1 > a_0$, the region is more likely to contain a TFBS, shown as a 'red' bar in Fig. 2.2; otherwise, it is a background region, shown as a 'blue' bar in Fig. 2. Read intensity is a major signature to help identify TFBS. The value of read intensity s in the sample data ('red' or 'blue' bars) and its differentiation against s_{input} from the input data ('gray' bars) altogether determine whether or not each binding event is true.

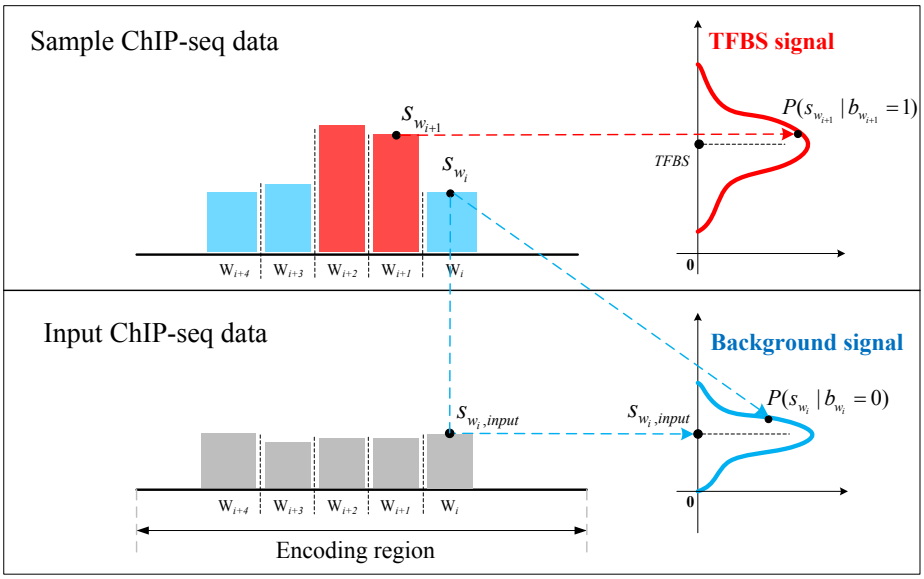
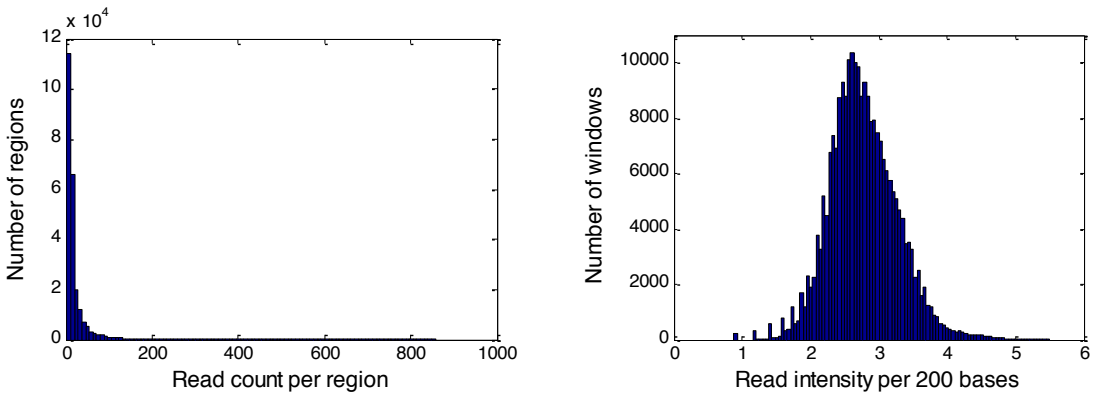


Figure 2.2 Model description of ChIP-BIT on peak detection and target gene identification. A 'red' bar represents the read intensity of a TFBS and a 'blue' bar represents the read intensity of a background region. Those 'gray' bars represent input signals at the same locations of any 'red' or 'blue' bars. For each window, read intensities from sample and input data are jointly analyzed for reliable TFBS identification.

2.2.2 Hypothesis on distributions in ChIP-BIT

Gaussian distribution hypothesis on read intensity

Traditionally, people directly model on the raw read count N_{region} at a region for TFBS prediction. We present the histogram of read count N_{region} for all regions in Fig. 2.3(a). It can be expected that a wider region is more possible to contain a relatively larger N_{region}. Therefore, it is not necessary that a region having high read count is more possible to contain a TFBS. Some wide and false positive regions will be included in the report list. Read intensity s for a 200 bps long window is defined as the natural log transformation of the accumulated read coverage. We present read intensity for all windows in Fig. 2.3(b), where a Gaussian distribution can be clearly observed. Using read coverage information in each window we can evaluate read enrichment fairly for peaks with different length. If a wide region is a true peak, it will have multiple consecutive highly enriched windows. In addition, as explain in Chapter 1, using the Gaussian distribution we successfully shrink the distance between strong and weak bindings. Since local input data is also used in a joint framework, we can still distinguish foreground bindings from background signals.

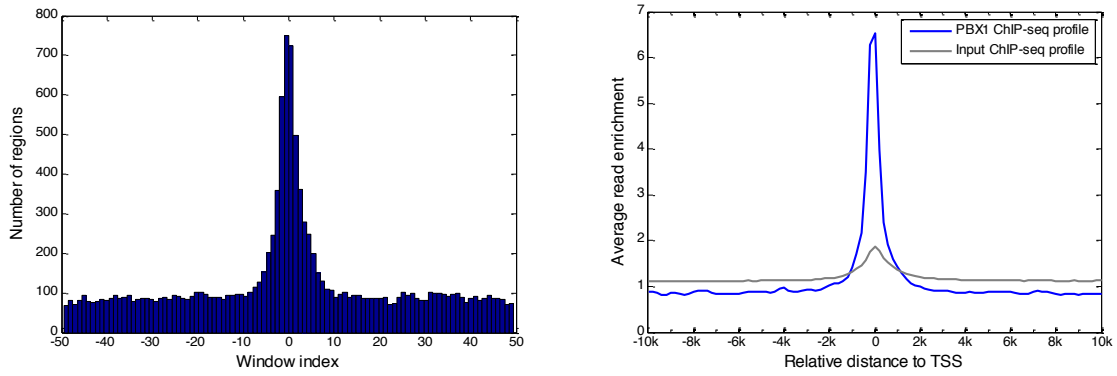


(a) (b)

Figure 2.3 Illustrations of read count and read intensity of TFBS. (a) read count per region; (d) read intensity per 200 bps window.

Exponential distribution hypothesis on binding locations at promoter regions

After literature review, we found that the regulatory effect of a TFBS that lays in promoter regions decays exponentially when the relative distance d increases [54, 55, 59-61]. This decaying effect is TF specific [59] and the parameter of its exponential distribution needs to be properly optimized so as to provide the most reliable gene prediction performance [54] (which has been demonstrated to be better than binary peak-gene assignment [55]). We have examined the distribution of binding locations of PBX1 in real ChIP-seq data using similar ways as in [60, 61], as shown in Fig. 2.4.



(a) (b)

Figure 2.4 PBX1 binding at gene promoter regions. (a) Number of regions in each window; (b) average read enrichment of each window.

From Fig. 2.4(b), it can be clearly seen that the average read enrichment in the sample ChIP-seq data follows an exponential distribution along promoter regions. When d is very large, the average read enrichment decreases significantly. However, in the input ChIP-seq data the average read enrichment is relatively uniformly distributed. Therefore, among regions with different distances, the probability of containing a foreground binding for a region with small d is assumed to be high. As illustrated in Fig. 2.5, the shorter d is, the more possible that observed read intensity s is sequenced from an effective TFBS.

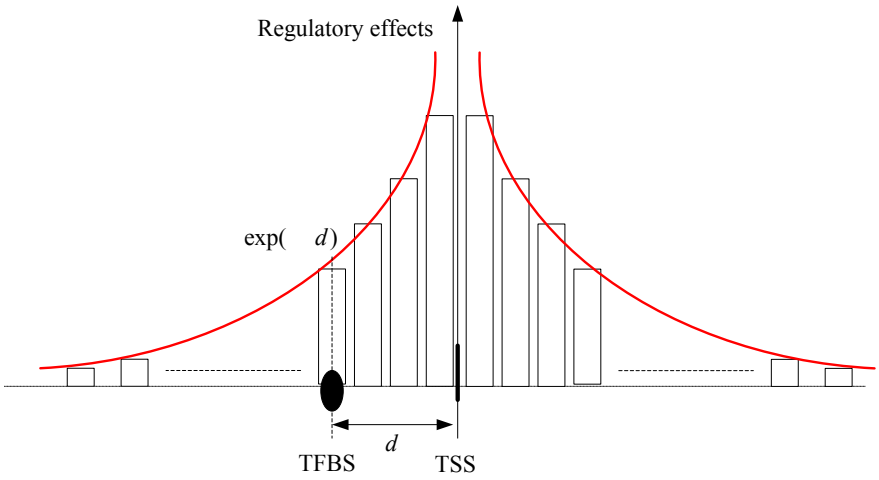


Figure 2.5 Exponentially distributed weights of TFBS enriched windows at gene promoter regions.

Uniform distribution hypothesis on binding location at enhancer regions

Enhancer regions are usually very distant from TSSs. From previous Fig. 2.4(b) it can be expected that in the sample ChIP-seq data, the distribution of binding signals should be uniform when they are far away from TSS. Here we have also examined average read enrichment at distant enhancer regions, as shown in Fig. 2.6. It can be seen that the average read enrichment in sample ChIP-seq data is relatively flat at ±1k region to the midpoints of enhancer regions. And within each window (200 bps), the average binding signal enrichment levels in sample and input ChIP-seq profiles are similar. Therefore, we use uniform distributions to model the weight of each window in sample and input ChIP-seq data.

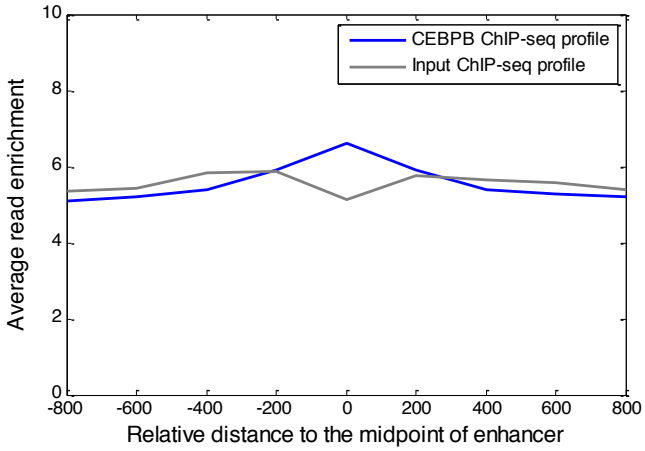


Figure 2.6 CEBPB average binding signal enrichment at enhancer regions.

2.2.3 Bayesian framework of ChIP-BIT

We use n to index encoding regions (promoter or enhancer regions) and w to index partitioned windows of each. A binary binding variable $b_{n,w}$ is defined as binding occurrence ($b_{n,w}=1$) and non-binding occurrence ($b_{n,w}=0$). And for each binding variable $b_{n,w}$, there are three observations as read intensity $s_{n,w}$ in the sample data, read intensity $s_{n,w,input}$ in the input data, and binding location information $d_{w,n}$ (referring to the midpoint of encoding region). For each state (1 or 0) of $b_{n,w}$, we define a posterior probability $a_{n,w,i}$ as:

$$a_{n,w,i} = P(b_{n,w} = i \,|\, s_{n,w}, d_{n,w}), i = 0,1. \tag{2-1}$$

Given a TFBS, the relationship between its binding signal intensity and binding location is assumed to be conditionally independent. Then, Eq. (2-1) can be extended as:

$$a_{n,w,i} = \frac{1}{C_{n,w}} P(s_{n,w} \,|\, b_{n,w} = i) P(d_{n,w} \,|\, b_{n,w} = i) P(b_{n,w} = i), \tag{2-2}$$

where $C_{n,w} = \sum_{b_{n,w}=0,1} P(s_{n,w}, d_{n,w} \,|\, b_{n,w}) P(b_{n,w})$ is a normalization factor.

Read intensity $s_{n,w}$ follows a Gaussian distribution [62], as $P(s_{n,w} \mid b_{n,w}=1)$ and $P(s_{n,w} \mid b_{n,w}=0)$ in Fig. 2.2. The conditional probability $P(s_{n,w} \mid b_{n,w})$ in Eq. (2-2) can be calculated as:

$$\begin{aligned} P(s_{n,w} \mid b_{n,w}=1) &\sim N(\ _{TFBS},\ ^{2}_{TFBS}), \\ P(s_{n,w} \mid b_{n,w}=0) &\sim N(s_{n,w,input},\ ^{2}_{input}). \end{aligned} \tag{2-3}$$

where for $b_{n,w}=1$, $s_{n,w}$ is sequenced from a TFBS so it follows a global Gaussian distribution with mean $_{TFBS}$ and variance $^{2}_{TFBS}$; while for $b_{n,w}=0$, $s_{n,w}$ is sequenced from background region so it follows a local Gaussian distribution with mean $s_{n,w,input}$ (its input signal) and variance $^{2}_{input}$. $^{2}_{input}$ is the variance of background signals, which can be directly calculated from the input ChIP-seq data. Values of $_{TFBS}$ and $^{2}_{TFBS}$ are unknown, which need to be estimated from all $s_{n,w}$ with $b_{n,w}=1$.

The second conditional probability in Eq. (2-2), $P(d_{n,w} \mid b_{n,w})$, is determined by the relative binding location $d_{n,w}$ as well as the binding state $b_{n,w}$. $d_{n,w}$ is defined as the distance between middle point of the w-th window to middle point of n-th encoding region.

$$d_{n,w} = (w + sign(w) * \frac{1}{2})\ \ d, \tag{2-4}$$

where d=200 bps denotes window size, and $w=\{0,\pm1,\pm2,...,\pm(W\ \ 1)\}$. The total number of windows at one side of encoding region, W, is defined by d_p / $\ d$. d_p represents the half length of each encoding region. We set d_p=10k bps for a promoter region and 4k bps for an enhancer region.

Different TFs may have decaying speed on their binding signal enrichment at promoter regions (as shown in Fig. 2.5). The binding enrichment follows an exponential distribution with parameter , as shown in Eq. (2-5). For background signal $s_{n,w}$ with state $b_{n,w}=0$,

$P(d_{n,w} \mid b_{n,w} = 0)$ is assumed to follow a uniform distribution with a discrete probability density function d / d_p.

$$P(d_{n,w} \mid b_{n,w} = 1) = \frac{1}{2} e^{\quad |w| \ d} (1 \quad e \quad ^{d}),$$
$$P(d_{n,w} \mid b_{n,w} = 0) = \frac{1}{2} \quad d / d_p. \tag{2-5}$$

where parameter needs to be estimated from all $d_{n,w}$ with $b_{n,w} = 1$.

At enhancer regions, the binding enrichment $s_{n,w}$ ($b_{n,w} = 1$) of TFBSs binding at different locations doesn't have a significant change (as shown in Fig. 2.6), so we use a uniform distribution to model the possibility $P(d_{n,w} \mid b_{n,w} = 1)$. For background signal $s_{n,w}$ with state $b_{n,w} = 0$, $P(d_{n,w} \mid b_{n,w} = 0)$ is assumed to follow a uniform distribution as well.

$$P(d_{n,w} \mid b_{n,w} = 1) = P(d_{n,w} \mid b_{n,w} = 0) = \frac{1}{2} \quad d / d_p. \tag{2-6}$$

For the prior probability $P(b_{n,w})$ in Eq. (2-2), since there is no prior knowledge about the proportion of regions that containing TFBSs, we define the prior probability as:

$$P(b_{n,w}) = \quad ^{b_{n,w}} (1 \qquad)^{(1 \ b_{n,w})} . \tag{2-7}$$

From the above discussion, we have formulated the TFBS detection problem as a parameter estimation problem: how to estimate parameters , $_{TFBS}$, $^{2}_{TFBS}$ and hence to calculate the posterior probability $a_{n,w,1}$. We assume uniform prior on $_{TFBS}$, uniform prior on , and inverse Gamma prior on $^{2}_{TFBS}$ (conjugate prior of Gaussian distribution) and Beta distribution [63] (conjugate prior of Binomial distribution) on as follows:

$$P(\ _{TFBS}) = 1 / C \ , \tag{2-8}$$

$$P(\ \ ) = 1 / C \ , \tag{2-9}$$

$$P(\ ^2_{TFBS}) = inverseGamma(\ ,\ ), \tag{2-10}$$

$$P(\ ) = Beta(\ _0,\ _1), \tag{2-11}$$

where and are hyper-parameters for inverse Gamma distribution and $_0$ and $_1$ are hyper-parameters for Beta distribution.

We assume inverse Gamma distribution on variance $^2_{TFBS}$ because we want to limit most binding signals around $_{TFBS}$. This assumption is consistent with the ChIP-seq data generation process, since most segments selected ('picked up') by the antibody are sequenced to a similar depth. Some background regions are also sequenced by the NGS machine but, due to the lack of antibody selection, their sequence depth is lower than binding regions. In addition, after read tag assembly, there are usually some segments showing extremely high coverage. As demonstrated in [64], these regions are highly possibly caused by segment duplication (high copy number), not TF-specific binding locations. Therefore, we assume an inverse Gamma distribution on $^2_{TFBS}$ to lower the impact of noise on TFBS binding signal distribution estimation.

To estimate the parameters described above, we define a second posterior probability as:

$$\begin{aligned} P(\ ,\ _{TFBS},\ ^2_{TFBS},\ \ |\mathbf{s},\mathbf{d}) &= \frac{1}{C_1} \prod_n \prod_w P(s_{n,w}, d_{n,w}|\ _{TFBS},\ ^2_{TFBS},\ ,\ ) \frac{1}{C}\frac{1}{C} P(\ )\ P(\ ^2_{TFBS}) \\ &= \frac{1}{C_2} \prod_n \prod_w \sum_{b_{n,w}=0,1} P(s_{n,w}\,|\,b_{n,w}) P(d_{n,w}\,|\,b_{n,w}) P(b_{n,w}) \quad P(\ )\ P(\ ^2_{TFBS}) \end{aligned}, \tag{2-12}$$

where C_1 and $C_2 = C_1 C\ C$ are constant values.

Based on Eq. (2-12), we can estimate all parameters (, $_{TFBS}$, $^2_{TFBS}$ and) and the posterior probability ($a_{n,w,1}$) using an EM approach. We set the initial values of all $a_{n,w,1}$ to 0.5, beta distribution parameters $_0$ and $_1$ to 5.0, and inverse gamma distribution parameters and to 1.0. Then, we carry out the E-step and the M-step iteratively to estimate all parameters until the improvement of the posterior probability defined in Eq. (11) is less than 1.0e-6. Note that the

distance distribution parameter is only estimated when ChIP-BIT is used to predict TFBSs at promoter regions. The E-step and M-steps are mathematically detailed as follows:

E-step:

$$\hat{a}_{n,w,i} = \frac{P(s_{n,w} \mid b_{n,w} = i)P(d_{n,w} \mid b_{n,w} = i)\ ^{i}(1\quad)^{(1\ i)}}{\sum_{b_{n,w}=0,1} P(s_{n,w} \mid b_{n,w})P(d_{n,w} \mid b_{n,w})\ ^{b_{n,w}}(1\quad)^{(1\ b_{n,w})}},\ i = 0,1. \qquad (2\text{-}13)$$

M-step:

$$= \frac{\hat{a}_{n,w,1} + (\ _{1}\ \ 1)}{T + \ _{0} + \ _{1}\ \ 2}, \qquad (2\text{-}14)$$

$$_{TFBS} = \sum_{n\ \ w} \hat{a}_{n,w,1} s_{n,w} \Big/ \sum_{n\ \ w} \hat{a}_{n,w,1}, \qquad (2\text{-}15)$$

$$^{2}_{TFBS} = \frac{2\ \ + \sum_{n\ \ w} \hat{a}_{n,w,1}(s_{n,w}\quad _{TFBS})^2}{(2\ \ + 2 + \sum_{n\ \ w} \hat{a}_{n,w,1})}, \qquad (2\text{-}16)$$

$$= \frac{1}{d} \ln \frac{\sum_{n\ \ w} \hat{a}_{n,w,1} + \sum_{n\ \ w} \hat{a}_{n,w,1} |w|}{\sum_{n\ \ w} \hat{a}_{n,w,1} |w|}. \qquad (2\text{-}17)$$

Based on our experiments on several ChIP-seq data sets, iteration of the E-step and the M-step usually converges within 50 rounds. For ChIP-seq data analysis, we select as high confident binding events those windows whose probabilities are over 0.9. Since each binding event has been associated with a target gene, we obtain a target gene list as identified for a particular TF under investigation. More details about the mathematical deviation for each variable can be found in **Appendix B**.

2.3 Simulation

2.3.1 TFBS simulation at proximal promoter regions

We generated a list of realistic genomic regions by applying PeakSeq [34] to MYC ChIP-seq data acquired from a leukemia study of cell line K562 in the ENCODE project (http://genome.ucsc.edu/ENCODE/). To simulated TFBSs at promoter regions, we selected those peaks falling on chromosome 1 and mapped them to the UCSC hg19 gene annotation file. Consistent with our model design, we set the promoter region as ±10k bps around TSS and about 6,000 proximal peaks in promoter regions were extracted for simulation data generation. We observed an exponential-like distribution of the relative distances of these peaks to the TSS (not shown here). We randomly selected half of the peaks as 'true' peaks and treated the remaining regions as 'background'. Using a simulation tool developed in [65], we simulated sample data and input data with the same total number of reads (denoted as Case 1). We called peaks using ChIP-BIT and other comparable peak calling tools; a successful call is counted if a detected peak overlaps with any 'true' peak by 50% in the promoter region. The precision/recall performances of ChIP-BIT and other competing methods on peak detection are shown in Fig. 2.7. F-measure was calculated to assess the overall performance of each method regarding its precision-recall performance, as summarized in Table 2.1 for all the methods in comparison. Note that the denominator used in the precision calculation is the number of peaks within promoter regions.

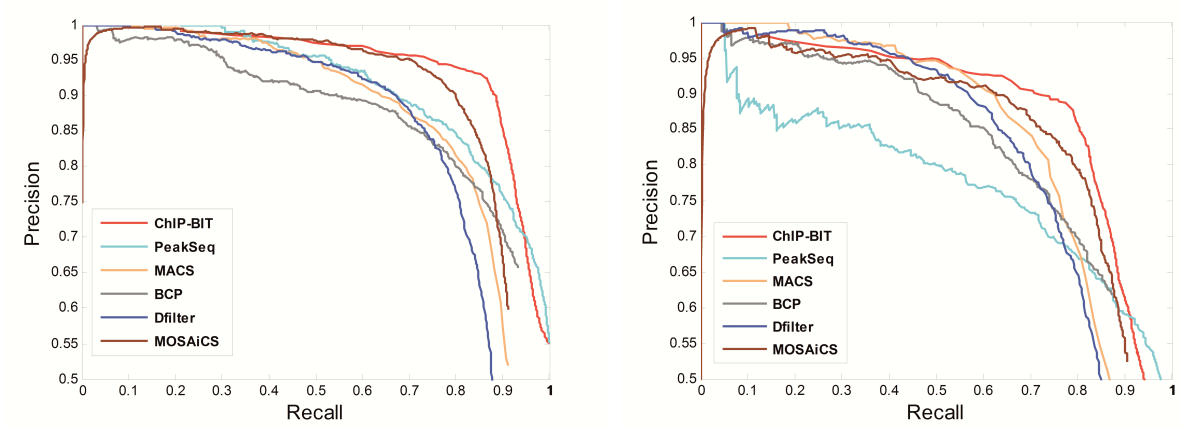


Figure 2.7 Precision and recall performance of ChIP-BIT and existing peak calling methods in simulation of Case 1. (a) Detection performance on all peaks; (b) detection performance on weak binding signals.

Table 2.1 Overall precision-recall performance, F-measure, on peak detection of Case 1.

METHOD	ChIP-BIT	MOSAiCS	PeakSeq	MACS	BCP	Dfilter
Case 1 (all peaks)	**0.905**	0.848	0.829	0.810	0.808	0.793
Case 1 (weak peaks)	**0.832**	0.799	0.735	0.770	0.749	0.745

From Fig. 2.7(a), we can see that the precision/recall performance of ChIP-BIT is the best, where the joint modeling of sample and input data and informative distance distribution significantly promote the precision of ChIP-BIT. As shown in Fig. 2.7(b), ChIP-BIT has a strong detection capability on weak binding signals (whose read intensity are lower than the mean value of read intensities in sample data), which promotes the overall performance of ChIP-BIT. From Fig. 2.7 and Table 2.1, it can be seen that the detection performance of ChIP-BIT on simulated peaks in the gene promoter region is better than existing tools, especially in terms of its detection performance on weak binding signals.

2.3.2 TFBS simulation at distant enhancer regions

To show that ChIP-BIT still works well to predict TFBSs at enhancer regions, when the distance distribution is relatively uniform at different locations, we simulated another pair of data sets (denoted as Case 2) based on a set of pre-selected enhancer regions enriched in breast cancer MCF-7 ER-α ChIP-seq data. In total we had 8,000 regions and we randomly selected half of them as foreground. In Case 2, the distribution of peaks in the enhancer region was relatively uniform and the read intensity difference between TFBSs and background regions was smaller than Case 1. We applied ChIP-BIT and competing peak callers to the simulated sample and input ChIP-seq data of Case 2. Their prediction-recall curves for strong and weak peak detection are shown in Fig. 2.8 and the F-measure of each competing method is summarized in Table 2.2

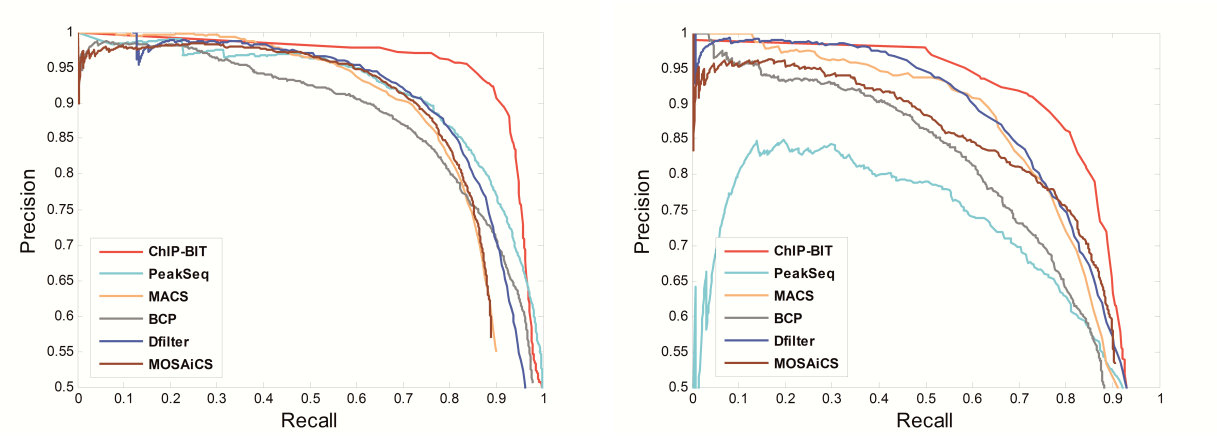


Figure 2.8 Precision and recall performance of ChIP-BIT and existing peak callers in simulation of Case 2. (a) Detection performance on all peaks; (b) detection performance on weak binding signals.

Table 2.2 Overall precision-recall performance, F-measure, on peak detection of Case 2.

METHOD	ChIP-BIT	MOSAiCS	PeakSeq	MACS	BCP	Dfilter
Case 2 (all peaks)	**0.908**	0.818	0.840	0.812	0.804	0.766
Case 2 (weak peaks)	**0.833**	0.778	0.708	0.769	0.725	0.723

As we can see from the figure and table, the overall detection performance of ChIP-BIT in Case 2 is better than competing methods, as shown in Fig. 2.8(a). The improvement is mainly due to the joint modeling of read intensities of the sample and input data. The Gaussian mixture model has enabled ChIP-BIT to detect more weak binding signals than any other existing method, as clearly shown in Fig. 2.8(b).

Since strong peaks, weak peaks and background signals are always mixed together in the sample ChIP-seq data. At either promoter or enhancer region, detection of weak peaks may include some background signals in the result list. In Fig. 2.9, we further present the false positive rate of weak peak detection of each method, when the overall precision (precision on all peaks) is fixed at the same level (in a range of 0.70 – 0.95) for all the methods in comparison. From the figures, we can clearly see that ChIP-BIT has the lowest false positive rate on weak

binding prediction, a major benefit of ChIP-BIT's joint modelling of read intensity in the sample and input ChIP-seq data. Consequently, in this stimulation study, the ability of ChIP-BIT in differing foreground bindings from background signals is better than the other competing methods, resulting in an improved performance of ChIP-BIT.

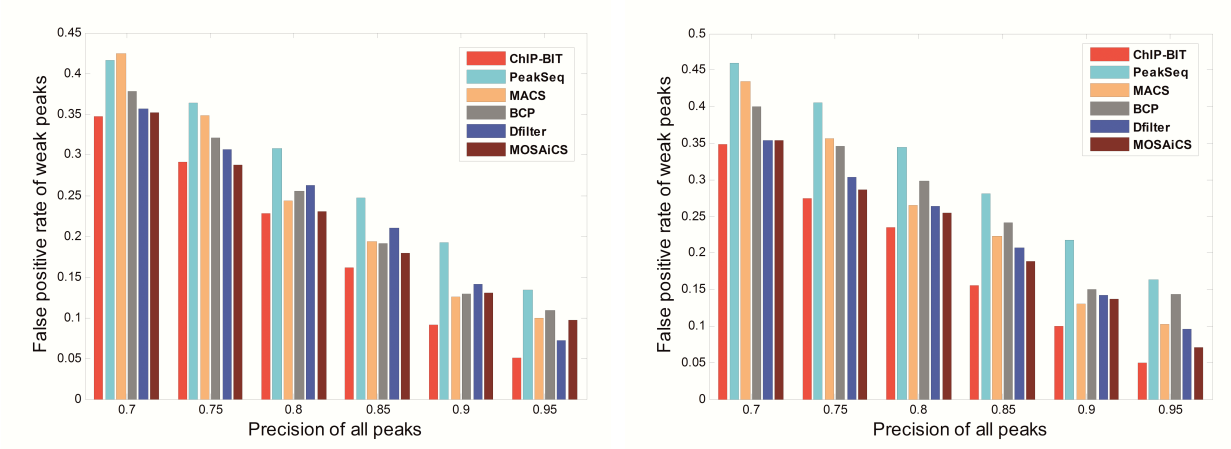


Figure 2.9 False positive rate of the detected weak binding signals of ChIP-BIT and existing peak calling methods. (a) Simulation Case 1; (b) simulation Case 2.

2.4 Breast Cancer MCF-7 cells ChIP-seq data analysis

2.4.1 NOTCH3 and PBX1 binding sites identification at promoter regions

NOTCH3-mediated signaling plays an important role in the proliferation of breast cancer cells and has emerged as a possible therapeutic target [66]. In Notch signaling pathway, previously we reported that PBX1 is a target of NOTCH3 in ovarian cancer [67]. An interaction of PBX1 and NOTCH3 in breast cancer cells is implied by the correlation of their gene expression data and the target genes identified from PBX1 ChIP-seq data [68]; NOTCH3 and PBX1 control the expression of a large number of genes associated with endocrine therapy resistance in breast cancer cells. We have acquired NOTCH3 ChIP-seq data from MCF7 cells of to further investigate the association between PBX1 and NOTCH3. The raw distributions of read intensity and relative distance to TSS for NOTCH3 or PBX1 can be found in Fig. 2.10.

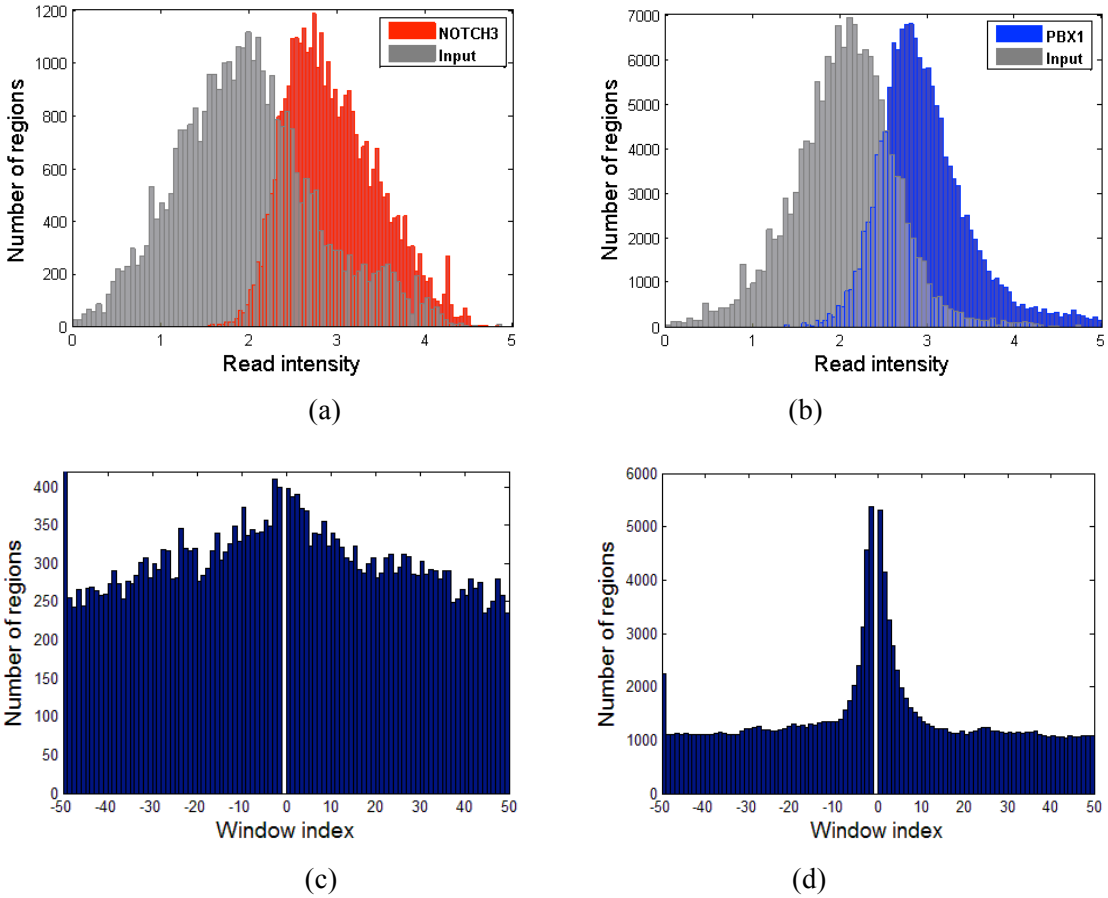


Figure 2.10 Raw distributions of NOTCH3 and PBX1 ChIP-seq data. (a) Histogram of read intensity at NOTCH3 candidate regions; (b) Histogram of read intensity at PBX1candidate regions; (c) histogram of relative distance of NOTCH3 candidate regions; (d) histogram of relative distance of PBX1 candidate regions.

Following the procedure shown in Fig. 2.1, we identified 2,871 TFBSs at promoter regions by applying ChIP-BIT to NOTCH3 ChIP-seq data. We also applied ChIP-BIT to a PBX1 ChIP-Seq data set acquired from MCF7 cells [68] and identified 5,280 TFBSs. In total we identified 621 common target genes with TFBSs from both factors. For comparison, we called peaks using PeakSeq or MACS and predicted target genes using GREAT by setting the promoter region as ±10k bps around TSS. TIP was applied to the sample ChIP-seq data only to directly predict target genes. Since PeakSeq reports read enrichment in sample or input ChIP-seq data for each detected peak, its read intensity distributions of the detected peaks are comparatively shown with those of ChIP-BIT for NOTCH3 or PBX1 in Fig. 2.11.

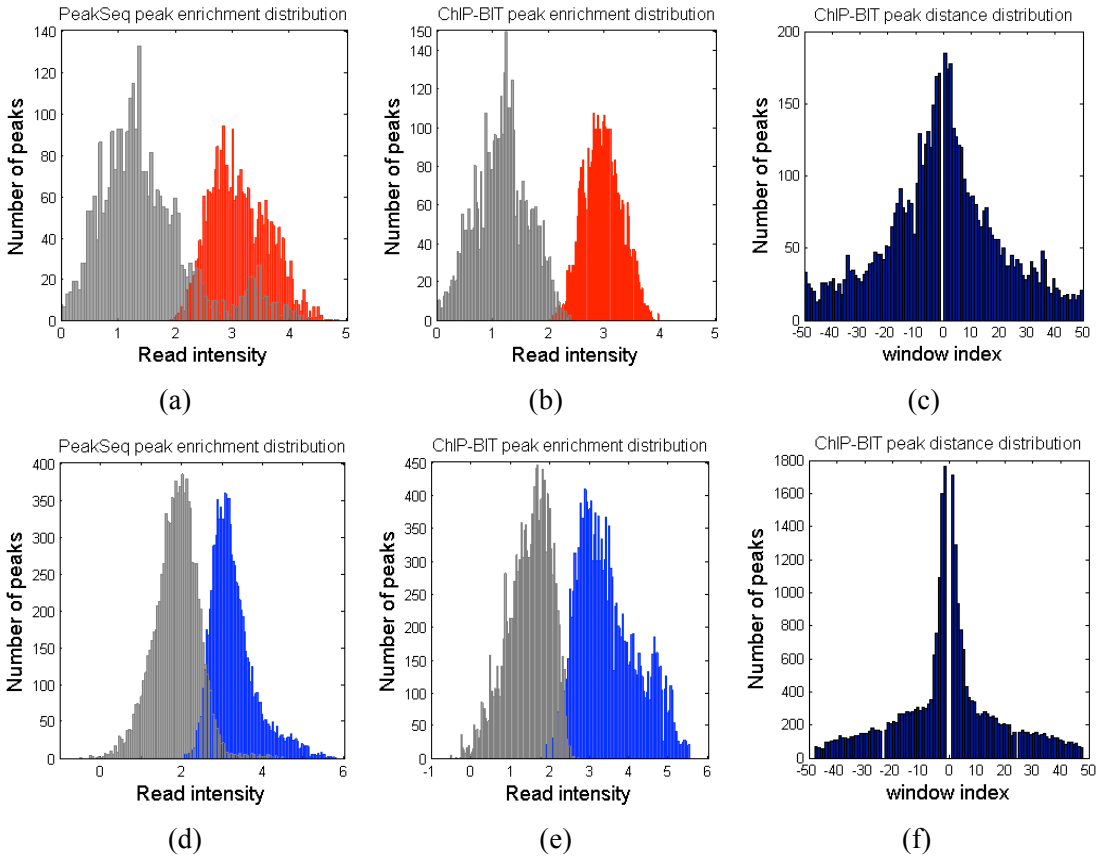
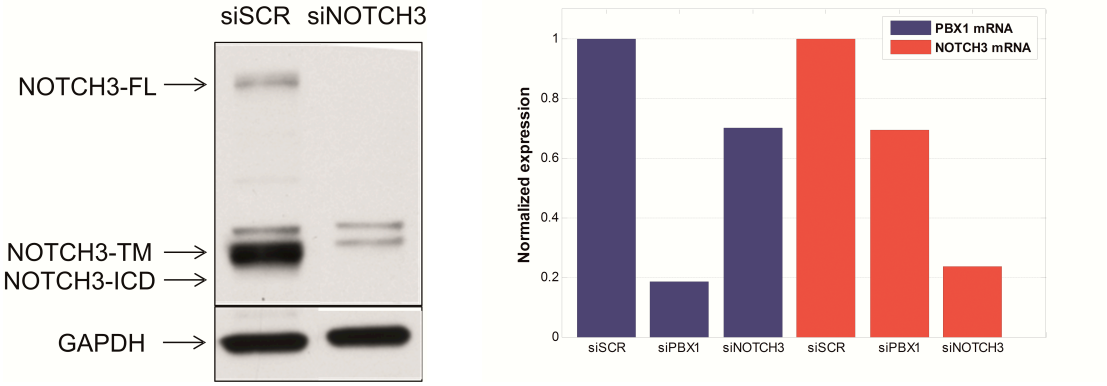


Figure 2.11 Peak calling results of NOTCH3 or PBX1 by using ChIP-BIT and PeakSeq. (a) Read intensities of PeakSeq detected NOTCH3 peaks in sample data (red) and input data (gray); (b) read intensities of ChIP-BIT detected NOTCH3 peaks; (c) relative distances of ChIP-BIT detected peaks to TSS; (d), (e) and (f) represent the same set of information obtained from PBX1 ChIP-seq data analysis as that of (a), (b) and (c).

Read intensity distributions of NOTCH3 peaks reported by PeakSeq and ChIP-BIT are shown in Fig. 2.11(a) and (b), respectively. We can see from Fig. 2.11(b) that ChIP-BIT separates binding signals from sample and input data sets quite well. Although some peaks have relative 'weak' enrichment in the NOTCH3 sample profile, their fold changes are large enough for peak detection. The average fold change of read enrichment between sample and input data is 6.57 for the NOTCH3 peaks identified by ChIP-BIT, which is higher than a commonly used fold change threshold of 4. Even though our exponential distribution assumption assigns higher weights to those peaks close to the TSS, as shown in Fig. 2.11(c), some distant peaks can still be detected if their enrichment in the sample data is significantly higher than that in input data. For PBX1, ChIP-BIT also provides a better separation of sample and input binding signals as well, as shown in Fig. 2.11(e). The average fold change of read enrichment between sample and input

data is 7.26 for the PBX1 peaks identified by ChIP-BIT, also much higher than a commonly used fold change threshold of 4. By comparing Fig. 2.11(f) - (c), we can see that TFBS location-wise distributions are different for different TFs.

As an initial step to validate the ChIP-BIT identified target genes co-regulated by NOTCH3 and PBX1, we used gene expression data acquired from inhibition and knockdown experiments of NOTCH3 and PBX1. First, we inhibited Notch signaling with 1 μM GSI (γ-secretase inhibitor I, EMD Chemicals, San Diego, CA), a well-known Notch signaling inhibitor [67], for 48 hours; both NOTCH3 and PBX1 are inhibited by GSI. Since most ChIP-seq target genes are marginally differentially expressed [69], we set a relatively low fold change threshold, 1.3 (0.3 in its log2 format), for differential expression analysis. 331 (53%) of ChIP-BIT identified NOTCH3-PBX1 common targets are differentially expressed. Second, we used small interfering RNA (siRNA) to knockdown NOTCH3 and PBX1 specifically, since siRNAs are more specific inhibitors of NOTCH3 and PBX1 than the GSI. PBX1 knockdown experiment using siPBX1 was performed previously and microarray gene expression data were made available to us by the authors [68]. In this study, we knocked down NOTCH3 using siNOTCH3 and acquired gene expression data for NOTCH3-PBX1 co-regulated target gene validation. For the knockdown experiment, MCF7 cells were transfected with NOTCH3 siRNA for 48 hours followed by Western blotting to confirm the knockdown efficiency as shown in Fig. 2.12(a). NOTCH3 expression was clearly inhibited with siNOTCH3 compared with its negative control.



(a) (b)

Figure 2.12 TF knockdown experiments for target gene validation. (a) Western blot of NOTCH3 protein expression after transfecting MCF7 cell with siRNA of NOTCH3 and scramble (SCR) for 48 hours, including full length (FL), transmembrane form (TM) and intracellular domain (ICD); (b) mRNA expression levels of PBX1 and NOTCH3 across siRNA samples.

We then performed microarray mRNA profiling by using Illumina HumanHT-12 v4 Expression BeadChip before and after siNOTCH3 is introduced to the cells. Under each condition, we generated two replicates. The normalized mRNA expression of NOTCH3 for each siRNA sample is shown in Fig. 2.12(b). NOTCH3 expression was down regulated significantly (*p*-value 9.7e-3) with siNOTCH3. Among the common genes targeted by NOTCH3-PBX1, gene transcription can be regulated by PBX1, NOTCH3 or both. Knockdown of either TF will provide a set of differentially expressed genes among 621 common target genes, but these two sets of genes are not necessarily the same. With a predefined fold change threshold of 1.3 (0.3 in log2 format), there are 62 genes differentially expressed in the siPBX1 sample, 81 genes differentially expressed in the siNOTCH3 sample and 50 genes differentially expressed in both siRNA samples. In total, we identified 193 differentially expressed target genes. It is expected that a majority of genes identified from siRNA experiments be also differently expressed after GSI treatment because GSI will inhibit both PBX1 and NOTCH3 simultaneously (as described before). After ccomparing to the differentially expressed genes (331 genes) with or without GSI treatment, we found 149 genes that overlapped.

Functional enrichment analysis was carried out on these 149 differentially expressed common target genes predicted by ChIP-BIT. As shown in Table 2.3, Notch signaling pathway and Wnt signaling pathway are enriched with *p*-values of 6.7e-4 and 3.2e-3, respectively. It has been known that Notch and Wnt signaling pathways share some common functions [70] and cooperate in breast and ovarian cancers [71, 72]. Both pathways are related to cancer stem cells, which are regarded as a main source of cancer recurrence and chemo-resistance [73]. Our computational analysis supports the hypothesis that crosstalk between these two pathways may exist downstream of transcription factors PBX1 and NOTCH3. The existence of such crosstalk would require further experiments for biological validation so as to establish the association of Notch and Wnt signaling pathways with the development and progression of breast cancer. Also from Table 2.3 it can be found that functional enrichment of target genes predicted by selected competing methods on these two specific pathways are not significant. Conventional tools only predict strong bindings and miss a lot of weak but still functional bindings. As a result, using conventional tools we can only capture a small proportion of functional genes which are not significant enough to claim biological importance.

Table 2.3 Functional enrichment analysis of target genes predicted by ChIP-BIT and competing methods.

METHOD	ChIP-BIT	PeakSeq	MACS	TIP
NOTCH signaling pathway	**11**	4	0	8
p-value	**6.7e-4**	>0.1	-	>0.1
Wnt signaling pathway	**11**	3	5	7
p-value	**3.2e-3**	>0.1	>0.1	>0.1

2.4.2 ER-α binding site identification at distant enhancer regions

The enhancer region is a specific kind of genomic region with specific epigenetic regulation of histone markers (HMs) [18]. As shown in Fig. 2.13(a), the enhancer region harbors enrichment of both H3K4me1 and H3K27ac with depletion of H3K4me3.

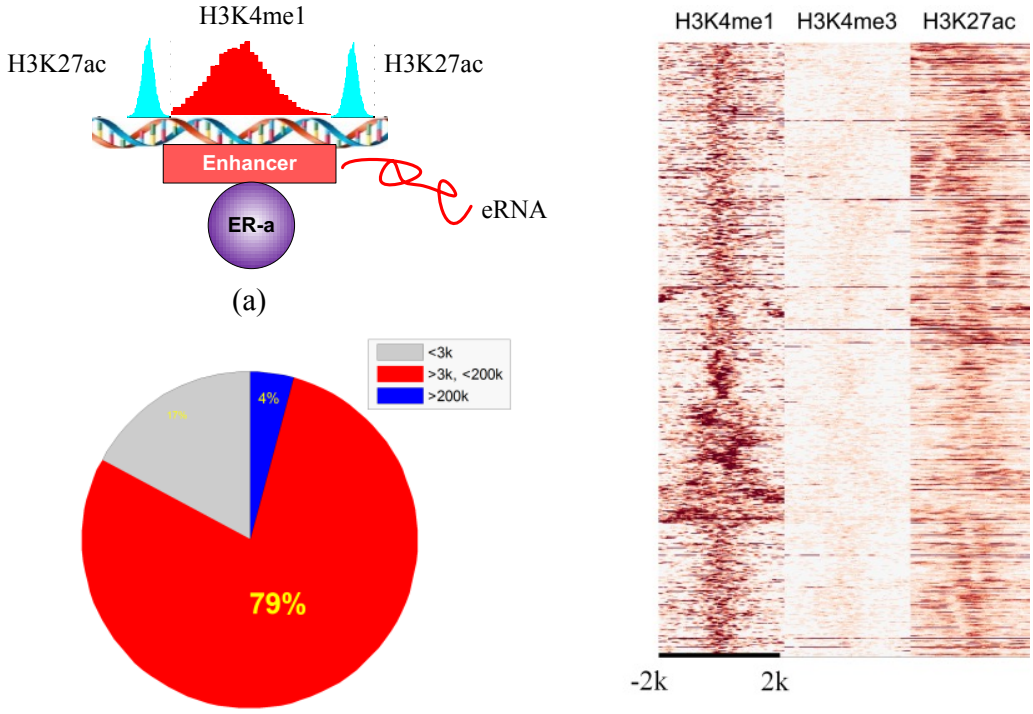
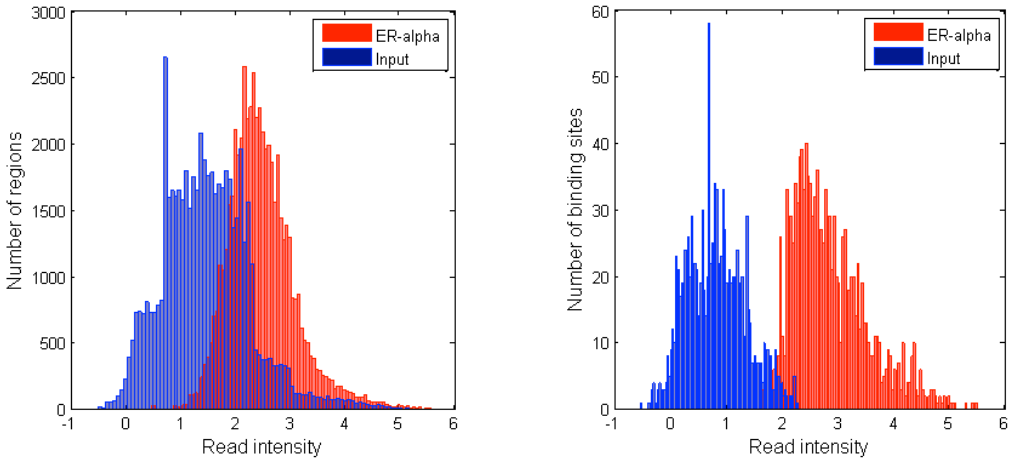


Figure 2.13 Candidate enhancer region identification using H3K4me1, H3K27ac and H3K4me3 ChIP-seq data. (a) Illustration of ERα activated enhancers; (b) relative distance to the nearest transcription starting site of candidate enhancer regions; (c) histone modification signal enrichment at candidate enhancer regions.

We downloaded E2 treated breast cancer MCF-7 ChIP-seq data of H3K4me1 and H3K4me3 (http://www.ncbi.nlm.nih.gov/geo, under accession number GSE23701) and H3K27ac (GSE45822). HM peaks are called using HOMER with wide peak parameter setting [74]. In total, we identified 7,961 H3K4me1-positive, H3K27ac-postive and H3K4me3-negative candidate enhancer regions, as shown in Fig. 2.13(c). We mapped each of these candidate enhancer regions to the nearest TSS and found that at least 79% of enhancer regions were located between 3k bps and 200k bps away from TSS, as shown in Fig. 2.13(b).

It has been known that ERα is a major activator of enhancer regions in breast cancer MCF-7 cells [18]. To identify enhancer regions with ERα binding sites, we further downloaded E2 treated MCF-7 ERα ChIP-seq and matched input data (GSE23893). The raw distribution of read intensity of sample and input data is shown in Fig. 2.14(a). Using ChIP-BIT with parameter setting for TFBS identification at enhancer regions, we identified in total 1,225 TFBSs (probability >0.9) which were located within ±2k bps from the midpoints of 1,168 candidate enhancer regions.

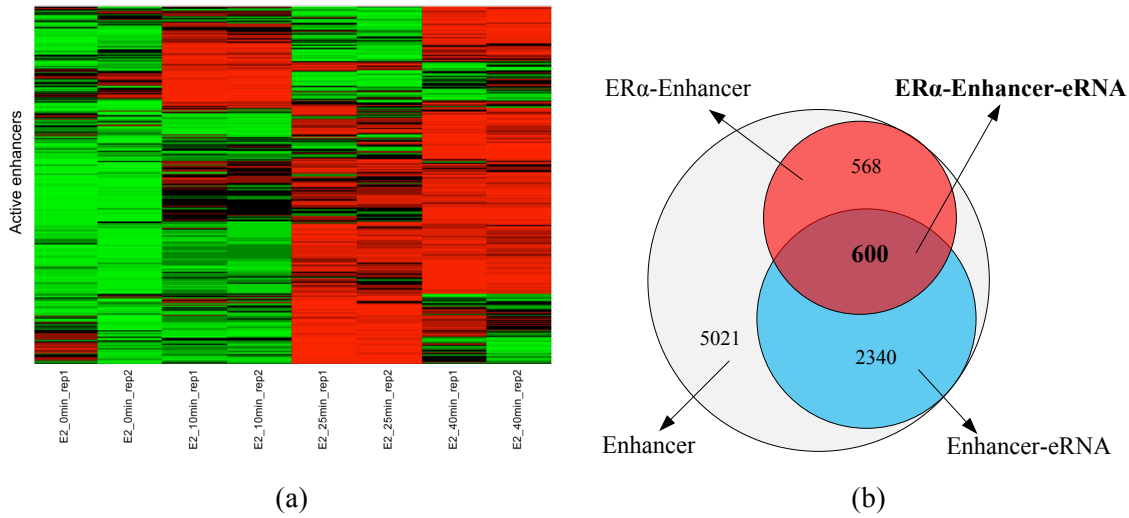


(a) (b)

Figure 2.14 Read intensity distributions of ERα ChIP-seq data. (a) Distributions of read intensities in ER-α and input ChIP-seq data; (b) Distributions of read intensities of ChIP-BIT identified ER-α binding sites in ER-α and input ChIP-seq data.

Validation of enhancer regions is challenging since the mapping relationship between each enhancer region and the target gene is still not clear. The enhancer region activation process is tightly linked to enhancer RNA (eRNA) profiling. eRNA is a specific type of non-coding RNA

measured by the GRO-seq data [75], which is a robust indicator of the functional loop between an activated enhancer region and the interacted promoter region. If an enhancer region is activated by an ERα binding site, it should also have high expression of eRNA. We downloaded a time course GRO-seq data with 8 samples measured at 4 time points (GSE27463). To identify overexpressed eRNA under E2 condition, strand-specific read counts from each GRO-Seq experiment were determined for each eRNA using HOMER. EdgeR [76] was then used to identify differentially expressed eRNAs by setting fold change >1.5-fold and false discovery rate <0.05. In total, we identified 8,921 eRNAs overexpressed under E2 treatment (10 mins, 25 mins or 40 mins), as shown in Fig. 2.15(a). Previously we identified 1,168 enhancer regions with ER-α binding sites, where 600 (significance *p*-value <<0.001) of them contained overexpressed eRNA, as shown in Fig. 2.15(b).



(a) (b)

Figure 2.15 Active enhancer region identification using GRO-seq data. (a) Heat map of time course eRNA expression of active enhancer regions; (b) common regions of candidate enhancer regions, ERα binding sites and overexpressed eRNA transcripts.

2.5 Discussion

In this chapter, a Bayesian approach, ChIP-BIT, was developed and applied to identify TFBSs at promoter and enhancer regions. In contrast to conventional peak detection methods, ChIP-BIT will compute a probability for each peak by joint modelling read intensities in sample and input ChIP-seq data. In addition, binding locations of TFBSs referring to the middle points of promoter regions are also modelled using an exponential distribution. Strong or weak binding

signals are defined according to their read intensities in the sample ChIP-seq data. Weak bindings have an overall lower read intensity in the sample ChIP-seq data but as TF specific binding sites, compared to the low local input signal, their fold changes are still significant. Such weak binding signals can be functional important on gene transcription.

ChIP-BIT is first applied to identify TFBSs at promoter regions. Each identified TFBS can be mapped to a unique target gene. We have validated identified bindings by examining target gene expression change after the knock down of the upstream TF. However, only about 15% of bindings can be successfully validated. A reasonable explanation for those remaining genes without any significant expression change is that, they are simultaneously regulated by multiple TFs so that the knock down of just one TF is not enough to inhibit gene transcription significantly. Running ChIP-BIT on multiple ChIP-seq data we can identify genes co-regulated by multiple TFs. The joint modelling of multiple TFs is very important to infer the complete regulatory map.

ChIP-BIT is then applied to identify TFBSs at distant enhancer regions. As an enhancer region is located far away from its target gene's TSS, enhancer regions are usually studied without an emphasis on target gene association. However, in recent years, predicting target genes of enhancer regions becomes more and more important since a lot of gene expression change cannot be fully explained using proximal binding theory. Although some new data like DNase-seq and ChIA-PET can be used to link an active enhancer region to the target gene, more computational and biological work is still needed to jointly analyze multiple ChIP-seq data, integrate different data types and validate identified genes or enhancer regions.

2.6 Conclusions

In summary, to identify TFBSs at encoding regions, we have developed a probabilistic method called ChIP-BIT. Each TFBS is annotated by a target gene and an enhancer region. And there is a probability reflecting the binding strength. ChIP-BIT is a novel Bayesian approach in that (1) a Gaussian mixture model is developed to help improve weak binding event detection; (2) an exponential distribution is incorporated to model the impact of TFBSs on their target genes. Through simulation studies, we have demonstrated that ChIP-BIT distinguishes foreground bindings especially those weak ones from background signals better than most

available tools. By using ChIP-BIT to identify common targets regulated by both PBX1 and NOTCH3 in breast cancer MCF-7 cells, we have found that there is a significant interaction between PBX1 and NOTCH3. Functional enrichment analysis of target genes co-regulated by PBX1 and NOTCH3 supports the existence of a crosstalk between the Wnt signaling and Notch signaling pathways. By further applying ChIP-seq to ERα ChIP-seq data we have found that active enhancer regions are intensely regulated by ERα in breast cancer MCF-7 cells.

3. BICORN: Bayesian integration of ChIP-seq and RNA-seq data for cis-regulatory module inference

3.1 Introduction

Recent technological advances have improved the molecular understanding of cancers and the identification of targets for therapeutic interventions. The diversity nature of tumors lowers the consistency of multiple biological replicates such that few reliable predictions can be expected if a single type of omics data is used [77]. Previous studies on regulatory mechanisms were mainly focused at the promoter region [78-81]. It has been known that only ~15% target genes predicted from ChIP-seq data of a single TF are differentially expressed (as identified from downstream microarray or RNA-seq data analysis) when the specific TF is knocked down [69]. It indicates that a large proportion of physical bindings are either not functional or not functional individually. Therefore, inferring functional bindings from observed physical bindings not only refers to limiting false positive predictions in ChIP-seq data, but also includes eliminating binding events that are actually inactive on target gene expression. In recent years, enhancer regulation has been widely studied due to the limited explanations that the promoter region focused study can provide for gene regulation [18, 82]. Through previous ChIP-seq data analysis, we have shown that enhancer regions are activated by specific TFs that bring enhancer regions to interact with promoter regions and regulate gene transcription. TF-enhancer-gene regulation can provide a novel and complementary picture to the conventional TF-gene regulation at promoter regions. An enhancer region can work from a distance, in either orientation, and do not necessarily regulate the closest gene. Therefore, for TFs working at enhancer regions, how to infer their target genes is a challenging task.

Early attempts for functional binding inference only used gene expression data [83, 84] because at that time prior binding knowledge was largely not available. With the accumulation of ChIP-chip data [85], many integrative approaches were proposed and developed. Liao *et al*. proposed network component analysis (NCA) to estimate activities of multiple TFs jointly by integrating binary binding information (from ChIP-chip data) and gene expression (from microarray data) [86]. Sabatti *et al.* improved NCA with a Bayesian framework to simultaneously infer hidden TF activities and functional bindings [78]. Chen *et al.* employed a

similar Bayesian hierarchical model called COGRIM to infer regulatory gene clusters [79]. Recently, large-scale ChIP-seq data generation made CRM inference progress much faster. Wang *et al.* developed a BETA package for functional target gene prediction by integrating single TF ChIP-seq data with target gene RNA-seq data [80]. Qin *et al.* developed an integrative approach, least absolute shrinkage and selection operator (LASSO), to infer regulatory networks of multiple TFs [81]. The importance of statistical integration of ChIP-seq and RNA-seq data was discussed in [41] for inferring gene regulatory mechanisms; Bayesian integration models were proposed as optimal approaches for causal inference of genome-wide regulatory networks. However, in existing Bayesian integration methods (e.g. BNCA and COGRIM), the functional evaluation of each TF-gene binding is based on an individual comparison between gene expression fitting performances under the existence of current binding or not. Contributions of other TFs to the occurrence of current binding event are not modelled.

In the real system, several specific TFs serve as major regulators (or activators) and recruit other co-factors (or so called partners) to achieve gene regulation. Binding signals of different TFs are enriched at the same location, which is referred to as a cis-regulatory module (CRM) [87]. TF co-binding or co-regulation has recently also been observed for the enhancer region activation [88]. To study the cooperation of multiple transcription factors on the gene regulation, we propose a novel Bayesian method (BICORN) to integrate ChIP-seq binding signals of multiple TFs and RNA-seq target gene expression and infer CRMs at encoding (promoter or enhancer) regions. In BICORN, each CRM has a unique combination of TFs and each of its target gene expression is modeled as a log-linear combination of the hidden TF activities (TFAs). We model each TF-gene binding event as a Bernoulli random variable (0 for 'none functional or background binding' or 1 for 'functional'). Then, each CRM can be defined by a vector containing a unique combination of multiple related Bernoulli random variables. A Gibbs sampling algorithm iterates between the estimation of hidden TFAs and that of posterior distribution of CRMs (also a joint distribution of multiple TFs). Notably, joint regulation of TFs in individual CRM on target genes are learned from the integrative analysis of binding network and target gene expression so that CRM-gene interactions can be directly predicted.

We demonstrate the robustness of BICORN using simulated data with respect to different noise levels. Then, BICORN is further tested on DREAM4 *in silico* networks with time course

gene expression data. The results show a significant precision-recall improvement over existing Bayesian integrative methods or gene expression-based approaches. For real data analysis, we apply BICORN to breast cancer MCF-7 cell ChIP-seq and RNA-seq data to infer CRMs at promoter and enhancer regions. Our integrative analysis shows a group of TFs is directly functional at proximal promoter regions while the other group of TFs distantly regulates the same set of genes through a possible enhancer-promoter looping mechanism. In each TF group, we identify several CRMs as well as their major regulators or activators. The breast cancer study highlights that it is necessary and important to investigate TF regulatory effects on both enhancer and promoter regions. Only in that way we can obtain a complete regulatory map for cancer-specific genes.

3.2 Methods

3.2.1 BICORN model description

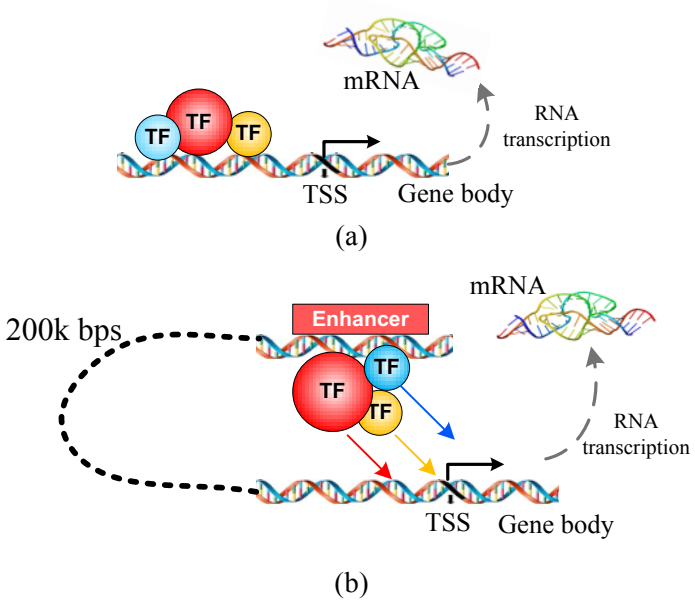


Figure 3.1 Illustration of regulation of cis-regulatory modules. (a) Cis-regulatory modules at promoter regions; (b) cis-regulatory modules at enhancer regions.

Identifying initial TF-gene interactions at promoter regions is straightforward. As shown in Fig. 3.1(a), we map each TFBS to the nearest transcription starting site (TSS) if their distance is within certain range, e.g., 10k bps. In order to identify functional target genes regulated by distal regulatory elements (shown in Fig. 3.1(b)), we associate 20 nearby genes (10 genes upstream and

10 genes downstream) with each enhancer region (enriched by histone markers H3K4me1 and H3K27ac). Although we realize that such a simple mapping approach cannot identify target genes that are farther than 10 genes away, we anticipate that many of enhancer regions would regulate a gene within this distance based on existing studies [19, 89]. We then build an initial physical binding network from multiple TFs to a number of target genes for CRM inference at promoter or enhancer regions.

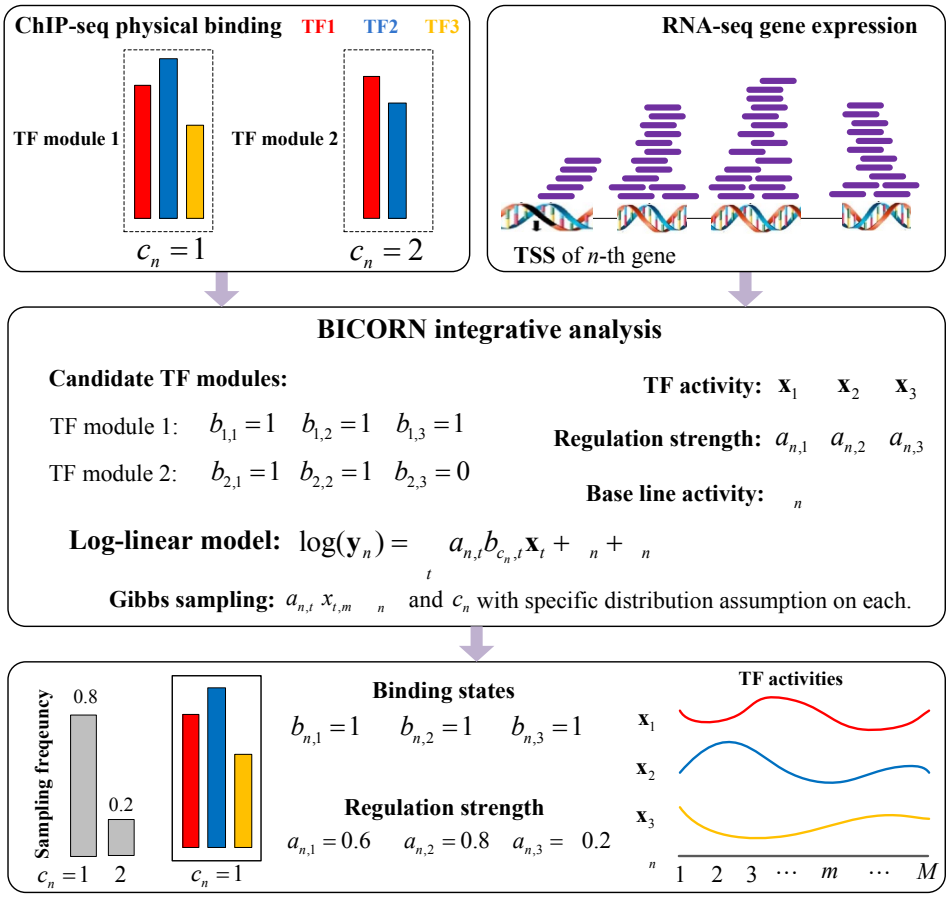


Figure 3.2 A flowchart of the proposed BICORN model.

Given the initial binding network and target gene expression, BICORN infers the most reliable CRM for each gene, as shown in Fig. 3.2. Specifically, each CRM is represented by a binary vector with a unique combination of TFs. Each gene can be regulated by different CRMs (e.g., c_1 and c_2), depending on observed binding signals at its encoding region. Using Bayesian integration, we jointly model ChIP-seq binding signals and RNA-seq gene expression such that

functional CRMs can be inferred. Specifically, the regulation relationship from multiple TFs in a CRM to individual target gene expression is modelled as a log-linear combination of TFAs [86]. The distribution of CRMs is a joint distribution of multiple TFs, which is difficult to model using a particular distribution shape. Using a Gibbs sampling algorithm we iteratively estimate the CRM distribution and hidden variables. After multiple runs of sampling, for each gene, we then select the most reliable CRM as the final prediction (e.g. c_1). Marginal distribution of individual TFs can also be calculated by summing up the probabilities of CRMs containing each particular TF, which indicates the major or partner role of each TF. Finally, genes under regulation of each CRM are clustered together and the number of genes is used to denote the importance of each CRM in the whole regulatory system.

3.2.2 Log-linear model and Gibbs sampling

Based on biological prior knowledge, we identify a candidate CRM pool, denoted by a matrix **B** including K rows and T columns. Each row represents a candidate CRM and each column represents a particular TF. Since TFs can only be clustered if they have potential biological interactions, we have $K << 2^T$. In the k -th row, a unique vector $[b_{k,1}, b_{k,2}, ..., b_{k,t}, ..., b_{k,T}]$ with binary states $b_{k,t}$ as '0' or '1' is defined according to prior biological knowledge. In real biological systems, there are always some 'background' genes without functional regulation. Hence, there is an all-zero vector $[b_{0,1}, b_{0,2}, ..., b_{0,t}, ..., b_{0,T}]$ in **B**.

In total we have N target genes with RNA-seq gene expression data under M time points or conditions; we have T TFs with baseline ChIP-seq data. We assume that n -th gene is regulated by the c_n -th CRM under M time points or conditions. c_n takes a value from $0 \sim K$ and represents which module (a row of **B**) is functional on n -th gene. Therefore, binary binding states $\left\{ b_{c_n,t} \mid t = 1 \sim T \right\}$ of n -th gene are jointly determined by unknown c_n and known **B**. To infer c_n, RNA-seq expression of n -th target, $\mathbf{y}_n$, is modelled using a log-linear model as defined in the following equation:

$$\log(\mathbf{y}_n) = \sum_t a_{c_n,t} b_{c_n,t} \mathbf{x}_t + \quad _n + \boldsymbol{\varepsilon}_n, \tag{3-1}$$

where vector $\mathbf{y}_n = [y_{n,1}, y_{n,2}, ..., y_{n,m}, ..., y_{n,M}]$ represents gene expression data observed at M time points or conditions with sample index m; vector $\mathbf{x}_t = [x_{t,1}, x_{t,2}, ..., x_{t,m}, ..., x_{t,M}]$ represents the hidden TFAs of t-th TF; variable $a_{c_n,t}$ represents the regulation strength of the binary binding variable $b_{c_n,t}$; variable $_n$ represents the baseline gene expression level across multiple time points; and vector $\boldsymbol{\varepsilon}_n = [\ _{n,1},\ _{n,2}, ...,\ _{n,m}, ...,\ _{n,M}]$ denotes the noise in gene expression data. Variables $x_{t,m}$, $a_{c_n,t}$, $_n$ and $_{n,m}$ are all unknown and needed to be estimated from data.

We make some hypotheses on hidden viable distributions. Specifically, we assume that $\mathbf{x}_t$ follows Gaussian random process and the prior distribution on each variable $x_{t,m}$ is defined as $N(0,\ _x^2)$. The regulation strength $a_{c_n,t}$ is conditional on the state of $b_{c_n,t}$. For $b_{c_n,t} = 0$, the regulation strength does not exist and $a_{c_n,t} = 0$. For $b_{c_n,t} = 1$, we assume a prior Gaussian distribution on $a_{c_n,t}$ as $N(0,\ _a^2)$. The baseline expression $_n$ is a constant value under multiple time points. The prior distribution on $_n$ is also assumed to be a Gaussian distribution as $N(0,\ ^2)$. Gene expression data measurement noise $_{n,m}$ is assumed to follow zero-mean Gaussian distribution as $N(0,\ ^2)$. To limit the scale of ^{2}, we assume that ^{2} follows inverse Gamma distribution as *inverseGamma*(,). Here, $_x^2$, $_a^2$, ^{2}, and are hyper-parameters.

The prior probability of binding event $b_{c_n,t} = 1$ is denoted as $_{n,t}$, which is estimated from the ChIP-seq data using ChIP-BIT. Then, a prior distribution for each binding variable $b_{c_n,t}$ can be defined in the following equation:

$$P(b_{c_n,t}) = \ _{n,t}^{\ b_{c_n,t}}(1 \quad _{n,t})^{1\ b_{c_n,t}}. \tag{3-2}$$

Furthermore, the prior probability of CRM c_n is defined as:

$$P(c_n) = \ _{t}\ _{n,t}^{\ b_{c_n,t}}(1 \quad _{n,t})^{1\ b_{c_n,t}}. \tag{3-3}$$

We define $\mathbf{Y}=\left\{y_{n,m}\,|\,n=1\sim N,\ m=1\sim M\right\}$, $\mathbf{A}=\left\{a_{c_n,t}\,|\,n=1\sim N,\ t=1\sim T\right\}$, $\mathbf{X}=\left\{x_{t,m}\,|\,t=1\sim T,\ m=1\sim M\right\}$ and $\boldsymbol{\eta}=\left\{\ _n\,|\,n=1\sim N\right\}$. To determine the functional regulatory network controlling the expression of genes under a certain condition, we infer CRM indexes for all genes as $\mathbf{C}=[c_1,c_2,...,c_n,...,c_N]^T$. To estimate variables in $\mathbf{A}$, $\mathbf{X}$, $\boldsymbol{\eta}$ and $\mathbf{C}$, a joint posterior probability is defined in the following equation:

$$P\left(\mathbf{A},\mathbf{C},\mathbf{X},\boldsymbol{\eta},\ ^2\,|\mathbf{Y},\mathbf{B}\right) \quad \prod_n P\left(\mathbf{y}_n\,|\mathbf{a}_n,c_n,\mathbf{X},\ _n,\ ^2,\mathbf{B}\right) \cdot P\left(\mathbf{a}_n\right)\cdot P\left(c_n\right)\cdot P\left(\mathbf{X}\right)\cdot P\left(\ _n\right)\cdot P\left(\ ^2\right). \tag{3-4}$$

Given the structure of Eq. (3-4), there is some conditional independence structure within each of these variables. To learn the distribution of each, we calculate the conditional probability of each variable and draw samples iteratively using a Gibbs sampler. A detailed derivation of the algorithm can be found in the **Appendix C**. Here we mainly describe the conditional distributions used in the Gibbs sampling algorithm:

Gibbs sampling of regulation strength

For each gene, if $b_{c_n,t}=0$, we sample $a_{c_n,t}=0$; if $b_{c_n,t}=1$, we sample $a_{c_n,t}$ according to its conditional probability as defined in the following equation:

$$P\left(a_{c_n,t}\,|b_{c_n,t}=1,\mathbf{y}_n,\mathbf{X},\ _n,a_{c_n,t'\ t},\ ^2\right) \quad P\left(\mathbf{y}_n\,|\,\mathbf{X},\ _n,a_{c_n,t},\ ^2\right)P\left(a_{c_n,t}\right) \quad \prod_m \frac{1}{}\exp\ \ \frac{1}{2\ ^2}\left(\ _{n,m}\right)^2\ \ \frac{1}{_a}\exp\ \ \frac{1}{2\ ^2_a}a^2_{c_n,t} \quad N\left(\ _a,\ ^2_a\right). \tag{3-5}$$

Gibbs sampling of TFA

We sample $x_{t,m}$, TFA of *t*-th TF under *m*-th time point, according to its conditional probability as defined in the following equation:

$$P\left(x_{t,m}\middle|\mathbf{Y},\mathbf{A},\mathbf{C},x_{t',t,m},\boldsymbol{\eta},\ ^2\right)\quad P\left(\mathbf{y}_m\middle|\mathbf{A},\mathbf{C},\mathbf{X},\boldsymbol{\eta},\ ^2\right)P\left(x_{t,m}\right)$$

$$_n\ \frac{1}{}\exp\quad\frac{1}{2\ ^2}\left(\ _{n,m}\right)^2\ \frac{1}{_x}\exp\quad\frac{1}{2\ ^2_x}x_{t,m}^2 \qquad (3\text{-}6)$$

$$N\left(\ _x,\ ^2_x\right).$$

Gibbs sampling of baseline gene expression

We sample $_n$, baseline expression of *n*-th gene, according to its conditional probability as defined in the following equation:

$$P\left(\ _n\middle|\mathbf{y}_n,\mathbf{X},c_n,\mathbf{a}_n,\ ^2\right)\quad P\left(\mathbf{y}_n\middle|\mathbf{X},c_n,\mathbf{a}_n,\ _n,\ ^2\right)P\left(\ _n\right)$$

$$_m\ \frac{1}{}\exp\quad\frac{1}{2\ ^2}\left(\ _{n,m}\right)^2\ \frac{1}{}\exp\quad\frac{1}{2\ ^2\ ^2_n} \qquad (3\text{-}7)$$

$$N\left(\quad,\ ^2\right).$$

Gibbs sampling of expression data noise variance

^{2} controls the overall fitting performance and can be sampled as follows:

$$P\left(\ ^2\middle|\mathbf{Y},\mathbf{A},\mathbf{C},\boldsymbol{\eta},\mathbf{X}\right)\quad P(\mathbf{Y}\middle|\mathbf{A},\mathbf{C},\boldsymbol{\eta},\mathbf{X},\ ^2)P\left(\ ^2\right)$$

$$_n\ _m\left(\ ^1\right)\exp\quad\frac{1}{2\ ^2}\left(\ _{n,m}\right)^2\cdot\left(\ ^2\right)\ ^1\exp\quad\frac{}{\ ^2} \qquad (3\text{-}8)$$

$$\text{inverse}\,\text{Gamma}\left(\ ',\ '\right).$$

Gibbs sampling of CRM

For CRM distribution learning, there is no particular distribution that can be used to model it. For *n*-th gene, we calculate the conditional probability given each CRM with c_n from 0 to *K* as follows:

$$P\left(c_n\middle|\mathbf{y}_n,\mathbf{X},\ _n,\mathbf{a}_n,\ ^2\right)\qquad _t\ P\left(\mathbf{y}_n,\mathbf{x},\ _n,\ ^2\middle|a_{c_n,t}\right)P\left(a_{c_n,t}\middle|c_n\right)P\left(c_n\right). \qquad (3\text{-}9)$$

Then, we draw a sample of c_n as c'_n according to its posterior probability distribution as defined in the following equation:

$$p(c_n = k) = \frac{P\left(c_n = k \middle| \mathbf{y}_n, \mathbf{X}, \ _n, \mathbf{a}_n, \ ^2\right)}{\sum_j P\left(c_n = j \middle| \mathbf{y}_n, \mathbf{X}, \ _n, \mathbf{a}_n, \ ^2\right)}. \tag{3-10}$$

3.3 Simulation

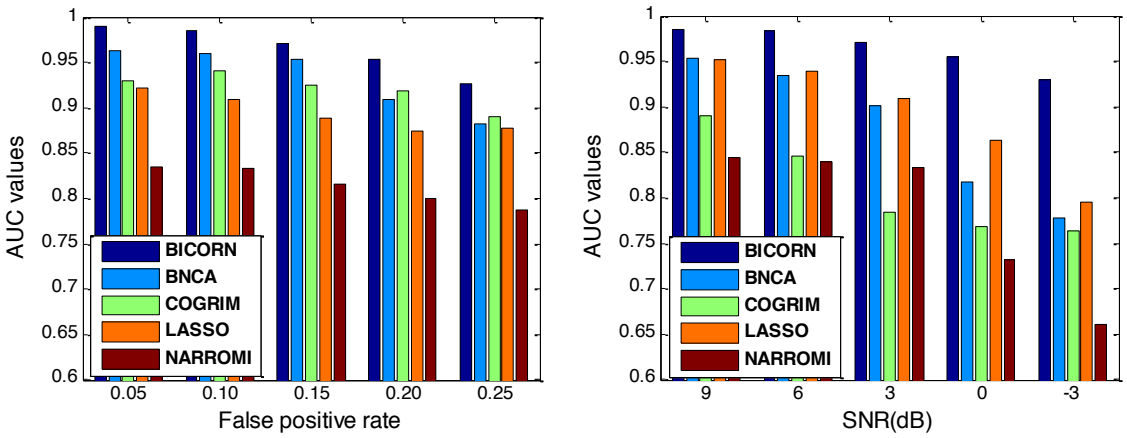
To demonstrate the advantage of BICORN over conventional methods on CRM inference, we simulated binding networks and gene expression data under different scenarios. First of all, we generated a regulatory network with 160 genes and 20 TFs. There were 80 foreground genes and on average 2 bindings of each gene. The regulation strength of each binding was generated as following Gaussian distribution. TFAs for individual TFs were simulated using Gaussian random process with mean and variance values as 0 and 1. Then, we simulated two different gene expression datasets using log-linear model as in Eq. (3-1). In Case 1, we generated 20 gene expression samples; in Case 2, we generated only 10 gene expression samples. Case 2 was more challenging since the number of TFs (20) is much larger than that of gene expression samples (10). To compare BICORN against conventional approaches on functional TF-gene binding prediction, we simulated different noise scenarios by varying the false positive rate of initial binding connections (from 0.05 to 0.25 with step 0.05) and the signal-to-noise ratio (SNR) of gene expression data (from 9 to -3 dB with step 3 dB).

Since conventional methods can only infer individual functional bindings rather than the whole module, here we want to show even on the functional binding prediction, BICORN can still provide advanced performance. We define sensitivity and specificity as follows:

$$\text{sensitivity} = \frac{\text{Number of true positive bindings}}{\text{Number of true positive bindings} + \text{Number of false negative bindings}},$$

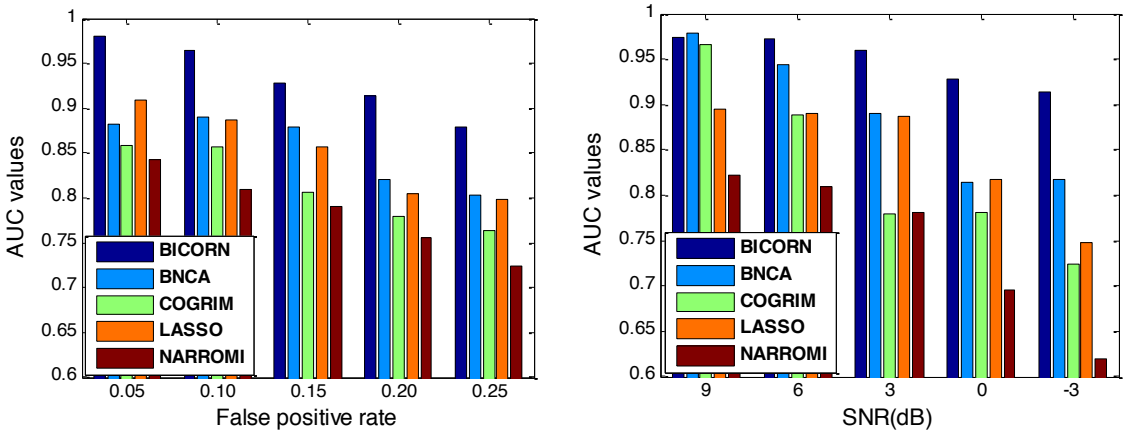
$$\text{sensitivity} = \frac{\text{Number of true negative bindings}}{\text{Number of true negative bindings} + \text{Number of false positive bindings}}.$$

Area under curve (AUC) values of receiver operating characteristic (ROC) of BICORN and competing methods are shown in Fig. 3.3 for performances in Case 1 and Fig. 3.4 for performances in Case 2.



(a) (b)

Figure 3.3 Binding network prediction performance of competing methods for Case 1. (a) Initial binding networks with different false positive rates; (b) gene expression data with different (Signal-to-Noise Ratio) SNR.



(a) (b)

Figure 3.4 Binding network prediction performance of competing methods for Case 2. (a) Initial binding networks with different false positive rates; (b) gene expression data with different (Signal-to-Noise Ratio) SNR.

It can be clearly seen from Fig. 3.3(a) that in Case 1 when the total number of expression samples is larger than the number of candidate TFs, BICORN works more robustly against false

positive connections in the initial network. In Fig. 3.3(b) when the noise of gene expression increases, performances of conventional methods degrade dramatically. The performance improvement of BICORN is more significant when the only observation of the whole system, gene expression data is noisy. In Case 2, due to the fewer number of expression data samples, estimation of TFAs will be very challenging. The overfitting issue cannot be overlooked since the number of candidate TFs is larger than expression data observations. As can be seen from Fig. 3.4, in both figures BICORN is much better than competing methods. In conventional Bayesian tools, for each gene they examine fitting performance of each TF iteratively; not enough expression data samples may bring ambiguity to determine which TF or TFs provide the best performance.

Performance evaluation in Fig. 3.3 and Fig. 3.4 was based on individual binding connection. It is necessary to compare the performances of competing methods on identifying the whole CRM. If only partial bindings of a module are predicted, the module or the combination relationship of multiple TFs will not be fully recovered. In that case, the module prediction is false. We changed the evaluation criterion to "*all functional bindings of a gene should be simultaneously identified correctly*" and summarized precision-recall performances of three Bayesian integration tools including the proposed BICORN, BNCA and COGRIM in Table 3.1. We define precision and recall as follows:

$$\text{precision} = \frac{\text{Number of true positive modules}}{\text{Number of true positive modules} + \text{Number of false positive modules}},$$

$$\text{recall} = \frac{\text{Number of true positive modules}}{\text{Number of true positive modules} + \text{Number of false negative modules}}.$$

Table 3.1 CRM identification with full recovery of all functional bindings.

METHOD	BICORN	BNCA	COGRIM
Precision	0.768	0.730	0.725
Recall	**0.663**	0.100	0.090

It can be seen from Table 1 that with similar precision performances, BICORN has identified more genes by fully recovering their upstream regulators. BNCA and COGRIM only recover complete bindings or the full modules of a small portion (~10%) of genes.

We also realized in some cases large scale ChIP-seq data experiments were not available and people had to use old type of binding information containing a lot of false positive bindings. In that case, people only cared about whether some genes with at least one functional binding can be captured. Those genes would be used in further downstream analysis. COGRIM was initially proposed to identify target genes based on the noisy initial TF-DNA binding network. We have further simulated more challenging scenarios where the physical binding network was very noisy (high false positive rate) including a lot of background genes. We mainly compared the performance of BICORN against BNCA and COGRIM. In each method, a foreground gene is called if there is at least one binding with sampling frequency over default threshold. We defined the sensitivity and specificity as follows:

$$\text{sensitivity} = \frac{\text{Number of true positive genes}}{\text{Number of true positive genes} + \text{Number of false negative genes}},$$

$$\text{sensitivity} = \frac{\text{Number of true negative genes}}{\text{Number of true negative genes} + \text{Number of false positive genes}}.$$

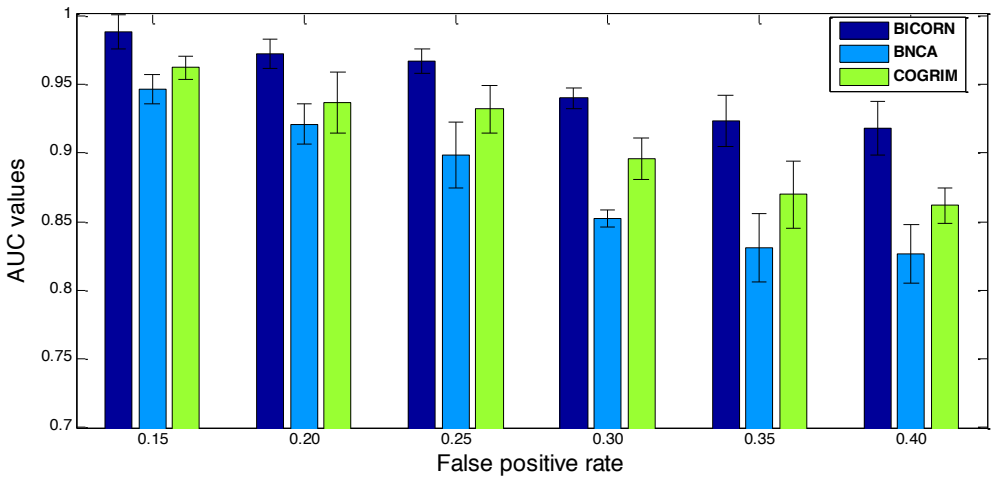


Figure 3.5 AUC performances of competing methods on target gene prediction using severely contaminated physical binding networks.

AUC values of ROC curves for foreground/background target prediction were shown in Fig. 3.5. It can be found that BICORN is quite robust to identify foreground genes even though the false connection rate in network increases. This robust performance is mainly due to the algorithm design of modelling TF modules directly. Because a background gene is not 'truly' regulated by any TFs, in the BICORN learned joint distribution of multiple TFs, there is not any combination of TFs showing a high enough probability. Hence, BICORN reports none binding for the background gene.

3.4 *in silico* network validation

In most of the tools for regulatory network inference, a log-liner model (as defined in Eq. (3-1)) is used to integrate TF-gene bindings and gene expression data. However, the log-liner model is an approximation, which may not hold in the real biological system. It is necessary to evaluate the performance of each method when gene expression data have a non-linear relationship with TF activities. We downloaded benchmark networks and training time course gene expression data from DREAM challenge 4 (http://gnw.sourceforge.net/dreamchallenge.html). There were 5 different networks simulated from 100 genes and ~15 TFs. On average, each TF regulated 11 genes and each gene had 2 regulators. For each benchmark network, 10 different time-course gene expression datasets with 21 time points in each were generated using Stochastic Differential Equations, which was a non-linear simulation of gene expression data based on the benchmark network. To evaluate the performance of BICORN and competing methods, we perturbed benchmark networks by adding 15% false negative connections and 30% false positive connections. Precision-recall performance on binding connection identification of each method under default setting is recorded. In total, we had 50 case studies (10 simulated gene expression datasets for each of 5 benchmark networks). A box plot of 50 F-measures (calculate by 2/(1/precision+1/recall)) of each competing method is presented in Fig. 3.6.

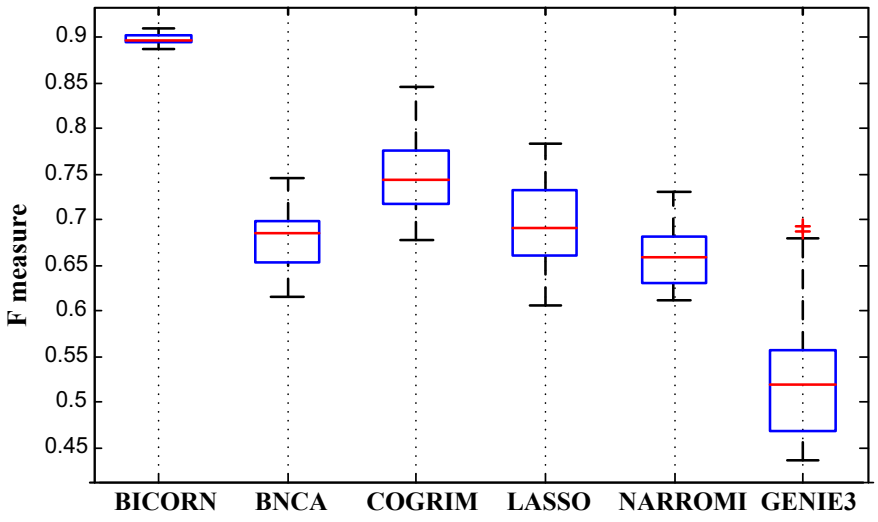


Figure 3.6 F-measure comparison of competing tools using DREAM 4 *in silico* benchmark regulatory networks and simulated time course gene expression data.

From Fig. 3.6 it can be seen that BICORN provides the best F-measure performance. The mean of its F-measures of in total 50 cases is the highest and the variance is the smallest. When the real network is very sparse and the log-linear relationship between gene expression and TF activities does not hold, methods using gene expression data only like NARPOMI or GENIE3 cannot provide reliable predictions. It can also be found that conventional Bayesian integration tools do not provide improved performance to the other methods. The advantage of BICORN is mainly due to the shrunk search space of candidate CRMs. Through an initial screening of possible CRMs on all candidate genes, any TF combinations never observed on any gene would not be selected. In comparison to the other Bayesian approaches, this initial CRM searching step narrowed the search space for functional CRM prediction and lowered the chance of BICORN on sampling false bindings.

Using this benchmark data we have further examined the efficiency of each method. We tested BICORN, LASSO, NARROMI, GENIE3 using MATLAB2012b and BNCA, COGRIM using R 2.15 on a desktop computer with 4-core CPU 3.20 GHz and RAM 6.0 GB. The average running time of key functions of each method was shown in Table 3.2. In general, the computational cost of Bayesian methods is higher than LASSO or the other regression based approaches. But among all three Bayesian approaches, BICORN has the shortest running time.

Table 3.2 Average computational time for each competing method.

METHOD	BICORN	BNCA	COGRIM	LASSO	NARROMI	GENIE3
Average running time (sec.)	**186.7**	198.2	493.6	0.3	10.3	73.7

3.5 Breast cancer ChIP-seq and RNA-seq integrative analysis

We finally applied BICORN to real ChIP-seq and RNA-seq data acquired from breast cancer MCF-7 cells. We downloaded ChIP-seq data of 39 TFs from ENCODE (https://genome.ucsc.edu/ENCODE/). Those ChIP-seq data samples were generated from baseline MCF-7 cells. To infer their regulatory effects on gene expression when MCF-7 cells are under fast progression, we downloaded two 17b-E2 treated MCF-7 RNA-seq data sets from the GEO database (accession numbers GSE62789 and GSE51403). GSE62789 data include 10 RNA-seq samples measured at baseline ('0' time point) and 9 different time points within 24hrs of 10nM 17b-E2 treatment. A candidate target gene was selected if at least at one time point its fold change to the '0' time point is larger than 1.3 (0.3 in its log2 format). GSE51403 data include 7 RNA-seq samples measured under vehicle condition and another 7 steady-state RNA-seq samples measured after 24hrs of 10nM 17b-E2 treatment. A candidate target gene was selected if its false discovery rate (FDR) is smaller than 10% after DeSeq2 [90] differential expression analysis. Finally, we identified 2,768 and 1,158 E2 up-regulated genes respectively from time course and steady state RNA-seq datasets. 275 common target genes were selected as candidate genes for regulatory mechanism study of 39 TFs in E2 treated breast cancer MCF-7 cells. BICORN CRM inference was carried out at promoter and enhancer regions separately since their regulatory mechanisms are very different.

3.5.1 Proximal CRM inference using breast cancer MCF-7 data

We identified TFBSs of 39 TFs at 25,802 promoter regions (annotated by UCSC RefSeq hg19) using ChIP-BIT. Each TFBS was uniquely associated with the nearest target gene. Based on the overall binding similarities among TFs, we selected 105 candidate CRMs including 32 TFs. TF symbols are summarized in Fig. 3.7(a). We used BICORN to integrate physical bindings with time course gene expression and steady state gene expression respectively for CRM

inference. We ran 1000 rounds of Gibbs sampling to iteratively estimate TFAs, regulation strengths and distributions of CRMs.

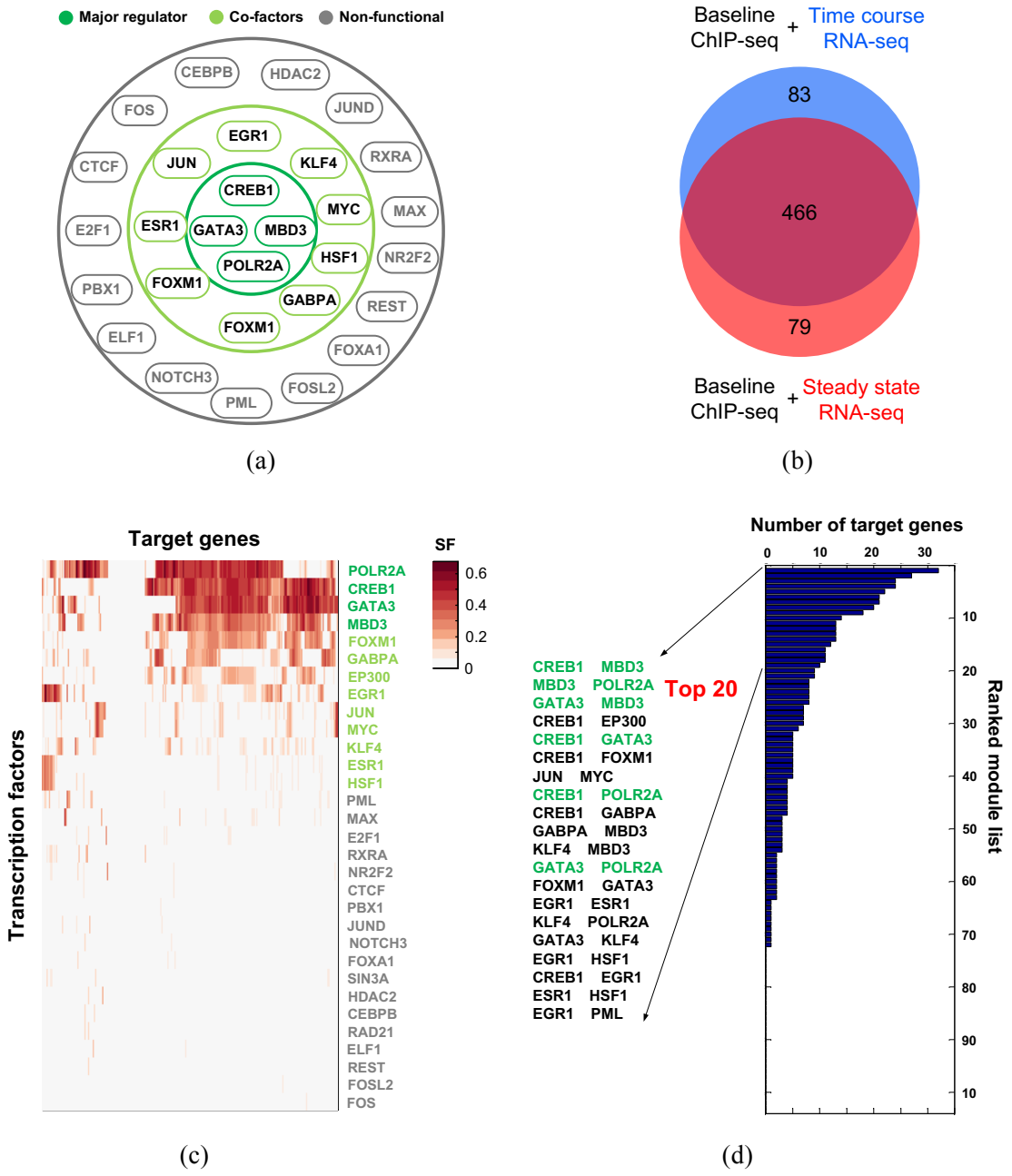


Figure 3.7 BICORN integrative analyses of 32 TFs at gene promoter regions. (a) Candidate TF symbols; (b) the similarity of BICORN inferred functional bindings from two different gene expression data sets; (c) a regulatory map of BICORN inferred TF-gene interactions (the color represents sampling frequency (SF)); (d) a CRM rank list sorted by the number of target genes regulated by each.

By setting sampling frequency threshold as quantile 0.85, for each gene, we predicted a most reliable CRM. In total we identified 549 CRM-gene interactions from time-course gene expression data and 545 CRM-gene interactions from steady-state gene expression. There were 466 common CRM-gene interactions, as shown in Fig. 3.7(b). The identified common TF-gene regulatory map included 32 TFs and 113 target genes, as shown in Fig. 3.7(c). Based on the number of target genes regulated by each module, 105 candidate CRMs were sorted and further presented in Fig. 3.7(d), where we only presented specific TF symbols in the top 20 CRMs. It can be seen from the TF-gene regulatory map in Fig. 3.7(c) that POLR2A, CREB1, GATA3, and MBD3 had very dense bindings and they served as major regulators. Another set of TFs from FOXM1 to HSF1 regulated diverse target genes, whose binding connections were sparse but the sampling frequency of each TF-gene interaction was still high. The remaining TFs were much less functional because each of them had few target genes with high sampling frequencies. Comparing TF symbols in top ranked modules in Fig. 3.7(d) to the ranked TF list in Fig. 3.7(c) we found that major TFs were clustered together with different weak factors and some strong-weak combination (labeled as black) were ranked higher than certain strong-strong combinations (labeled as green). This observation supports our hypothesis that the regulatory effects of weak binding signals are still significant on target gene transcription. Conventional tools can only provide a regulatory map similar to Fig. 3.7(c). There is no indicator showing that how TFs collaborate with each other on target gene regulation.

3.5.2 Distant CRM inference using E2-treated breast cancer MCF-7 data

To infer CRMs functional at enhancer regions, we first identified 10,434 candidate enhancer regions using histone modification ChIP-seq data by following the procedure in [91]. Then, we identified TFBSs of 39 TFs at enhancer regions using ChIP-BIT. Based on the overall binding similarities among TFs we identified 59 candidate CRMs including 22 TFs. TF symbols are shown in Fig. 3.8(a). For each enhancer region, we associated it with 20 target genes (10 upstream gene and 10 downstream genes). We found that although the number of TFs under investigation (22 TFS) was smaller than that of the promoter region focused study (32 TFs), the density of current regulatory map was much higher due to the multiple-multiple mapping relationship between enhancer regions and target genes.

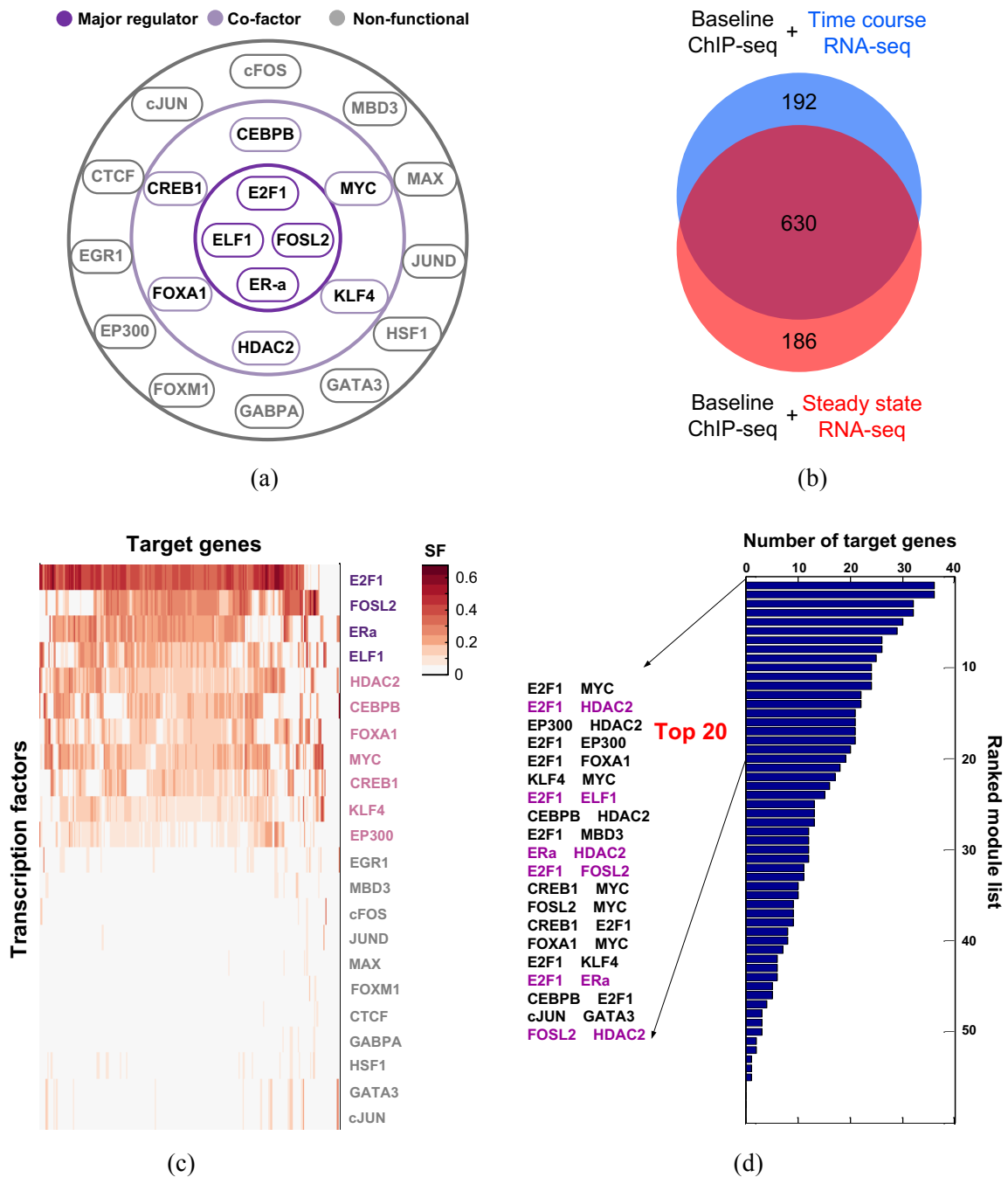


Figure 3.8 BICORN integrative analyses of 22 TFs at enhancer regions. (a) Candidate TF symbols; (b) the similarity of BICORN inferred functional bindings from two different gene expression data sets; (c) a regulatory map of TF-gene interactions (the color represents sampling frequency (SF)); (d) a CRM rank list sorted by the number of target genes regulated by each.

BICORN was then used to integrate observed physical bindings with time-course and steady-state gene expression data respectively for CRM inference. After 1000 runs of Gibbs sampling, setting sampling frequency threshold as quantile 0.85, for each gene, we predicted the

most reliable CRM which was functional at a matched enhancer region. In total we identified 822 CRM-gene interactions from time-course gene expression and 816 CRM-gene interactions from state-gene expression. There were 630 common CRM-gene interactions covering 22TFs and 99 target genes, as shown in Fig. 3.8(b) and Fig. 3.8(c). According to the number of target genes regulated by each module, 59 candidate CRMs were sorted and then presented in Fig. 3.8(d), where we only presented TF symbols in the top 20 CRMs. As can be seen from Fig. 3.8(c), TFs from E2F1 to EP300 were more functional. Although there was no clear separation of 'strong' or 'weak' TFs, comparing TF symbols in the top 20 CRMs in Fig. 3.8(d) to the TF list in Fig. 3.8(c) we could find that a higher ranked TF in Fig. 3.8(c) was more frequently combined with a lower ranked TF. For example, E2F1 was clustered together with MYC. E2F1 was a top ranked TF in Fig. 3.8(c) while MYC was ranked much lower. But their combination was ranked as the first in CRM analysis. It can also be found from Fig. 3.8(d) that pairwise combinations of the top 5 TFs with very dense bindings in Fig. 3.8(c) were not highly ranked in the top 20 CRMs (labeled as purple). The strong-weak TF combinations in the inferred CRMs further support our hypothesis that the regulatory effects of weak bindings are important in the activation of enhancer regions.

Comparing above two CRM studies at promoter and enhancer regions we found that active TFs in each study were very different, as illustrated in **Appendix D**. There were 6 TFs including E2F1, FOSL2, ELF1, HDAC2, CEBPB and FOXA1 highly active at enhancer regions, as shown in Fig. D.1(a). However, they almost had no strong functional bindings at promoter regions, as can be seen from in Fig. D.1(b). Conversely there were 7 TFs including HSF1, cJUN, EGR1, GABPA, FOXM1, MBD3 and GATA3 with intensive bindings at gene promoter regions. Their bindings at enhancer regions were almost non-functional. There were only 4 TFs functionally working at both regions. These two complimentary figures together provided a complete picture of the regulatory effect of transcription factors on gene transcription.

3.6 Discussion

A novel computational tool, namely BICORN, is proposed and developed for CRM inference by integrating protein-DNA binding and target gene expression. BICORN makes it feasible to analyze a large number of TFs simultaneously and identify potential TF clusters as

CRMs and target genes for each. As a unique feature of BICORN, each CRM (or a combination of TFs) is directly modelled and learned from data. Through simulation study, we have demonstrated that modelling TF modules than individual TFs can provide improved robustness against false positive connections in network and noise in gene expression data. In practical applications, BICORN can not only provide solutions to conventional regulatory network inference at gene promoter regions, but also predicts target genes for TF modules working at enhancer regions. We used a breast cancer focused study to demonstrate that it is necessary to study both enhancer and promoter regions for transcriptional regulation research. Regulatory maps observed from each type of region may be biased because only partial TFs are active at each type of region.

In recent years, more data types appear like DNA-methylation array measuring methylation site, DNase-seq measuring DNA segments with enrichment of general binding signals, CLIP-seq measuring RNAs bound by RNA binding protein, ChIA-PET measuring enhancer-promoter interactions connected by a specific TF, etc. Although BICORN only integrates ChIP-seq and RNA-seq data, it can be easily extended under a Bayesian framework to incorporate more regulation information. DNA Methylation plays an important role for epigenetic gene regulation [92]. Hyper-methylation will greatly repress gene transcription. Therefore, observed methylation signal at promoter or enhancer regions may be used as a prior to represent the possibility that TFs can regulate the transcription process of current gene. Specific for the enhancer region analysis, ChIA-PET provides an experimental measurement for enhancer-gene loop identification [93]. In addition, ROADMAP project (http://www.roadmapepigenomics.org/) provides a public resource of human epigenetic data, including DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts. All these information can be integrated with TF ChIP-seq data and RNA-seq gene expression data by extending the proposed Bayesian framework for a large scale study of transcriptional regulation.

As indicated in the results of breast cancer MCF-7 cells study, TFs functional at promoter or enhancer regions are very different. Spatial interactions of TFs across two types of regions should be further investigated. Although ChIA-PET technology [94] can be used to experimentally measure enhancer-promoter loops, it is a TF specific technology and can only

provide enhancer-promoter loops bound by the same TF at both regions. As shown in active TF comparison at promoter and enhancer regions, only a few TFs work on both types. Therefore, computational efforts are needed for spatial interaction prediction of different TFs forming or maintaining long-range enhancer-promoter loops.

Our study is mainly focused on the regulatory effects of protein-DNA interactions. In very recent years, RNA binding proteins (RBPs) has been demonstrated as key players in gene post-transcriptional regulation through protein-RNA interactions. There are a large number (~1,500) of RBPs and we still do not have a clear understanding of the identity and number of genes involved in post-transcriptional regulation. CLIP-seq [95] has been proposed and applied to identify interactions of RBPs and target RNAs. Currently, its data resources are still limited. When it becomes more cost-effective and be applied to a wide selection of human disease, post-transcriptional gene regulation can be also well investigated by integrating other types of omics data.

3.7 Conclusion

A Bayesian integration approach, BICORN, has been developed for CRM inference by jointly analyzing ChIP-seq data of multiple TFs and target gene RNA-seq expression data. As a major difference from conventional methods, BICORN directly learns TF combinations in the form of CRMs from data, providing a deep-level understanding of TF cooperation beyond only predicting common target genes for multiple TFs. Using realistic simulation case studies, we have demonstrated that BICORN is more robust against noisy bindings and/or gene expression than conventional approaches. Applying BICORN to breast cancer MCF-7 ChIP-seq and RNA-seq data, we clearly show that strong and weak TFs can be clustered together as CRMs and their functional effects on gene transcription can be even more significant than some CRMs with all strong TFs. We have also found that the regulatory maps obtained from enhancer and promoter regions are complimentary, with distinct active TF members in each. This observation provides evidence of the existence of spatial interactions between two groups of TFs functionally existed since they work together to regulate the transcription of the same set of genes.

4. PSSV: A novel pattern-based probabilistic approach for somatic structure variation identification

4.1 Introduction

Genomic mutation analysis has been accelerated with the accumulation of DNA sequencing data acquired from exome regions to whole genome. Somatic mutations are likely to be critical factors determining how tumors progress and respond to treatments. To help optimize cancer therapy and predict long-term prognosis, it is important to identify somatic mutations at functional genomic regions like gene coding, promoter or enhancer regions. Structural variation (SV), a major type of genomic mutation [42], has been characterized in several studies [43, 44]. Increasing volumes of whole genome sequencing (WGS) data generated from paired tumor-normal samples makes its feasible to identify somatic SVs by comparing the sequence data of a tumor sample with that of its matched normal sample [45, 46]. Nevertheless, the lack of powerful computational methods poses challenges to the accuracy of somatic SV identification.

Earlier SV detection tools (like BreakDancer [47] and GASVPro [48]) could be used in a two-step approach to predict somatic SVs. These tools identify SVs in a tumor sample and its matched normal sample independently. SVs uniquely identified from the tumor sample can be called somatic. However, there are at least three factors that lower the confidence of SVs predicted using the two-step approach. Firstly, it is well known that germline mutations significantly outnumber somatic mutations [96]. High sensitivity for somatic SV prediction requires accurate differentiation between somatic and germline SVs, rather than a requirement only for somatic SVs. Secondly, the impurity of tumor samples (contamination of normal cells in sampled tumor tissue) increases the difficulty in differentiating somatic mutations from germline mutations [97]. Although the proportion of normal cells in the tumor sample can be estimated by joint analyzing tumor-normal WGS data [49], probabilistic models, as previously proposed for detecting somatic single nucleotide variations [98, 99], are still needed for the identification of those noisy SVs. Finally, cells with heterozygous mutations (only one copy of chromosome mutated) may function normally with the wild type copy until the latter becomes somatically mutated as a homozygous mutation (both copies of chromosome mutated). Thus, chromosome diploid feature should be also considered in modelling SVs [48], which was previously utilized

in computational modeling for single nucleotide variation [98] and copy number variation detection [100].

We propose a pattern-based probabilistic approach (PSSV) for somatic SV prediction by jointly modelling discordant and concordant read counts from paired samples in a Bayesian framework. PSSV is specifically designed to predict somatic deletions, inversions, and insertions by considering their different formation mechanisms. In detail, we define the 'true' mutation pattern (mutation statuses in a paired tumor and normal samples) of a SV as one of six states including three 'somatic', two 'germline' and one 'none' (as a special case of 'germline'). Under each state, the discordant and concordant read counts are assumed to follow specific Poisson distributions. Since read count observations from tumor samples are quite noisy, each SV is modeled as a mixture of hidden states. A Poisson mixture distribution is then introduced to each type of read count. Through an Expectation-Maximization (EM) algorithm, we iteratively estimate the prior probability for each state as well as the parameters of Poisson mixture distributions. For an individual SV, we can then estimate a most reliable hidden state to represent its mutation pattern, which is finally used to cluster the SV to either somatic or germline catalogs.

The proposed PSSV approach is first evaluated by simulation studies and further tested on TCGA breast cancer WGS data. Our simulation results show that PSSV is robust against the noise in WGS data caused by mapping errors and imperfect classification of discordant reads. The precision-recall performance of PSSV is better than that of conventional approaches, especially on detecting somatic changes from heterozygous status in normal sample to homozygous status in tumor sample. Among somatic SVs predicted by PSSV from the patient data set, we have not only identified several frequent somatic SVs at gene promoter/coding regions (which are reported by another TCGA breast cancer study [101] using whole exome sequencing data), but have also observed some somatic SVs on a high fraction of Polycomb-group (PcG) genes. PcG genes can act as tumor suppressors, and their malfunction can trigger cancer progression [102]. After large scale function enrichment analysis of genes with at least one somatic SV at gene coding, promoter and enhancer regions, we find mutations in genes affecting chemokine, ErbB, apoptosis/P53 and MAPK signaling. Abnormal activities or malfunctions of these cellular processes may play an important role in breast cancer progression.

4.2 Methods

4.2.1 Overview of PSSV

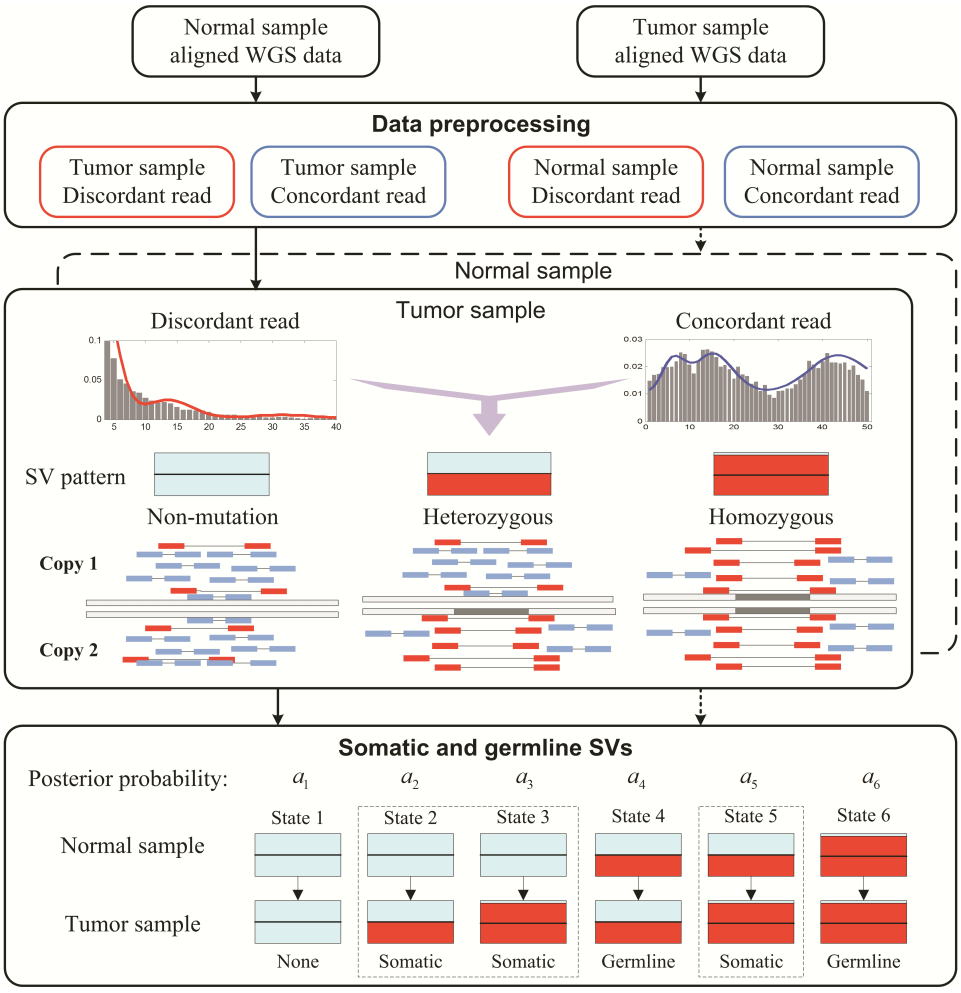


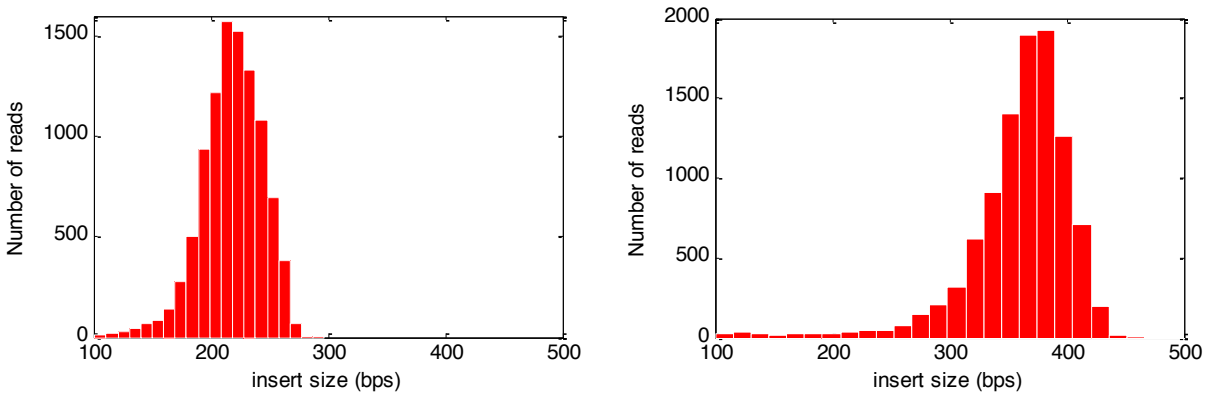
Figure 4.1 Flowchart of PSSV to identify somatic SVs from paired tumor-normal samples. PSSV features (i) read counts that are modeled by a mixture Poisson distribution with three components representing non-mutation, heterozygous and homozygous mutations; (ii) modeling of each SV as a mixture of hidden states representing different somatic and germline mutation patterns.

The workflow of the proposed PSSV approach is shown in Fig. 4.1. First, we identify a candidate SV pool from a pair of tumor and normal samples by hierarchical clustering of discordant reads in each. For each candidate SV, the concordant reads mapped around the mutation region are also counted. Each type of read count is assumed to follow a Poisson mixture distribution with three components respectively representing non-mutation, heterozygous and homozygous mutations. Non-mutation, heterozygous and homozygous

mutations in each sample can be modeled according to local alignments of discordant and concordant reads. Considering the mutation situations in a pair of tumor and normal samples, patterns of somatic or germline SVs can be characterized by six hidden states. Using a Bayesian framework, PSSV analyzes discordant and concordant read counts jointly from paired samples, and iteratively estimates the prior probability and the Poisson distribution parameters for each state, and the posterior probability for an individual SV using the EM algorithm. Finally, highly confident SVs of States 2, 3 and 5 are reported as somatic SVs.

4.2.2 Candidate structural variation detection

For each WGS data, we use the first 10000 paired-end reads to calculate the mean and variance of read insert size distribution. It can be found from Fig. 4.2 that reads come from two different read libraries and in each library their insert size follows a Gaussian like distribution with mean and standard deviation . Usually there is one library with a small mean insert size and another one with a large mean insert size. To predict discordant reads in each library, we set the upper and lower thresholds as ±3 , respectively. Paired-end reads with insert size larger than +3 are called discordant reads and are mainly used to predict deletions, as shown in Fig. 4.3(a). Paired-end reads with insert size smaller than -3 are also called discordant but they are mainly used to predict insertions, as shown in Fig. 4.3(b). There are also some reads with paired ends mapping into the same orientation (normally paired ends of each read should be mapped in opposite directions). Such reads are called as discordant as well and are mainly used to predict inversions, as shown in Fig. 4.3(c). Remaining reads with insert size close to the mean value of each read library and paired-ends mapped in opposite directions are call concordant. Concordant reads are still important in SV prediction, especially for deletion. As shown in Fig. 4.3(a), read coverage of a deletion region is significantly lower than that of flank regions. The read coverage change can be used to facilitate deletion prediction. For the other two types, although their read coverage does not change, it is still important in data normalization, which will be further introduced in the following content.



(a) (b)

Figure 4.2 Paired-end read insert size distribution of WGS data. (a) Read library #1 with mean insert size 218 bps and standard deviation 26 bps; (b) read library #2 with mean insert size 360 bps and standard deviation 46 bps.

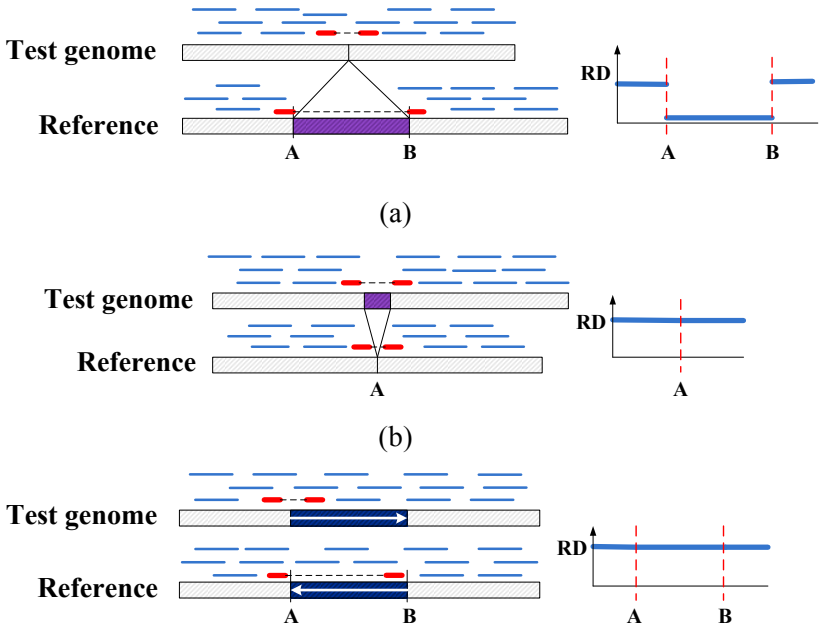


Figure 4.3 Discordant and concordant read alignments around SVs. (a) Paired ends of each read covering a deletion region are mapped at a significant longer insert size and there is a drop of read coverage within deletion region; (b) paired ends of each read covering a insertion region are mapped at a significant shorter insert size; (c) paired ends of each read covering an inversion region are mapped in the same orientation.

To identify candidate SVs of above three major types we apply a hierarchical clustering approach BreakDancer [47] to WGS data of a pair of tumor and normal samples. In each sample,

BreakDancer predicts a list of discordant read clusters as well as the overall average read depth (ARD) of a tumor or normal sample as C_T or C_N. We filter detected SVs by setting minimum number of discordant reads in the tumor sample as 4. Then, we use GATK [103] to calculate the average read depth of concordant reads within each candidate SV in both samples. To model read information (denoted by k_T for tumor and k_N for normal) from all candidate SVs, in the proposed PSSV framework we need do two rounds of read count normalizations. Since tumor impurity (contamination of normal cells in tumor sample) severely impacts the prediction of somatic regions, in the first round we calculate the proportion (denoted by) of normal cells in the tumor sample using THetA [49] and remove contaminations in read counts observed from tumor sample as $(k_T \quad k_N)/(1 \quad)$. In the second round, to eliminate the impact of local GC bias and copy number change at different regions, we calculate the read coverage at flank regions ($k_{T,FLANK}$ or $k_{N,FLANK}$) of each SV in both samples. We assume there are no mutations at flank regions and each value is proportional to the copy number of current genome segment. By assuming copy number '2' in both samples we normalize k_T as $k_T / (k_{T,FLNAK} / C_T)$ and k_N as $k_N / (k_{N,FLNAK} / C_N)$.

4.2.3 Hypothesis on read count distributions

To get a prior knowledge of read count or read coverage distributions of candidate SVs, we download from The Cancer Genome Atlas (TCGA) Data Portal (http://cancergenome.nih.gov/) a pair of tumor and normal WGS bam files sequenced from a breast cancer patient (TCGA-A2-A04P). We follow the data preprocessing steps as introduced previously to identify candidate SV regions and normalize observed discordant read count and concordant read coverage of each SV. For the selected breast cancer patient sample, the proportion of normal cells in the tumor sample is estimated using THetA as 0.35. Using read count information of deletion regions in the tumor sample, we draw a histogram of normalized discordant read counts in Fig. 4.4(a) and another histogram of normalized concordant read coverage in Fig. 4.4(b).

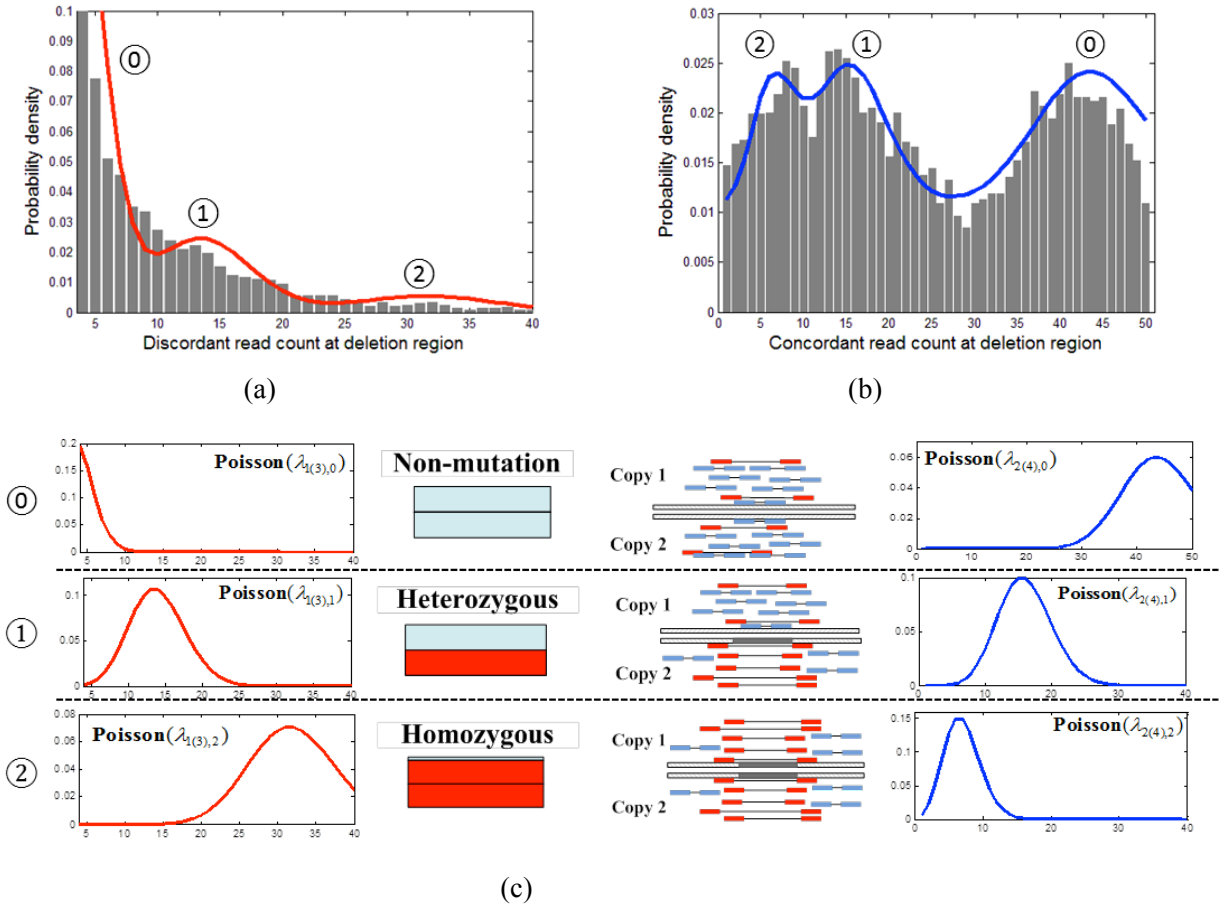


Figure 4.4 Distribution of discordant and concordant read counts at deletion regions. (a) Discordant read count fitting with a Poisson mixture distribution; (b) Concordant read coverage fitting with a Poisson mixture distribution; (c) Poisson distribution of discordant read count or concordant read coverage of each component (non-mutation, heterozygous or homozygous).

It can be found from Fig. 4.4(a) or (b) that for either discordant or concordant reads, their distribution contains multiple modes. And these two distributions show complementary features. A reasonable explanation is the existence of different mutation patterns, as illustrated in Fig. 4.4(c). Since all read counts have been normalized to copy number 2 (each chromosome has only two copies), each region should be formed by non-mutation, heterozygous mutation (one copy of chromosome deleted) or homozygous mutation (both two copies deleted). For non-mutation (component '0'), the number of discordant reads is very small, which are highly possible to false positive discordant read predictions, and the read coverage is quite high. For heterozygous mutation (component '1'), there is a significant cluster of discordant reads as well as a loss of read coverage within mutation region. For homozygous mutation (component '2'), both

chromosome copies are mutated. The number of discordant reads is much larger than that of heterozygous and the coverage of concordant reads is close to zero. It is obvious that discordant and concordant read counts should be jointly modelled to clearly model different mutation status in each sample. Noisy read count observations of each SV can be modelled as a mixture of multiple patterns. Poisson distribution has been widely used to model read count distribution in NGS data analysis [26, 104]. Hence, we use a Poisson mixture distribution to model discordant read count or concordant read coverage in each sample.

4.2.4 PSSV model

After having identified candidate SVs from WGS data of a pair of tumor and normal samples, for the n-th SV, we generate four read counts including $k_{n,D,T}$ and $k_{n,C,T}$, discordant (D) and concordant (C) read counts in a tumor (T) sample, and $k_{n,D,N}$ and $k_{n,C,N}$, discordant and concordant read counts in the matched normal (N) sample. We use index m to denote each type of read and define a read count vector $\mathbf{k}_n = [k_{n,1},...,k_{n,m},...,k_{n,4}]$. As illustrated in Fig. 4.4, we model each type of read count with a Poisson mixture distribution of three components as:

$$\begin{aligned} P(k_{n,m} \mid i=0) &= \mathbf{Pois}(\ _{m,0}),\ \text{'Non-mutation'} \\ P(k_{n,m} \mid i=1) &= \mathbf{Pois}(\ _{m,1}),\ \text{'Heterozygous'} \\ P(k_{n,m} \mid i=2) &= \mathbf{Pois}(\ _{m,2}),\ \text{'Homozygous'} \end{aligned} \quad , \tag{4-1}$$

where index i represents different mutation status in a single sample.

For a pair of samples, we jointly model four types of read count stored in $\mathbf{k}_n$ to identify somatic or germline SVs, as shown in Fig.4.5. Here, we use index i (0~2) and j (0~2) to denote mutation statuses in tumor and normal samples respectively. Since somatic or germline mutations refer to genomic changes from normal to tumor samples, only those mutation states with i j are biologically meaningful. In total we define six states as State 1 ~ 6. States with $i > j$, like States 2, 3 and 5, represent somatic mutations. States with $i = j$, like States 1, 4 and 6, represent germline mutations. Non-mutation ($i = j = 0$) is a special case of germline mutation. For each state, a unique joint Poisson distribution is used to model four types of read information observed from paired samples. Then, each SV is modelled as a mixture of six somatic or

germline states. Under a probabilistic framework, the proposed PSSV learns the joint Poisson distribution of each state using read information from all candidate SVs in paired sample and predicts the most reliable state for each SV.

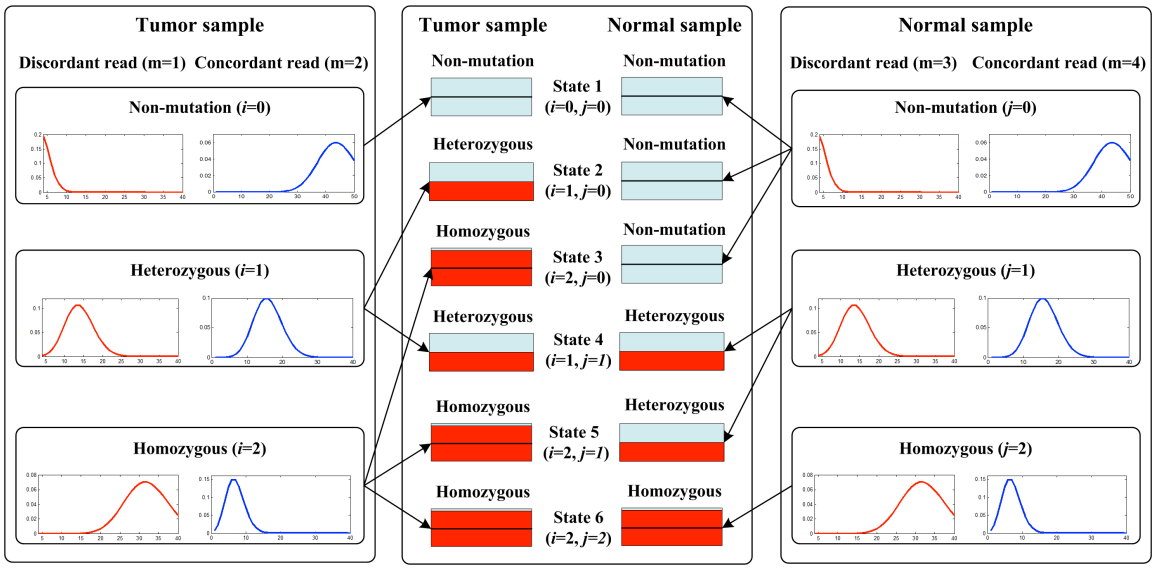


Figure 4.5 Relationship between mutation status in each sample and six hidden states. Somatic mutation occurs during the process of transition from normal cell to neoplastic cell. Only six states (combined pattern of mutation statuses in paired samples) are biological meaningful, including three 'somatic', two 'germline' and one 'none' (as a special case of 'germline').

The mutation pattern of the n-th SV is modeled by a mixture of six hidden states with a posterior probability vector $\mathbf{a}_n = \{a_{n,(i,j)} \mid 0 \quad j \quad i \quad 2\}$. Each posterior probability $a_{n,(i,j)}$ is defined by:

$$a_{n,(i,j)} = \frac{P(\mathbf{k}_n \mid i,j)P(i,j)}{\underset{j,i \quad j}{P(\mathbf{k}_n \mid i,j)P(i,j)}}, \tag{4-2}$$

where n is the index of SV, $P(\mathbf{k}_n \mid i,j)$ is the joint likelihood function for State (i,j), and $P(i,j)$ is the prior probability for State (i,j).

For the likelihood function $P(\mathbf{k}_n \mid i,j)$ in Eq. (4-2), given the State (i,j), the distribution of each read type is conditionally independent. Therefore, $P(\mathbf{k}_n \mid i,j)$ can be calculated as:

$$P(\mathbf{k}_n \mid i,j) = P(k_{n,1} \mid i)P(k_{n,2} \mid i)P(k_{n,3} \mid j)P(k_{n,4} \mid j) \\ = \mathbf{Pois}(k_{n,1} \mid \ _{1,i})\mathbf{Pois}(k_{n,2} \mid \ _{2,i})\mathbf{Pois}(k_{n,3} \mid \ _{3,j})\mathbf{Pois}(k_{n,4} \mid \ _{4,j}) \quad (4\text{-}3)$$

Here, we define a Poisson parameter matrix λ as $\left\{ \ _{m=1(2),i}, i=0\sim 2; \ _{m=3(4),j}, j=0\sim 2 \right\}$. Selection of λ will determine the final performance for somatic SV detection. For deletion or insertion, the classification of discordant and concordant reads is based on the insert size distribution of their paired ends, where the value of the cut-off threshold will affect the number of read counts and the overall distribution of each type. Therefore, the actual value of λ is unknown and must be estimated. We assume a Gamma prior $P(\ )$ on each as shown in Eq. (4-4), which is a conjugate prior of the Poisson likelihood.

$$P(\ _{m,i}) = \mathbf{Gamma}(\ _{m,i}, \ _{m,i}) \quad m=1,2 \\ P(\ _{m,j}) = \mathbf{Gamma}(\ _{m,j}, \ _{m,j}) \quad m=3,4 \quad (4\text{-}4)$$

where mean value (/) of each Gamma distribution is determined by the whole genome average read depth (ARD) in the tumor or normal sample, and its mutation status; the variance (/ ^{2}) is designed to control the scale of possible parameter values by avoiding severe overlap between different components. Initial mean value for / of each Poisson component for somatic deletion detection is set as Table 4.1. For the other two subtypes, their parameter settings for discordant read count modelling are the same as deletion. However, at insertion or inversion regions there is no read coverage change even if heterozygous or homozygous mutations occur. Hence, we model their concordant read coverage distribution using a single Poisson distribution with the same parameter setting as component '0'.

Table 4.1 Initial Poisson distribution parameter settings for somatic deletion detection

Component i or j	**Tumor sample**		**Normal sample**	
	discordant read	concordant read	discordant read	concordant read
0		C_T		C_N
1	$C_T / 2$	$C_T / 2$	$C_N / 2$	$C_N / 2$
2	C_T		C_N	

For the prior probability $P(i,j)$ (denoted as $_{(i,j)}$) in Eq. (4-2), we assume that $_{(i,j)}$ follows a Dirichlet prior as shown in the following equation with modes $(\ _{(i,j)}\ \ 1)/(\ _{j,i\ \ j}\ _{(i,j)}\ \ 6)$:

$$P(\boldsymbol{\pi}) = \frac{1}{B(\boldsymbol{\gamma})}\ _{j,i\ \ j}\ ^{(i,j)\ \ 1}_{(i,j)}, \tag{4-5}$$

where $B(\boldsymbol{\gamma})$ is a constant normalization term, $\boldsymbol{\pi}$ = $\left\{\ _{(i,j)} \mid 0 \quad j \quad i \quad 2\right\}$ and $\boldsymbol{\gamma}$ = $\left\{\ _{(i,j)} \mid 0 \quad j \quad i \quad 2\right\}$. For the parameters setting $_{(i,j)}$, we need to consider the practical situation for SV detection. First, our candidate SV list is generated using discordant reads clustering, so there should be very few regions with State 1 ($i = 0, j = 0$). Second, the probability for 'homozygous' (mutation on both copies) is much lower than that for 'heterozygous' (mutation on one copy). Therefore, the numbers of mutations belonging to States 3, 5 and 6 ($i = 2$) should be relatively small. Third, germline mutation outnumbers somatic mutation significantly. Thus, the total number of mutations of States 2, 3 and 5 ($i > j$) is much smaller than that of State 4 ($i = 1, j = 1$). Overall, it is reasonable to assume that the prior probability of State 4 is the highest with $_{1,1} = G + 1$, where $G >> 1$, and $_{i,j} = G/10$ for the other states.

With the probabilistic model formulated, the problem of determining whether the *n*-th SV is a somatic or germline mutation can be mathematically stated as follows: given the model parameters ($\boldsymbol{\lambda}$;) and read counts $\mathbf{K} = \{\mathbf{k}_n \mid n = 1 \sim N\}$, how to estimate the posterior probabilities $\{\mathbf{a}_n \mid n = 1 \sim N\}$. Since both $\boldsymbol{\lambda}$ and are unknown, we define a posterior probability in the following equation to estimate these parameters:

$$\begin{aligned} P(\boldsymbol{\lambda}, \boldsymbol{\pi} \mid \mathbf{K}) \quad & \prod_n P(\mathbf{k}_n \mid \boldsymbol{\lambda}, \boldsymbol{\pi}) P(\boldsymbol{\lambda})\ P(\boldsymbol{\pi}) \\ & \prod_n\ _{j,i\ \ j}\ \prod_{(i,j)} \prod_{m=1,2} P(k_{n,m} \mid i) P(\ _{m,i}) \prod_{m=3,4} P(k_{n,m} \mid j) P(\ _{m,j}) \quad P(\boldsymbol{\pi}) \end{aligned} \tag{4-6}$$

In the following EM algorithm, we use the E-step and the M-step iteratively to update each parameter until the improvement of estimated posterior probability $P(\boldsymbol{\lambda}, \boldsymbol{\pi} \mid \mathbf{K})$ is smaller than 10^{-4}. The EM algorithm can be summarized as follows:

E-step:

$$\hat{a}_{n,(i,j)} = \frac{\prod_{m=1,2}^{(i,j)} P(k_{n,m} \mid i) P(\quad_{m,i}) \prod_{m=3,4} P(k_{n,m} \mid j) P(\quad_{m,j})}{\sum_{j,i \quad j} \prod_{m=1,2}^{(i,j)} P(k_{n,m} \mid i) P(\quad_{m,i}) \prod_{m=3,4} P(k_{n,m} \mid j) P(\quad_{m,j})} \tag{4-7}$$

M-step:

$$\quad_{(i,j)} = \frac{\hat{a}_{n,(i,j)} + (\quad_{(i,j)} \quad 1)}{N + \sum_{j,i>j} (\quad_{(i,j)} \quad 1)}. \tag{4-8}$$

$$\begin{aligned} \quad_{m,i} &= \sum_{n \quad j=0}^{i} a_{n,(i,j)} \left(k_{n,m} \quad 1 + \quad_{m,i}\right) \Big/ \sum_{n \quad j=0}^{i} a_{n,(i,j)} \left(1 + \quad_{m,i}\right) \\ \quad_{m,j} &= \sum_{n \quad i=j}^{2} a_{n,(i,j)} \left(k_{n,m} \quad 1 + \quad_{m,j}\right) \Big/ \sum_{n \quad i=j}^{2} a_{n,(i,j)} \left(1 + \quad_{m,j}\right) \end{aligned} \tag{4-9}$$

Note that in our experiments testing PSSV on both simulated and real data, the EM algorithm will usually converge within 20 iterations. After convergence, for the n-th SV, we select the state with the highest $a_{n,(i,j)}$ to represent the mutation pattern of the region. If $a_{n,(i,j)} > 0.5$ and $i > j$ (States 2, 3 or 5), we will predict the n-th SV as somatic. More details about the estimation of each model parameter can be found in **Appendix E**.

4.3 Simulation

To evaluate PSSV's performance on somatic SV prediction, especially when discordant and concordant reads are imperfectly classified, for each SV subtype (deletion, insertion, or inversion), we simulated 100 SVs for each mutation state. Discordant and concordant read counts were generated using Poisson mixture distributions. The first component represented non-mutation. The mean value was non-zero due to the existence of tumor impurity. It reflected the residual contaminations after data normalization and affects the detection accuracy of mutation

regions. To test the robustness of PSSV on differentiating mutation and non-mutation regions, we set the average read depth in each sample as 40 (20 on each copy of chromosome) and varied the mean parameter of the first component from 1 to 5 in order. In addition, we added Gaussian distributed noise to all three components to mimic inaccuracy occurring at discordant read prediction, even though all reads were correctly aligned. We simulated multiple data sets with different noise levels and applied PSSV to each data set to predict somatic SVs. The prediction accuracy performances of PSSV on different SV subtypes are shown in Fig. 4.6.

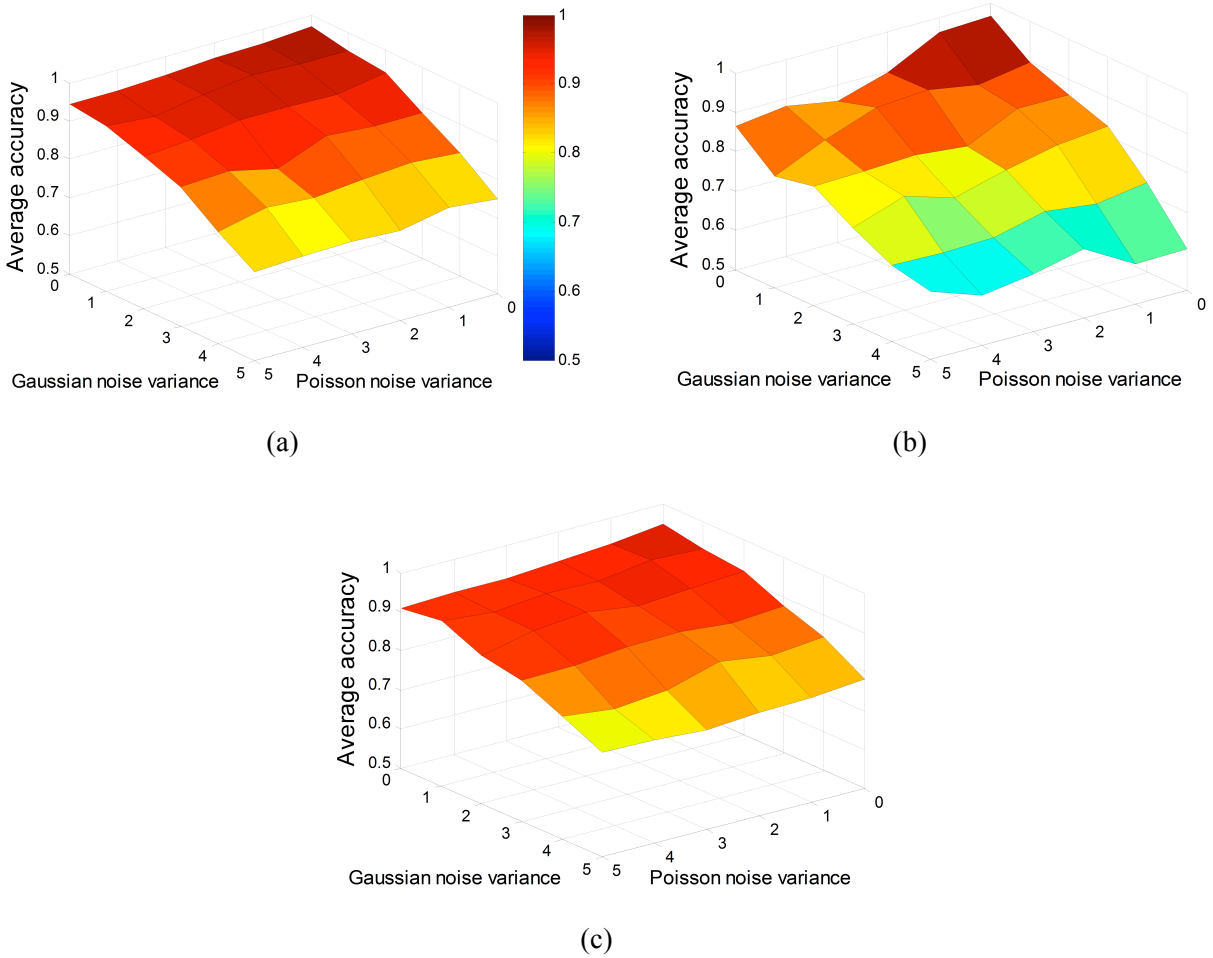


Figure 4.6 Performance evaluation of PSSV on simulated data. (a), (b) or (c) represents the accuracy of PSSV on detecting somatic deletions (a), somatic insertions (b) or somatic inversions (c) under different noise levels.

For deletion SV detection, the performance of PSSV is robust against noise even though the mean value of the first component of Poisson distribution is high (Fig. 4.6(a)). The impact of

Gaussian noise is evident but PSSV provides a high accuracy even when the level of Gaussian noise is high. For insertion SV detection, only discordant read counts can be utilized since there is no concordant read coverage change at insertion regions. However, only partial discordant reads can be recorded because some discordant reads with each read end mapped to the inserted genomic segment cannot be successfully mapped to the reference genome. The accuracy of insertion detection degrades much faster when the level of noise, especially the Gaussian distributed noise, increases (Fig. 4.6(b)). If the discordant read count is accurate and the Gaussian distributed noise is small, PSSV can achieve a high accuracy on insertion SV detection. For inversion detection, as shown in Fig. 4.6(c), PSSV is very robust under different noise conditions. The impact of Gaussian noise is much lower than that on deletion or insertion SV detection because the insert size is not used to define inverted discordant reads, whereas the orientation of two ends of a read is considered.

Average read depth (ARD) has been demonstrated as the major technical factor of WGS data affecting mutation detection. We further simulated WGS data with different ARD. Here, we used a SV simulation tool, RSVSIM [105], to simulate a pair of mutated 'tumor' and 'normal' genomes refereeing to a segment of human reference genome (GRCh37). For each type of SV, 300 germline (100 for State 1, 4 or 6) and another 300 somatic SVs (100 for State 2, 3 or 5) were implanted to the 'tumor' and 'normal' genomes. We sequenced each mutated genome using WGSIM [106] by setting the read end length as 100 bps, the mean of insert size as 300 bps and the standard deviation of insert size as 30 bps. We varied ARD from 10, 20 to 40 and in total simulated three pair of 'tumor' and 'normal' WGS samples. Paired-end reads of simulated WGS data were aligned to human reference genome (GRCh37) using BWA [107]. We applied PSSV to WGS data of each pair of 'tumor' and 'normal' samples. Curves of Receiver operating characteristic (ROC) of PSSV were presented in Fig. 4.7. It can be seen that when the read coverage comes to 40, the area under the curve (AUC) of ROC curves of PSSV can reach up to 0.94. And its performance does not drop much when ARD comes to 20. However, if ARD further comes down to 10, the performance of PSSV drops dramatically, especially for insertion detection. In real data like TCGA WGS data, ARD is usually between 30 ~ 40 and sometimes even higher. Within this range, PSSV can provide a good performance on all kinds of somatic structural variation detection.

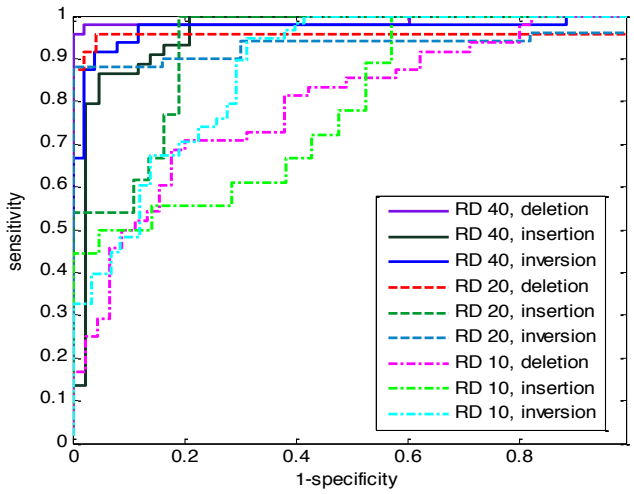


Figure 4.7 ROC performance evaluation of PSSV on simulated data with realistic settings.

We further compared PSSV with two existing tools (BreakDancer [47] and GASVPro [48] using Poisson statistics to score each SV) which could be used to call somatic regions. BreakDancer only used discordant reads to identify SVs in each sample. For a fair comparison, I also calculated the concordant read coverage of each SV, especially for identified deletion regions. Then, k-means was used to cluster BreakDancer predicted SVs into six somatic or germline groups based on their discordant and concordant read observations from paired samples. GASVPro jointly modelled discordant and concordant reads and precited heterozygous or homozygous mutation status for each SV in each sample. Comparing mutations statuses in paired samples, we assigned a somatic or germline state to each GASVPro predicted SV. BreakDancer and GASVPro only report SV status in each sample with very strong mutation pattern. Therefore, we only examined how many somatic mutation events can be captured by each method. We set the ARD as 20 and used each comparable method to identify somatic SVs. The F-measures (a joint measurement of precision and recall) of competing methods are shown in Fig. 4.8. Specific precision and recall values of competing methods on State 5 somatic SV (heterozygous mutation in the normal sample and homozygous mutation in the tumor sample) detection are summarized in Table 4.2.

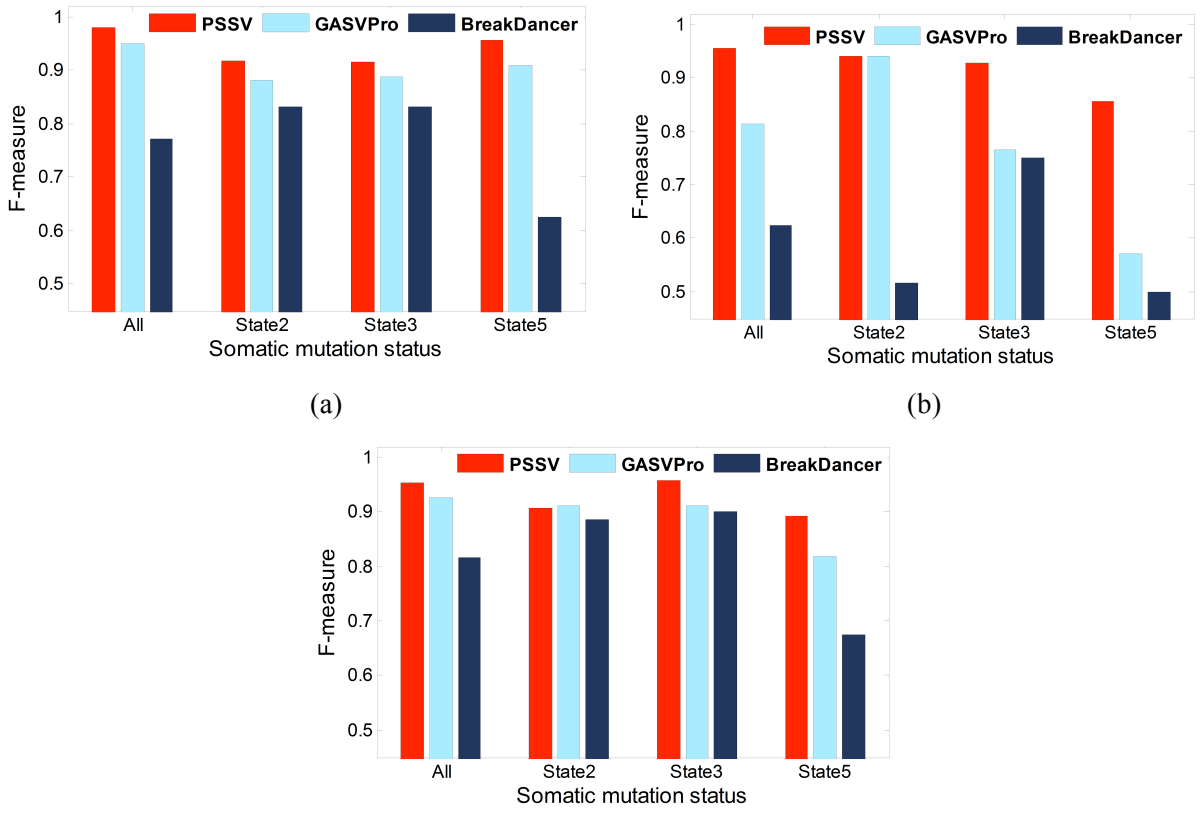


Figure 4.8 Performance comparison of PSSV, BreakDancer and GASVPro on simulated WGS data. (a), (b) or (c) represents the F-measures of competing methods on detecting somatic deletions (a), somatic insertions (b) or somatic inversions (c).

Table 4.2 Precision-recall performances of competing methods for somatic SV detection.

METHOD	PSSV		GASVPro		BreakDancer	
SV type	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
DEL State 5	0.978	**0.936**	0.976	0.851	0.758	0.532
INS State 5	0.857	**0.857**	0.571	0.571	0.571	0.444
INV State	0.844	**0.948**	0.878	0.766	0.867	0.553

Note: DEL – deletion; INS – insertion; INV – inversion.

As shown in Fig. 4.8, PSSV achieved a better detection performance on each type of somatic SVs than the other two competing methods. Clearly, differentiating heterozygous and homozygous mutation statues (as done in both PSSV and GASVPro) significantly improved

performance on somatic SV detection than that of using a universal model (as with BreakDancer). The improvement is especially clear for the detection of somatic SVs with State 5 (homozygous in tumor and heterozygous in normal). Further, by comparing the performances of PSSV and GASVPro, a joint modelling of paired samples by estimating multiple components parameters makes PSSV outperform GASVPro, especially for the more challenging insertion SV detection cases where the number of discordant reads is relatively small. We can see from Table 4.2 that PSSV, without loss of precision, has a higher recall performance than the other two methods.

4.4 TCGA Breast cancer patient WGS data analysis

To demonstrate the capability of PSSV on real WGS data analysis, we analyzed 14 TCGA estrogen receptor negative (ER-) breast cancer patients with paired tumor-normal samples. In total we identified 9520 somatic deletions (State 2: 8414, State 3: 286, State 5: 820), 1233 somatic insertions (State 2: 1169, State 3: 32, State 5: 32) and 3732 somatic inversions (State 2: 3409, State 3: 295, State 5: 28). To study the functional roles of somatic SVs, we mapped PSSV identified somatic SVs to gene coding, promoter and enhancer regions and then did functional enrichment analysis on genes.

4.4.1 PSSV identified somatic SVs at promoter and coding regions

We mapped somatic SVs to promoter (±10 kbps from each transcription starting site) and gene coding regions according to UCSC human RefSeq 19 since these two types of regions have overlap. In total, we identified 3,374 deletions, 515 insertions, and 1,166 inversions with State 2; 105 deletions, 8 insertions, and 87 inversions with State 3; 386 deletions, 15 insertions, and 11 inversions with State 5. We compared the results of different methods for somatic deletion, insertion and inversion, where a common SV is counted if identified by at least two methods on the same gene from the same sample (Fig. 4.9(a)-(c)). For each SV subtype (deletion, insertion, or inversion), PSSV captured novel somatic SVs.

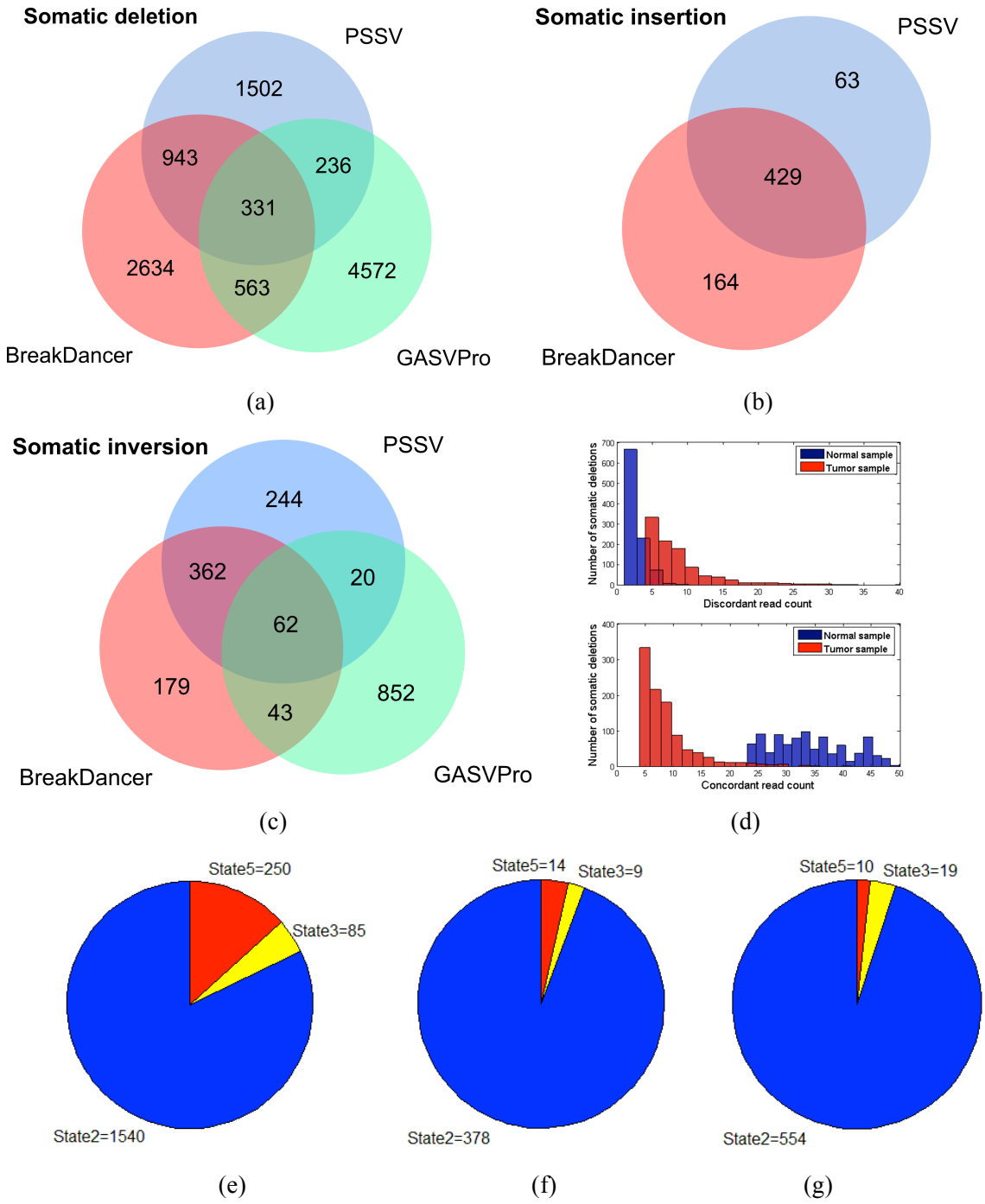
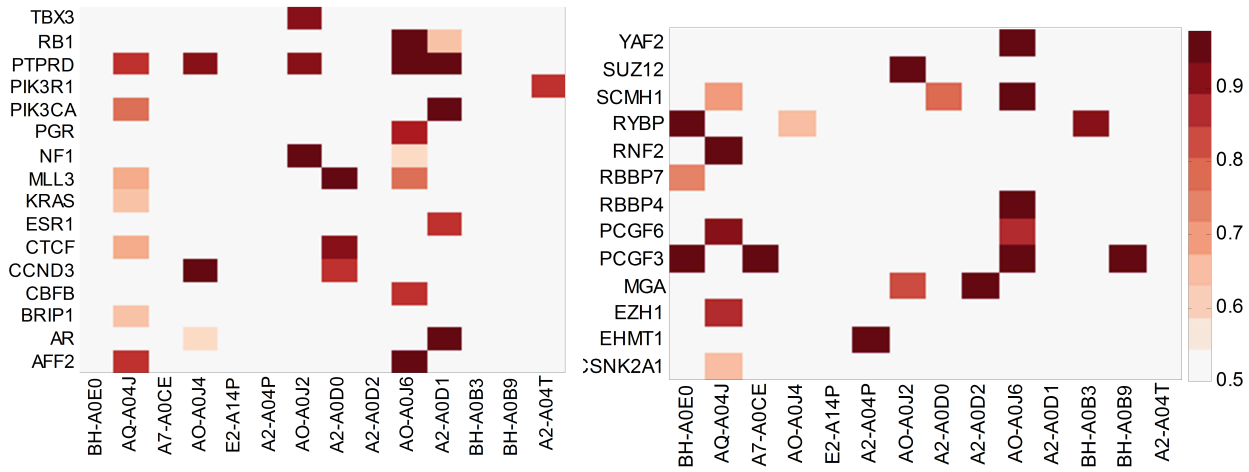


Figure 4.9 Summary of somatic SVs detected from the TCGA breast cancer WGS data at promoter and gene coding regions. (a) Comparison of somatic deletions detected by each method; (b) comparison of somatic insertions detected by each method; (c) comparison of somatic inversions detected by each method; (d) histograms of discordant and concordant read counts of somatic deletions detected by PSSV only; (e) number of PSSV detected somatic deletions with each somatic state; (f) number of PSSV detected somatic insertions with each somatic state; (g) number of PSSV detected somatic inversions with each somatic state.

For those somatic deletions detected by PSSV only (Fig. 4.9(a)), we plotted the histogram of their discordant and concordant read counts of paired samples in Fig. 4.9(d). While the number of discordant reads is non-zero in the normal sample, the number is much smaller than that in the tumor sample. Concordant read counts are significantly higher in the normal sample than in the tumor sample, indicating the existence of somatic deletions at these regions. We then focused on those genes containing at least one PSSV-detected somatic SV. Since one gene can be hit by somatic SVs from multiple samples, we merged gene lists from all samples and obtained 1,620 genes with somatic deletions, 389 genes with somatic insertions, and 558 genes with somatic inversions. From Fig. 4.9(e)-(g), we can see that a proportion of genes have homozygous mutations (17.8% for somatic deletion, 5.7% for somatic insertion, 5% for somatic inversion). Among all somatically mutated genes (regardless of SV subtype), there are 2,186 genes containing heterozygous somatic SVs (State 2, blue part in Fig. 4.9(e)-(g)) and 345 genes of homozygous somatic SVs (State 3 and 5, red and yellow parts in Fig. 4.9(e)-(g)). Homozygous mutations are more likely to cause gene malfunction [108]. After performing functional annotation on these homozygous genes using DAVID [109], we identified 16 cancer related genes (*p*-value 3e-3), 13 genes enriched in the MAPK signaling pathway (*p*-value 9.1e-3) and 8 genes enriched in the Wnt pathways. Notably, our results are consistent with the available genomic knowledge that mutations of MAPK signaling pathway members are frequently found in human breast tumors [110].

We proceeded to investigate the PSSV-identified genes in available databases to check whether PSSV had captured somatic SVs occurring at any known breast cancer specific genes. In a previous study [45, 46], researchers reported 20 breast cancer related genes with somatic copy number variations using whole exome sequencing data of 507 TCGA breast cancer patients, including all individuals used in our analysis. PSSV successfully captured 10 genes including PIK3CA, MLL3, RB1, TBX3, CBFB, AFF2, PIK3R1, PTPRD, NF1 and CCND3. We further compared our results with available literature and identified another 6 breast cancer related genes including PGR, KRAS, ESR1, CTCF, CBFB and AR. The posterior probabilities of the somatic SVs occurring at these breast cancer specific genes are shown in Fig. 4.10(a). As major factors in epigenetic regulation, somatic genetic alteration of the Polycomb-group (PcG) genes predisposes normal cells to various types of disease [102]. PcG genes are known tumor suppressor genes,

whose target genes have been demonstrated to be predictive for breast cancer prognosis [111]. Among the 30 PcG genes reported in [102], PSSV identified 13 genes with somatic SVs, of which the posterior probabilities are shown in Fig. 4.10(b).



(a) (b)

Figure 4.10 Detected somatic SVs in breast cancer specific genes and in Polycomb genes. 'Red' color represents the posterior probability reported by PSSV for each somatic SV in each sample.

Due to the limited number of breast cancer patient samples used in this study, some predictions of PSSV may not be breast cancer specific. The Structure Variation Analysis Group (SVAG) published and validated 22,025 deletions identified from 179 unrelated individuals [45]. If a somatic SV matches a variant identified from the normal human population, such calls are more likely to be non-somatic, or non-tumor related. We compared 1,620 identified genes with somatic deletions (Fig. 4.9(e)) with any known variants in the normal human population to exclude spurious somatic calls. We found 659 somatic genes without any overlap with previously reported variants [45]. Functional annotations on these somatic genes using DAVID [112] show that 12 genes are significantly enriched in the chemokine signaling pathway (*p*-value 2.5e-2). Apoptosis and the MAPK signaling pathway are also enriched.

4.4.2 PSSV identified somatic SVs at enhancer regions

Candidate enhancer regions were extracted from three different breast cancer cells including MCF-7 (estrogen responsive), LCC-1 (estrogen independent) and LCC-9 (drug

resistant). For each cell type, we generated candidate enhancer regions from vehicle and E2 treated conditions by joint analyzing public and in house ChIP-seq data of H3K4me1, H3K27ac and H3K4me3. We merged regions from three types of cells under two conditions together because we do not have clear knowledge about which cell line model can better represent the cells in real tumors. In total we identified 22,979 candidate enhancer regions. We mapped PSSV predicted somatic SVs to enhancer regions and identified in total 178 enhancer regions with overlap of 186 somatic deletions, 29 somatic insertions and 37 inversions. For each SV subtype, the number of somatic SVs under each state is shown in Fig. 4.11.

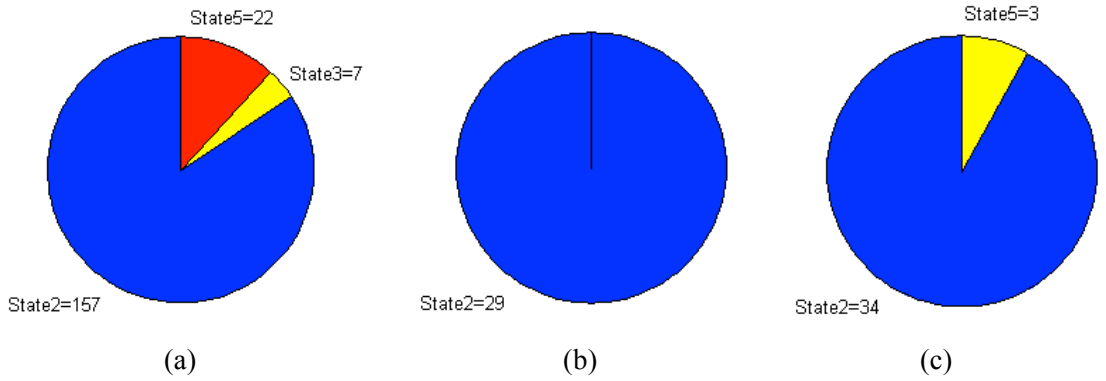


Figure 4.11 Summary of somatic SVs detected from the TCGA breast cancer WGS data at enhancer regions. (a) Number of PSSV detected somatic deletions with each somatic state; (b) number of PSSV detected somatic insertions with each somatic state; (c) number of PSSV detected somatic inversions with each somatic state.

Functional study of enhancer regions is very hard due to the lacking of clear mapping relationship between enhancer regions and target genes. First, following the strategy as proposed in [91], we only selected the closest target gene for each enhancer and limited the distance within 200k bps. For 178 enhancer regions containing somatic SVs, we identified 165 target genes. Functional enrichment analysis using DAVID [113] showed that apoptosis was enriched including MNT1, ARHGEF18, BIRC7, DAP3, HSPA9, HIP1, MKL1, PIK3CA, POLB and ZMAT3. Low expression of DAP3 has been observed in human breast cancer [114] but its upstream regulatory mechanism is still not clear. Our study showed that DAP3 would be a potential target gene of enhancer regulation and somatic mutations were found at this enhancer region. This founding may provide an explanation to the low expression of DAP3. HSPA9 was reported as a new target for the reactivation of tumor-suppressive signaling of the pathway in

cancer [115]. In a previous promoter-focused study, we already found somatic SVs at gene coding/promoter regions of PIK3CA. Here, we further identified a somatic SV at the enhancer region of PIK3CA. This founding is consistent with the reported coexistence of mutations of PIK3CA and other PI3K-enhancing mechanisms [116]. It is also interesting to find SVs of MKL1 because its activity is enhanced in estrogen-independent breast cancer cell lines like LCC-1 and LCC-9 cells and it is active in ER negative breast cancer [117].

Second, we changed the distance limit on enhancer-gene mapping to 10M bps and selected at most 10 upstream genes and 10 downstream genes for each enhancer region, which is assumed to be the largest scale that each enhancer region can regulate [118]. In total we identified 901 genes. Functional enrichment analysis using DAVID showed that ErbB signaling pathway (8 genes), Pathways in cancer (19 genes), Jak-STAT signaling pathway (9 genes), p53 signaling pathway (5 genes) and MAPK signaling pathway (12 genes) were all enriched. Besides previously mentioned genes, we found BRCA2, CHEK2, E2F1, EGF, ELK4, MAPK8, LAMA family genes and STAT family genes. These genes have been frequently reported in breast cancer research, especially using TCGA breast cancer patient data analysis.

4.5 Discussion

A pattern based probabilistic model has been developed for somatic SV prediction using paired tumor-normal WGS data (PSSV). Based on both simulation and real TCGA breast cancer WGS data studies, we have demonstrated that PSSV can predict highly confident somatic SVs using a probabilistic model with multiple hidden states. We have also shown that PSSV can capture somatic SVs missed by conventional methods such as BreakDancer and GASVPro, for example, somatic mutations with State 5 (heterozygous mutation in normal sample while homozygous mutation status in tumor sample). All the model parameters are learned from the input data during the EM iteration procedure of PSSV. The only prior knowledge incorporated is the ARD of each sample, which can be easily obtained.

Correctly identifying insertions is a particularly challenging task. Reads around insertion regions cannot be fully mapped to the reference genome because a certain segment is unknown. The number of available reads used to detect insertion SVs is so limited that, in our simulation studies, all methods have a relatively low performance on somatic insertion prediction. In

addition, using discordant reads with significant small insert size can only identify insertions with small size. A lot of insertions with longer size may be missed. That's the major reason for the smaller number of insertions than that of deletions or inversions. Some methods use reads to construct longer contigs first and then map each contig to the reference genome to find longer insertions. However, the insertion detection problem is still not fully solved. Reads with longer insert size in NGS data, high read coverage, more advanced detection methods, and a deeper understanding of the insertion formation mechanism are needed to overcome the current bottle neck of insertion SV detection.

We have observed that a large portion of detected somatic SVs represent heterozygous mutation (State 2). The fraction of somatic SVs with homozygous mutation is relatively low. These observations are biologically reasonable, but bring difficulties to reliably establishing whether heterozygous somatic mutations on one copy of chromosome will affect gene transcription. Even though using gene expression data like RNA-seq we can check gene different expression between tumor and normal samples, if a genomic mutation is not located exactly at the coding region of a gene, it cannot be used as a deterministic factor to explain the gene transcription change. As introduced in previous chapters, transcription factors binding at promoter or enhancer regions also control the gene expression process. Furthermore, for identified somatic mutations in promoter or enhancer regions, without mapping them with binding signals of transcription factors, it is hard to claim them as functional genomic mutations. All in all, a joint data analysis of both binding signals and genomic mutation at gene coding, promoter and enhancer regions is needed to infer the causal factors on gene expression change. Those factors, if validated in biological laboratory and targetable by existing or new drugs, can improve the treatment of tumor patients.

4.6 Conclusion

We developed a novel pattern-based probabilistic approach, namely PSSV, to identify somatic SVs from WGS data. Specifically, we modeled discordant and concordant read counts from paired samples jointly in a Bayesian framework. Each type of read counts at SV regions is modelled to follow a mixture of Poisson distributions in tumor or normal samples and each SV is modelled as a mixture of hidden states representing different somatic and germline mutation

patterns. Using simulated WGS data we demonstrated that PSSV can more effectively detect those somatic SVs with heterozygous mutations in normal samples and homozygous mutations in tumor samples. PSSV was then applied to WGS data acquired from multiple TCGA breast cancer paired samples. Functional enrichment analysis on genes with at least one somatic SV at gene coding, promoter or enhancer regions showed that PSSV successfully captured the somatic SVs of key factors responsible for breast cancer development.

5. Contribution, Future work and Conclusion

5.1 Summary of original contribution

We have developed new computational methods to model or integrate ChIP-seq, RNA-seq and DNA-seq data for gene transcriptional regulation analysis. In this chapter, we briefly summarize the original contributions of this dissertation research.

5.1.1 Transcription factor binding site identification using ChIP-seq data

A novel Bayesian approach, ChIP-BIT, has been proposed to identify transcription factor binding sites (TFBSs) at promoter or enhancer regions. In the approach, read intensities of sample and input ChIP-seq data are jointly analyzed so that weak binding signals can be reliably identified. ChIP-BIT uses a Gaussian mixture model (consisting of global and local Gaussian components) to capture both foreground and background signals in the sample ChIP-seq data. A unique feature of ChIP-BIT is that the component modeling background signals is specially designed as a local Gaussian distribution that can be estimated accurately from the input data. Specific for TFBS inference at the promoter region, an exponential distribution is used to model the relative distance of TFBS to the nearest TSS. Using an Expectation-Maximization algorithm, a posterior probability is estimated for each TFBS.

We have applied ChIP-BIT to simulated ChIP-seq data with known binding sites, and compared the performance of ChIP-BIT with those of several existing tools. In this study, we have mainly simulated two cases by considering different distribution features of TFBSs at gene promoter or enhancer regions. Specifically in Case 2, considering that enhancer regions have many weak bindings, we simulate binding signals with a distribution close to that of background signals. Simulation results show that ChIP-BIT can identify weak binding signals with a high accuracy.

We finally apply ChIP-BIT to real ChIP-seq data acquired from an in-house profiling study of NOTCH3 and PBX1 to help understand their functional role in breast cancer cells, particularly at promoter regions. We have validated target genes of NOTCH3 and PBX1 by performing TF knock down experiments. The functional enrichment analysis shows that NOTCH3 and PBX1

are involved in the regulation of Notch and Wnt signaling pathways. In addition, we apply ChIP-BIT to ER-α ChIP-seq data and identify binding sites at enhancer regions. GRO-seq data measuring enhancer RNA (eRNA) is used to examine whether enhancer regions are activated by ER-α. Results show that over 50% of enhancer regions with ER-α binding sites are active. ER-α is a major activator of enhancer regions in breast cancer cells.

5.1.2 Cis-regulatory module inference by integrating ChIP-seq and RNA-seq data

We have developed a novel Bayesian integration method, BICORN, to infer cis-regulatory modules (CRMs) at promoter or enhancer regions by jointly modelling ChIP-seq binding signals of multiple TFs and RNA-seq target gene expression. In this method, we directly model each CRM (a combination of multiple TFs) as a variable. Target gene expression is modelled as a log-linear combination of hidden activities of TFs in the functional CRM. We use a Gibbs sampling technique to learn the posterior distribution of CRMs as well as the model parameters including TF activities. BICORN provides TF association maps at a deeper level than existing methods.

We apply BICORN to simulated data with respect to different experimental settings of noise in gene expression data and binding network. Results clearly show that BICORN is more effective than existing tools to identify CRMs for each gene even though bindings are weak. In addition, BICORN shows its robust performance against gene expression noise by reducing the overfitting effect for large-scale regulatory network identification. We further validate the performance of BICORN using benchmark time course gene expression data and regulatory networks. BICORN has a significant precision-recall improvement over existing tools for functional binding prediction.

We finally apply BICORN to breast cancer MCF-7 cell ChIP-seq and RNA-seq data to infer CRMs at promoter or enhancer regions. Results show that a group of TFs is functional at proximal promoter regions while another different group of TFs is functional at distant enhancer regions. Only a few TFs are shared working at both regions. In each group of TFs, BICORN predicts several CRMs as well as target genes of each. The results highlight the necessity of investigating regulatory effects of TFs at both enhancer and promoter regions. Using BICORN we can construct a complete regulatory map including major TFs, co-factors, their associations as CRMs and functional target genes.

5.1.3 Somatic structural variation detection using WGS data

We have proposed a probabilistic approach, PSSV, for somatic SV prediction. In a Bayesian framework, we jointly model read count information in whole genome DNA-seq (WGS) data from a pair tumor and normal samples. Specifically, potential mutation patterns (mutation statuses in a paired tumor and normal samples) including three 'somatic', two 'germline', and one 'none' (as a special case of 'germline') are defined as hidden states using the diploid feature of human chromosome. The observed pattern of each SV is modeled as a mixture of hidden states and its read count observation is modelled using a Poisson mixture distribution. Using an Expectation-Maximization algorithm, for an individual SV, we estimate a most reliable hidden state to represent its mutation pattern, which is finally used to cluster this SV to either the somatic or germline category.

We have evaluated the performance of PSSV using simulated WGS data. Results show that PSSV is quite robust against the noise in WGS data caused by the read mapping errors and the imperfect classification of discordant reads. With average read coverage between 20x ~ 40x in both tumor and normal samples, PSSV can predict somatic SVs with a high accuracy. Further comparison with existing SV detection methods shows that PSSV significantly improves the detection accuracy for a particular type of somatic SVs with heterozygous mutation status in the normal sample and homozygous mutation status in the tumor sample.

We finally apply PSSV to TCGA breast cancer patient WGS data and identify somatic SVs at gene coding, promoter and enhancer regions. We have not only identified several frequent somatic SVs at gene promoter/coding regions (which are reported by another TCGA breast cancer study [101] using whole exome sequencing data), but have also identified some somatic SVs on a high fraction of Polycomb-group (PcG) genes. PcG genes can act as tumor suppressors, and their malfunction can trigger cancer progression [102]. Functional enrichment analysis of genes with at least one somatic SV at gene coding, promoter or enhancer regions reveals that chemokine, ErbB, apoptosis/P53 and MAPK signaling pathways are highly enriched. Literature survey of genes involved in the above-mentioned signaling pathways indicates that abnormal activities or malfunctions of these cellular processes may play an important role in breast cancer progression.

5.2 Future work

5.2.1 Transcriptional regulation analysis

Differential binding signal detection using ChIP-seq data of TF knockdown experiment

The proposed ChIP-BIT method is mainly used to detect TFBSs of a single TF. In results validation, only 15% of genes with TFBSs have been successfully validated. This is also observed in other studies using similar validation strategy [119]. It indicates that binding signals of those non-differentially expressed genes are not fully silenced by the TF knockdown experiment and/or those genes are not regulated by the selected TF alone. However, using gene expression as binding signal validation cannot provide more detailed explanations. In recent years, as the cost of ChIP-seq experiment is coming down, researchers can measure binding signals directly associated with one specific TF by generating two ChIP-seq experiments: TF sample and TF knockdown sample [120]. Binding signals unique in the TF sample ChIP-seq experiment can be used to locate very specific binding sites for the TF under investigation.

Instead of examining differential signals between TF sample and input ChIP-seq data for one TF, the proposed ChIP-BIT method can be extended to model differential signals between TF sample and TF knockdown sample. ChIP-BIT currently uses a two component Gaussian mixture model to analyze binding and background signals. It can be extended to include multiple components; each component represents a certain level of binding signal enrichment. Genomic regions with different levels of binding signal enrichment are more possible to be direct binding sites. The advantage of using the extended ChIP-BIT method is still on detecting weak binding signals. If a genomic region has a weak binding signal in the TF sample experiment and no binding signal enrichment in the TF knockdown experiment, it will be detected by ChIP-BIT and further predicted as a direct binding site. We have already generated ChIP-seq data of a few TFs as well as their knockdown experiments. Extension and application of ChIP-BIT to the new data set will provide more specific binding signals for those TFs of interest.

Enhancer region identification by integrating multiple data types

In this dissertation research we develop BICORN for cis-regulatory module inference by integrating ChIP-seq and RNA-seq data, especially at distant enhancer regions. However, the

initial enhancer-gene mapping is arbitrary and the ambiguity of mapping each enhancer region to 20 target genes is very high. There are also some other data types that can be used to facilitate the identification of enhancer-gene promoter interactions, e.g. DNase-seq and ChIA-PET. DNase-seq technology is a more general measurement of binding signals along the whole genome compared to ChIP-seq [121]. In one DNase-seq experiment, binding signals of multiple TFs can be simultaneously captured. At an enhancer region, if we observe a binding signal enrichment in DNase-seq data and in the meanwhile, we observe binding signal enrichment in one or a few promoter regions within the scale of 10 upstream genes and 10 downstream genes, we can match them together as enhancer-promoter pairs. Using DNase-seq data we can greatly narrow down the search space of BICORN. The ChIA-PET technology provides DNA 3D structure measurement [122], particular for the enhancer-promoter interactions activated by a specific TF, which is usually the major activator of enhancer regions. Experimentally measured enhancer-promoter interactions by ChIA-PET can be used as an input to BICORN for functional enhancer-promoter interaction prediction.

Current design of BICORN is under a Bayesian framework. The Bayesian design clearly models the dependency between hidden variable, brings convenience to extend the method, and includes more variables which can be measured using other types of data. Raw signals in DNase-seq or ChIA-PET can be modeled in a probabilistic way to provide prior probabilities for candidate enhancer-gene mappings. These data sets can greatly lower the false positive rate in the initial enhancer-gene mapping network, narrow down the search space of BICORN, speed up the convergence of Gibbs sampling and finally improve the prediction accuracy of the proposed algorithm.

5.2.2 Functional genomic mutation identification

Identify somatic intra or inter-chromosome translations

The proposed PSSV method is mainly used to predict somatic deletions, insertions and inversions. Structural variation is not limited to these three types, but also includes intra or inter-chromosome translocations that frequently occur in tumor cells. Such type of SVs can be identified using discordant reads with paired ends mapped to distant locations in the same chromosome (for intra) or two different chromosomes (for inter). We check the distribution of

discordant read count associated with inter or intra chromosome translations and find that it can also be modelled using a mixture of Poisson distributions but with only two components: 'non-mutation' and 'heterozygous'. The possibility that both copies of chromosome are simultaneously translocated is very low. To detect more complex mutation patterns, a mixture of Poisson distributions (of two components) can be used to model discordant read counts in each tumor or normal sample and each chromosome translocation is modeled as a mixture of three hidden states as previously defined: State 1 (non-mutation), State 2 (somatic heterozygous mutation) and State 4 (germline heterozygous mutation). Only State 2 represents the somatic mutation status.

Jointly modelling multiple samples for recurrent somatic SV identification

Current design of PSSV is for somatic SV detection using one pair of tumor and normal samples. With the accumulation of WGS data, for each cancer type, we can now collect tens or hundreds of sample pairs. Although somatic mutation is quite diverse, there are some recurrent mutations serving as main drivers in cancer development [123]. Therefore, it is necessary to further extend the current framework of PSSV to identify recurrent somatic mutations across multiple samples. A possible solution is to add another layer over current framework and statistically calculate a probability for each SV based on its posterior probability of somatic state under each sample. Then, a SV predicted as somatic with higher posterior probabilities in more samples will be also ranked higher in the final prediction list. For each recurrent somatic SV at gene coding, promoter or enhancer regions, we can map it to the target gene and examine gene expression change between the sample set containing current SV and the other sample set without current SV. Functional enrichment analysis on differentially expressed genes containing driver mutations would provide insights about what functions or signaling pathways are driven by the somatic SVs.

5.3 Conclusions

In this dissertation research, we have developed novel Bayesian methods to infer transcriptional regulation mechanisms. Using distribution learning, we achieve the final predictions of foreground signals from each type of NGS data. Specific for promoter and enhancer regions, we have proposed or developed (1) a Gaussian mixture model to detect

transcription factor binding sites; (2) a Gibbs sampling framework to infer cis-regulatory modules; (3) a Poisson mixture model to identify somatic structural variations. We have demonstrated the advantages of the proposed methods over existing approaches using simulated NGS data and further applied our methods to real NGS data of breast cancer. Results of each method have been initially validated using biological experiments, independent data recourses or available databases.

Appendix A. Journal manuscript in preparation and conference publication

Manuscript in preparation

X. Chen, J. Gu, X. Wang, A. N. Shajahan-Haq, L. Hilakivi-Clarke, R. Clarke, and J. Xuan, "CRNET: An efficient sampling approach to infer regulatory networks by integrating ChIP-seq and RNA-seq data," to be submitted to ***Bioinformatics.***

X. Chen, J. Xuan, et al., "ChIP-GSM: A ChIP-seq data based Gibbs Sampling approach for cis-regulatory Module inference," in preparation.

J. Jung, **X. Chen**, et al., "Defining NOTCH3 transcriptional network in breast and ovarian cancer cells by genome-wide global mapping of Notch3 binding sites," in preparation.

Conference publication

A. N. Shajahan-Haq, L. Jin, A. K. Cheema, S. M. Boca, Y. Gusev, K. Bhuvaneshwar, D. M. Demas, H. Ressom, R. Michalek, **X. Chen**, J. Xuan, S. Madhavan and Robert Clarke. "Abstract B1-23: Early growth response (EGR1) is a critical regulator of cellular metabolism and predicts increased responsiveness to antiestrogens in breast cancer," Cancer Research, November 15, 2015 75:B1-23.

X. Chen, Xu Shi, A. N. Shajahan-Haq, L. Hilakivi-Clarke, and R. Clarke and J. Xuan, "BSSV: Bayesian based Somatic Structural Variation identification with whole genome DNA-Seq data," in Proc. of the 36th International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Chicago, IL, USA, August 2014.

X. Chen, X. Shi, A. N. Shajahan-Haq, L. Hilakivi-Clarke, R. Clarke, and J. Xuan. "Statistical identification of co-regulatory gene modules using multiple ChIP-Seq experiments," In Proc. of the International Conference on Bioinformatics Models, Methods and Algorithms, pages 109-116, Angers, France, March 2014.

X. Chen, J. Xuan, X. Shi, A. N. Shajahan-Haq, L. Hilakivi-Clarke, and R. Clarke. "A novel statistical approach to identify co-regulatory gene module," In Proc. of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 16-18, 2013.

X. Chen, C. Wang, A. N. Shajahan, R. B. Riggins, R. Clarke, and J. Xuan. "Reconstruction of transcription regulatory networks by stability-based network component analysis," ISBRA 2012, Lecture Notes in Computer Science, Volume 7292, pp. 36-47, 2012.

Appendix B. ChIP-BIT model parameter estimation

By introducing Jensen's inequality with $\sum_{i=0,1} \hat{a}_{n,w,i} = 1$ to Eq. (2.12), we obtain the following inequality:

$$\begin{aligned} &\log P(\ ,\ _{TFBS},\ ^2_{TFBS},\ |\mathbf{s},\mathbf{d}) \\ =\ &\sum_{n\ \ w} \log \sum_{b_{n,w}=0,1} P(s_{n,w}\,|\,b_{n,w})P(d_{n,w}\,|\,b_{n,w})\ ^{b_{n,w}}(1\ \ )^{(1\ b_{n,w})}\ \ \log P(\boldsymbol{\pi})\ \ P(\ ^2_{TFBS}) + \log C_2 \\ &\sum_{n\ \ w}\ \sum_{b_{n,w}=0,1} \hat{a}_{n,w} \log \frac{1}{\hat{a}_{n,w}} P(s_{n,w}\,|\,b_{n,w})P(d_{n,w}\,|\,b_{n,w})\ ^{b_{n,w}}(1\ \ )^{(1\ b_{n,w})}\ \ \log P(\boldsymbol{\pi})\ \ P(\ ^2_{TFBS}) + \log C_2. \end{aligned} \tag{B-1}$$

Estimate $_i$

Extending (S-5) and excluding items independent of $_i$, we obtain $f(\boldsymbol{\pi})$ as

$$f(\boldsymbol{\pi}) = \sum_{n\ \ w\ \ i} \hat{a}_{n,w,i} \log\ _i\ \ \log P(\boldsymbol{\pi}),\quad \sum_i\ _i = 1. \tag{B-2}$$

Minimizing the upper bound of Eq. (B-1) is equivalent to finding $_i$ to meet $\frac{f(\boldsymbol{\pi})}{\ _i} = 0$. To address the constraint of $\sum_i\ _i = 1$, we introduce a Lagrange parameter to Eq. (B-2) as:

$$f(\boldsymbol{\pi}) = \sum_{n\ \ w\ \ i} \hat{a}_{n,w,i} \log\ _i\ \ \log P(\boldsymbol{\pi}) + \ (\ \ \sum_i\ _i\ \ 1). \tag{B-3}$$

The derivative of Eq. (B-3) in term of $_i$ is computed as:

$$\frac{f}{\ _i} = \sum_{n\ \ w} \hat{a}_{n,w,i} \frac{1}{\ _i}\ \ (\ _i\ \ 1)\frac{1}{\ _i} + \ \ = 0. \tag{B-4}$$

Since $\sum_i\ _i = 1$ and $\sum_i \hat{a}_{n,w,i} = 1$, we sum Eq. (B-4) using the value(s) of i and obtain $= T + \ _1 + \ _2\ \ 2$. Bring back to Eq. (B-4), we can estimate $^*_i$ as:

$$_i = \frac{\hat{a}_{n,w,i} + (\ _i \quad 1)}{T + \ _1 + \ _2 \quad 2}. \tag{B-5}$$

Estimate $_{TFBS}$

Extending Eq. (B-1) and excluding items independent on $_{TFBS}$, we obtain $f(\ _{TFBS})$ as defined in the following equation:

$$f(\ _{TFBS}) = \quad _{n \quad w} \hat{a}_{n,w,1} \quad \log \frac{1}{\sqrt{2}\ _{TFBS}} \quad \frac{1}{2}\left(\frac{s_{n,w} \quad _{TFBS}}{\ _{TFBS}}\right)^2 \quad . \tag{B-6}$$

Minimizing the upper bound of Eq. (B-6) is equivalent to finding $_{TFBS}$ to meet $\frac{df(\ _{TFBS})}{d\ _{TFBS}} = 0$. The derivative of Eq. (B-6) in terms of $_{TFBS}$ is computed as:

$$\frac{df(\ _{TFBS})}{d\ _{TFBS}} = \quad _{n \quad w} \hat{a}_{n,w,1} \frac{1}{\ _{TFBS}^2}\left(\ _{TFBS} \quad s_{n,w}\right) = 0. \tag{B-7}$$

Finally, we can estimate $_{TFBS}$ as:

$$_{TFBS} = \quad _{n \quad w} \hat{a}_{n,w,1} s_{n,w} \Big/ \quad _{n \quad w} \hat{a}_{n,w,1}. \tag{B-8}$$

Estimate $_{TFBS}$

Extending Eq. (B-1) and excluding items independent on $_{TFBS}$, we obtain $f(\ _{TFBS})$ as defined in the following equation:

$$f(\ _{TFBS}) = \quad _{n \quad w} \hat{a}_{n,w,1} \log P(s_{n,w} \mid b_{n,w} = 1, \ _{TFBS}, \ _{TFBS}^2) \quad \log P(\ _{TFBS}^2). \tag{B-9}$$

Minimizing the upper bound of Eq. (B-9) is equivalent to finding $_{TFBS}$ to meet $\frac{df(\ _{TFBS})}{d\ _{TFBS}} = 0$. The derivative of Eq. (B-9), relative to $_{TFBS}$, is computed as:

$$\frac{1}{\ _{TFBS}^{3}}\ \ \ \hat{a}_{n,w,1}(s_{n,w}\ \ \ _{TFBS})^2 + \frac{1}{\ _{TFBS}}\ \ \ \hat{a}_{n,w,1} + 2(\ \ +1)\frac{1}{\ _{TFBS}}\ \ 2\frac{}{\ _{TFBS}^{3}} = 0. \quad \text{(B-10)}$$

Finally, we can estimate $_{TFBS}^{2}$ as

$$_{TFBS}^{2} = \frac{2\ \ +\ \ \ \ \hat{a}_{n,w,1}(s_{n,w}\ \ \ _{1})^2}{(2\ \ +2+\ \ \ \ \hat{a}_{n,w,1})}. \quad \text{(B-11)}$$

Estimate specifically for TFBS prediction at promoter regions

To identify TFBSs at promoter regions, we need estimate the exponential distribution parameter . Extending Eq. (B-1) and excluding items independent on , we obtain $f(\ \ )$ as

$$\begin{aligned} f(\ \ ) &= \ \ \ \hat{a}_{n,w,1}\log P(d_{n,w} \mid b_{n,w} = 1,\ \ ) \\ &= \ \ \ \hat{a}_{n,w,1}\log\ \frac{1}{2}\exp\big(\ \ \ |w|\ \ d\big)\big(1\ \ \exp(\ \ \ \ d)\big)\ . \end{aligned} \quad \text{(B-12)}$$

Minimizing the upper bound of Eq. (B-12) is equivalent to finding to meet $\frac{df(\ \ )}{d} = 0$. The derivative of Eq. (B-12), relative to , is computed as follows:

Let $x = \exp(\ \ \ \ d)$, then we have $f(x) = \ \ \ \hat{a}_{n,w,1}\log\ \frac{1}{2}x^{|w|}\big(1\ \ x\big)\ .$

$$\frac{df(x)}{dx} = \ \ \ \hat{a}_{n,w,1}|w|\frac{1}{x} + \ \ \ \hat{a}_{n,w,1}\frac{1}{1\ \ x} = 0. \quad \text{(B-13)}$$

$$x = \frac{\hat{a}_{n,w,1}\,|w|}{\hat{a}_{n,w,1} + \hat{a}_{n,w,1}\,|w|} \qquad \text{(B-14)}$$

Finally, we can estimate as

$$= \frac{1}{d}\ln\frac{1}{x} = \frac{1}{d}\ln\frac{\hat{a}_{n,w,1} + \hat{a}_{n,w,1}\,|w|}{\hat{a}_{n,w,1}\,|w|} \quad . \qquad \text{(B-15)}$$

Appendix C. BICORN model parameter estimation

A hierarchical model with previously defined variables is shown Fig. C.1. The most important variable is CRM index c_n, a row index of the candidate module matrix B. c_n controls which binding ($b_{c_n,t}$) is 'functional' on the n-th target gene. For each binding event $b_{c_n,t}$, using ChIP-BIT we can calculate a prior probability as $_{n,t}$. If $b_{c_n,t}$ = 1, it is a functional binding and will regulate target gene expression with certain strength. Then the activity of the t-th TF will affect the abundance of gene expression profiling through the TF-gene interaction. Sometimes even without any TF regulation, a gene still has a certain amount of sRNA. We call this expression as 'baseline' expression and model it in the proposed framework. There is always data noise in the measured gene expression data, and we define the noise power as ^{2}.

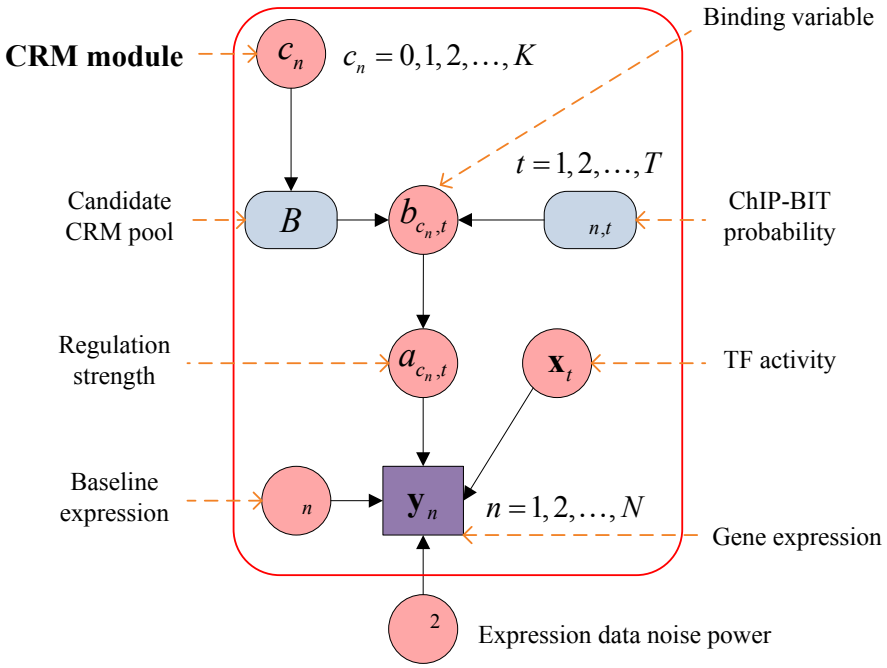


Figure C. 1 A hierarchical model of defined hidden variables in BICORN.

Considering above dependency relationship between hidden variables, Gibbs sampling based on conditional probability calculation is an optimal approach to estimate each variable and finally learn the distribution of CRMs. Note, each CRM is a unique combination of TFs and the distribution of CRMs is actually a joint distribution of multiple TFs. It is hard to model a joint

distribution using a particular distribution shape. Thence, we need generated enough samples for CRMs and learn the distribution from data directly.

The posterior probability defined in Eq. (3-4) can be further extended as follows:

$$
\begin{aligned}
P\left(\mathbf{A},\mathbf{C},\boldsymbol{\eta},\mathbf{X},\ {}^{2}\middle|\mathbf{Y},\mathbf{B}\right)\ & P\left(\mathbf{Y}\middle|\mathbf{A},\mathbf{C},\mathbf{B},\boldsymbol{\eta},\mathbf{X},\ {}^{2}\right)\cdot P\left(\mathbf{A},\mathbf{C},\boldsymbol{\eta},\mathbf{X},\ {}^{2}\right) \\
& {}_{n} P\left(\mathbf{y}_n\middle|\mathbf{a}_n,c_n,\mathbf{B},\ {}_n,\mathbf{X},\ {}^{2}\right)\cdot P(\mathbf{a}_n)\cdot P(c_n)\cdot P(\ {}_n)\cdot P(\mathbf{X})\cdot P\left(\ {}^{2}\right) \\
& {}_{n}\ {}_{m}\left(\ {}^{1}\right)\exp\ \ \frac{1}{2\ {}^{2}}\ \log(y_{n,m})\ \ {}_{t} a_{c_n,t} b_{c_n,t} x_{t,m}\ \ {}_{n}\ {}^{2} \\
& \cdot\ {}_{n}\ {}_{t}\left(\ {}_{a}^{1}\right)\exp\ \ \frac{1}{2\ {}_{a}^{2}} a_{c_n,t}^{2}\ \cdot\ {}_{n} P(c_n) \\
& \cdot\ {}_{n}\left(\ {}^{1}\right)\exp\ \ \frac{1}{2\ {}^{2}}\ {}_{n}^{2}\ \cdot\ {}_{t}\ {}_{m}\left(\ {}_{x}^{1}\right)\exp\ \ \frac{1}{2\ {}_{x}^{2}} x_{t,m}^{2} \\
& \cdot\left(\ {}^{2}\right)\ {}^{1}\exp\ \ \frac{}{2}\ .
\end{aligned}
\tag{C-1}
$$

Step 1: Gibbs sampling of $a_{c_n,t}$

For each gene, if $b_{c_n,t}=1$, we sample $a_{c_n,t}$ by considering $a_{c_n,t}\sim N(0,\ {}_{a}^{2})$; if $b_{c_n,t}=0$, we sample $a_{c_n,t}=0$.

$$
\begin{aligned}
& P\left(a_{c_n,t}\middle|b_{c_n,t}=1,\mathbf{y}_n,\mathbf{X},\ {}_n,a_{c_n,t'\ t},\ {}^{2}\right) \\
& {}_{m}\ \frac{1}{}\frac{1}{{}_{a}}\exp\ \ \frac{1}{2\ {}^{2}}\ \log(y_{n,m})\ \ {}_{t} a_{c_n,t} b_{c_n,t} x_{t,m}\ \ {}_{n}\ {}^{2}\ \ \frac{1}{2\ {}_{a}^{2}} a_{c_n,t}^{2}\ .
\end{aligned}
\tag{C-2}
$$

The posterior distribution of $a_{c_n,t}$ is a Gaussian distribution with mean and variance parameters as follows:

$$
{}_{a}=\frac{{}_{a}^{2}\ \ {}_{m}\ (\log(y_{n,m})\ \ {}_{t'\ t} a_{c_n,t'} b_{c_n,t'} x_{t',m}\ \ {}_{n}) x_{t,m}}{{}_{a}^{2}\ \ {}_{m} x_{t,m}^{2}+M\ {}^{2}},\quad {}_{a}^{2}=\frac{{}_{a}^{2} M\ {}^{2}}{{}_{a}^{2}\ \ {}_{m} x_{t,m}^{2}+\ {}^{2} M}.
\tag{C-3}
$$

Note: we need iteratively sample $a_{c_n,t}$ for $t=1\sim T$ because this probability is conditional on all $a_{c_n,t'}(t'\ \ t)$. We assign non-informative large value to hyper-parameter ${}_{a}$(e.g., ${}_{a}=100$) to assume no knowledge of the model parameter.

Step 2: Gibbs sampling of $x_{t,m}$

$$P\left(x_{t,m}\middle|\mathbf{Y},\mathbf{C},\mathbf{A},x_{t'\ t,m},\boldsymbol{\eta},\ ^2\right)$$
$$_n\ \frac{1}{\ }\frac{1}{\ _x}\exp\ \ \frac{1}{2\ ^2}\ \log(y_{n,m})\ \ _t\ a_{c_n,t}b_{c_n,t}x_{t,m}\ \ \ _n\ ^2\ \ \frac{1}{2\ _x^2}x_{t,m}^2\ . \quad \text{(C-4)}$$

The posterior distribution of $x_{t,m}$ is a Gaussian distribution with mean and variance parameters as follows:

$$_x=\frac{^2_x\ \ (\log(y_{n,m})\ \ \ _{t'\ t}\ a_{c_n,t'}b_{c_n,t'}x_{t',m}\ \ \ _n)a_{c_n,t}b_{c_n,t}}{^2_x\ \ _n\ a_{c_n,t}^2b_{c_n,t}^2+\ ^2N},\quad _x^2=\frac{^2N\ \ _x^2}{^2_x\ \ _n\ a_{c_n,t}^2b_{c_n,t}^2+\ ^2N}. \quad \text{(C-5)}$$

Note: we iteratively sample $x_{t,m}$ because this probability is condition on $x_{t',m}(t'\ \ t)$. For the hyper parameter ^2_x, we set an informative prior on x with $^2_x=1$.

Step 3: Gibbs sampling of $_n$

$$P\left(\ _n\middle|\mathbf{y}_n,\mathbf{X},c_n,\mathbf{a}_n,\ ^2\right)$$
$$_m\ \frac{1}{\ }\frac{1}{\ }\exp\ \ \frac{1}{2\ ^2}\ \log(y_{n,m})\ \ _t\ a_{c_n,t}b_{c_n,t}x_{t,m}\ \ \ _n\ ^2\ \ \frac{1}{2\ ^2}\ _n^2\ . \quad \text{(C-6)}$$

The posterior distribution of $_n$ is a Gaussian distribution with mean and variance parameters as follows:

$$=\frac{^2\ \ (\log(y_{n,m})\ \ \ _t\ a_{c_n,t}b_{c_n,t}x_{t,m})}{M(\ ^2+\ ^2)},\quad ^2=\frac{^2\ \ ^2}{^2+\ ^2} \quad \text{(C-7)}$$

We use non-informative prior for $_n$ by setting hyper-parameter $=100$.

Step 4: Gibbs sampling of ^{2}

$$P\left(\ ^2 \middle| \mathbf{Y}, \mathbf{A}, \mathbf{C}, \boldsymbol{\eta}, \mathbf{X} \right)$$

$$_n \ _m \left(\ ^1 \right) \exp \ \frac{1}{2 \ ^2} \log(y_{n,m}) \ _t a_{c_n,t} b_{c_n,t} x_{t,m} \ _n \ ^2 \cdot \left(\ ^2 \right) \ ^1 \exp \ \frac{}{2} . \quad \text{(C-8)}$$

The prior distribution of ^{2} is assumed as inverse Gamma distribution, which is the conjugate prior of its likelihood function (Gaussian distribution), therefore the posterior distribution of ^{2} also follows inverse Gamma:

$$a' = \ + \frac{1}{2}, \quad ' = \ + \frac{1}{2MN} \ _n \ _m \log(y_{n,m}) \ _t a_{c_n,t} b_{c_n,t} x_{t,m} \ _n \ ^2 . \quad \text{(C-9)}$$

where N is the number of genes and M is the total number of gene expression experiments. We set hyper-parameters $\ = 1$ and $\ = 1$ to make the prior distribution of $\frac{1}{2}$ non-informative.

Step 5: Gibbs sampling of c_n

$$P\left(c_n \middle| \mathbf{y}_n, \mathbf{X}, \ _n, \mathbf{a}_n, \ ^2 \right) \ _t P\left(\mathbf{y}_n, \mathbf{x}, \ _n, \ ^2 \middle| a_{c_n,t} \right) P\left(a_{c_n,t} \middle| c_n \right) P\left(c_n \right)$$

$$= \ _t \frac{1}{\sqrt{2}} \exp \ \frac{1}{2 \ ^2} \ _m \log(y_{n,m}) \ _t a_{c_n,t} b_{c_n,t} x_{t,m} \ _n \ ^2 \quad \text{(C-10)}$$

$$\cdot \ _t \frac{1}{\sqrt{2}} \ _a \exp \ \frac{1}{2 \ _a^2} (a_{c_n,t})^2 \cdot \ _t \ _{n,t}^{b_{c_n,t}} (1 \ _{n,t})^{1 \ b_{c_n,t}} .$$

We calculate above probability for all candidate values of c_n including *0,1,2,...,K*. Then, we sample c_n according to

$$p\left(c_n = k \right) = \frac{P\left(c_n = k \middle| \mathbf{y}_n, \mathbf{X}, \ _n, \mathbf{a}_n, \ ^2 \right)}{_j \ P\left(c_n = j \middle| \mathbf{y}_n, \mathbf{X}, \ _n, \mathbf{a}_n, \ ^2 \right)} . \quad \text{(C-11)}$$

Appendix D. Comparison of active TFs at promoter and enhancer regions

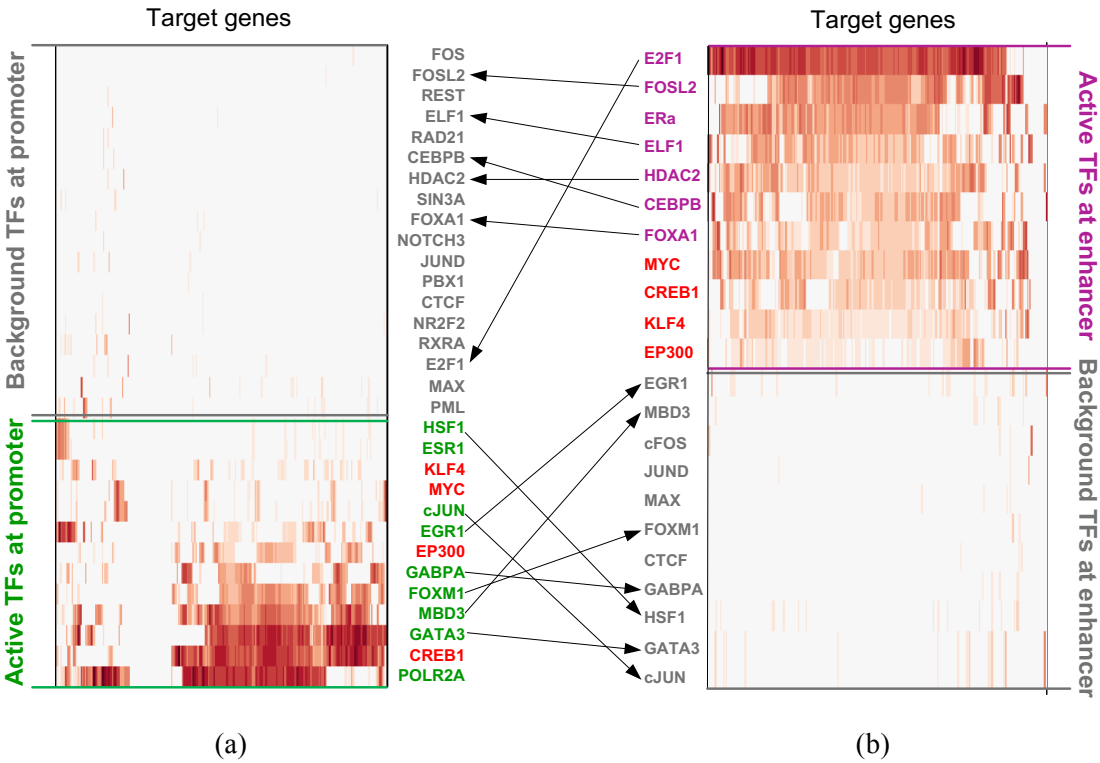


Figure D.1 TFs functional at promoter or enhancer regions of E2 responsive target genes in breast cancer MCF-7 cells. (a) Foreground (green) and background (grey) TFs at promoter regions; (b) foreground (purple) and background (grey) TFs at enhancer regions. Common foreground TFs at both types of regions are labeled as 'red'.

Appendix E. PSSV model parameter estimation

The posterior probability $P(\boldsymbol{\pi}, \boldsymbol{\lambda} \mid \mathbf{K})$ can be extended as follows:

$$\begin{aligned}
\log P(\boldsymbol{\pi}, \boldsymbol{\lambda} \mid \mathbf{K}) &= \log \frac{1}{C} P(\mathbf{K}, \boldsymbol{\lambda} \mid \boldsymbol{\pi}) P(\boldsymbol{\pi}) \\
&= \log \frac{1}{C} \prod_{n} P(\mathbf{k}_n \mid \boldsymbol{\lambda}) P(\boldsymbol{\lambda}) P(\boldsymbol{\pi}) \\
&= \log \frac{1}{C} \prod_{n} \prod_{j,i} \prod_{j} \prod_{(i,j)} \prod_{m=1,2} P(k_{n,m} \mid i) P(\quad_{m,i}) \prod_{m=3,4} P(k_{n,m} \mid j) P(\quad_{m,j}) \quad P(\boldsymbol{\pi}) \\
&= \log C \prod_{n} \log \prod_{j,i} \prod_{j} \prod_{(i,j)} \prod_{m=1,2} P(k_{n,m} \mid i) P(\quad_{m,i}) \prod_{m=3,4} P(k_{n,m} \mid j) P(\quad_{m,j}) \quad \log P(\boldsymbol{\pi}).
\end{aligned} \tag{E-1}$$

By constraining $\sum_{j,i} \sum_{j} a_{n,(i,j)} = 1$ and applying Jensen's inequality to Eq. (E-1), we can obtain the following inequality:

$$\begin{aligned}
\log P(\boldsymbol{\pi}, \boldsymbol{\lambda} \mid \mathbf{K}) &= \log C \prod_{n} \log \prod_{j,i} \prod_{j} \frac{a_{n,(i,j)}}{a_{n,(i,j)}} \prod_{(i,j)} \prod_{m=1,2} P(k_{n,m} \mid i) P(\quad_{m,i}) \prod_{m=3,4} P(k_{n,m} \mid j) P(\quad_{m,j}) \quad \log P(\boldsymbol{\pi}) \\
&\log C \prod_{n} \prod_{j,i} \prod_{j} a_{n,(i,j)} \log \frac{(i,j)}{a_{n,(i,j)}} \prod_{m=1,2} P(k_{n,m} \mid i) P(\quad_{m,i}) \prod_{m=3,4} P(k_{n,m} \mid j) P(\quad_{m,j}) \quad \log P(\boldsymbol{\pi}) \\
&= \log C \prod_{n} \prod_{j,i} \prod_{j} a_{n,(i,j)} \log \prod_{(i,j)} \prod_{n} \prod_{j,i} \prod_{j} a_{n,(i,j)} \log \prod_{m=1,2} P(k_{n,m} \mid i) P(\quad_{m,i}) \prod_{m=3,4} P(k_{n,m} \mid j) P(\quad_{m,j}) \\
&+ \prod_{n} \prod_{j,i} \prod_{j} a_{n,(i,j)} \log a_{n,(i,j)} \quad \log P(\boldsymbol{\pi}).
\end{aligned} \tag{E-2}$$

Maximizing the equation in Eq. (E-1) is equivalent to minimizing the upper bound of the inequality in Eq. (E-2). Thus, we estimate each parameter based on Eq. (E-2) using an iterative Expectation-Maximization (EM) approach.

Estimate $a_{n,(i,j)}$

To minimize the upper bound of the above inequality, with $\sum_{j,i} \sum_{j} a_{n,(i,j)} = 1$, we seek $a_{n,(i,j)}$ to meet the equality condition.

$$\log \frac{{}^{(i,j)}}{a_{n,(i,j)}} \quad {}_{m=1,2} P(k_{n,m} \mid i) P({}_{m,i}) \quad {}_{m=3,4} P(k_{n,m} \mid j) P({}_{m,j})$$
$$= \log {}_{j,i \quad j} \quad {}^{(i,j)} \quad {}_{m=1,2} P(k_{n,m} \mid i) P({}_{m,i}) \quad {}_{m=3,4} P(k_{n,m} \mid j) P({}_{m,j}) \quad . \tag{E-3}$$

Then, $a_{n,(i,j)}$ can be obtained as:

$$a_{n,(i,j)} = \frac{{}^{(i,j)} \quad {}_{m=1,2} P(k_{n,m} \mid i) P({}_{m,i}) \quad {}_{m=3,4} P(k_{n,m} \mid j) P({}_{m,j})}{{}_{j,i \quad j} \quad {}^{(i,j)} \quad {}_{m=1,2} P(k_{n,m} \mid i) P({}_{m,i}) \quad {}_{m=3,4} P(k_{n,m} \mid j) P({}_{m,j})}. \tag{E-4}$$

Estimate ${}_{(i,j)}$

Searching for ${}_{(i,j)}$ to minimize the upper bound of Eq. (E-2) is equivalent to minimizing $f(\boldsymbol{\pi}) = {}_{n \quad j,i \quad j} \hat{a}_{n,(i,j)} \log {}_{(i,j)} \quad \log P(\boldsymbol{\pi})$ by deleting those unrelated items in Eq. (E-2), with the constraint of ${}_{j,i \quad j} {}^{(i,j)} = 1$.

We define the Lagrangian equation L by introducing Lagrange parameter as follows:

$$L = {}_{n \quad j,i \quad j} \hat{a}_{n,(i,j)} \log {}_{(i,j)} \quad \log P(\boldsymbol{\pi}) + \quad ({}_{j,i \quad j} {}^{(i,j)} \quad 1). \tag{E-5}$$

Then, searching for ${}_{(i,j)}$ to minimize the upper bound of Eq. (E-2) is equivalent to calculating ${}_{(i,j)}$ to meet the equation condition of the following equation.

$$\frac{L}{{}_{(i,j)}} = \quad {}_{n} \hat{a}_{n,(i,j)} \frac{1}{{}_{(i,j)}} \quad ({}_{(i,j)} \quad 1) \frac{1}{{}_{(i,j)}} + \quad = 0. \tag{E-6}$$

Considering the constraints of ${}_{j,i \quad j} {}^{(i,j)} = 1$ and ${}_{j,i \quad j} a_{n,(i,j)} = 1$, we sum up Eq. (E-6) in term of (i, j) and finally get the solution of Lagrange parameter as

$$= N + \sum_{j,i} \left(\log_{(i,j)} - 1 \right). \tag{E-7}$$

We bring back to Eq. (E-6) and update $\log_{(i,j)}$ with previously updated $\hat{a}_{n,(i,j)}$ as follows:

$$\log_{(i,j)} = \frac{\sum_{n} \hat{a}_{n,(i,j)} + \left(\log_{(i,j)} - 1 \right)}{N + \sum_{j,i>j} \left(\log_{(i,j)} - 1 \right)}. \tag{E-8}$$

Estimate $\log_{m,i}$ and $\log_{m,j}$

Searching for $\log_{m,i}$ and $\log_{m,j}$ to minimize the upper bound of Eq. (E-2) is equivalent to minimizing $f(\boldsymbol{\lambda}) = \sum_{n} \sum_{j,i} \hat{a}_{n,(i,j)} \log \frac{\prod_{m=1,2} P(k_{n,m} \mid i) P(\log_{m,i})}{\prod_{m=3,4} P(k_{n,m} \mid j) P(\log_{m,j})}$ by deleting those unrelated items in Eq. (E-2).

Function $f(\boldsymbol{\lambda})$ can further extended as follows:

$$\begin{aligned} f(\boldsymbol{\lambda}) = \sum_{n} \sum_{j,i} \hat{a}_{n,(i,j)} & \log\left(P(k_{n,1} \mid i) P(\log_{1,i}) \right) + \log\left(P(k_{n,2} \mid i) P(\log_{2,i}) \right) \\ & + \log\left(P(k_{n,3} \mid j) P(\log_{3,j}) \right) + \log\left(P(k_{n,4} \mid j) P(\log_{4,j}) \right) . \end{aligned} \tag{E-9}$$

Now, searching for $\log_{1,i}$ to minimize the upper bound of Eq. (E-2) is equivalent to calculating $\log_{1,i}$ to meet the equation condition of the following equation.

$$\frac{f(\boldsymbol{\lambda})}{\log_{1,i}} = \frac{\sum_{n} \sum_{0} \sum_{j} \hat{a}_{n,(i,j)} \log\left(P(k_{n,1} \mid i) P(\log_{1,i}) \right)}{\log_{1,i}} = 0 \tag{E-10}$$

Then, $\log_{1,i}$ can be updated as:

$$\log_{1,i} = \sum_{n} \sum_{j=0}^{i} \hat{a}_{n,(i,j)} \left(k_{n,1} - 1 + \log_{1,i} \right) / \sum_{n} \sum_{j=0}^{i} \hat{a}_{n,(i,j)} \left(1 + \log_{1,i} \right). \tag{E-11}$$

Similarly, all $_{m,i}$ or $_{m,j}$ for $m = 1 \sim 4$ can be updated as:

$$
\begin{aligned}
_{m,i} &= \sum_{n} \sum_{j=0}^{i} a_{n,(i,j)} \left(k_{n,m} \quad 1 + \quad _{m,i} \right) / \sum_{n} \sum_{j=0}^{i} a_{n,(i,j)} \left(1 + \quad _{m,i} \right) \\
_{m,j} &= \sum_{n} \sum_{i=j}^{2} a_{n,(i,j)} \left(k_{n,m} \quad 1 + \quad _{m,j} \right) / \sum_{n} \sum_{i=j}^{2} a_{n,(i,j)} \left(1 + \quad _{m,j} \right)
\end{aligned}
\qquad \text{(E-12)}
$$

Bibliography

[1] H. Kitano, "Computational systems biology," *Nature,* vol. 420, pp. 206-10, Nov 14 2002.

[2] J. P. Mathew, B. S. Taylor, G. D. Bader, S. Pyarajan, M. Antoniotti, A. M. Chinnaiyan*, et al.*, "From bytes to bedside: data integration and computational biology for translational cancer research," *PLoS Comput Biol,* vol. 3, p. e12, Feb 23 2007.

[3] C. Tomasetti and B. Vogelstein, "Variation in cancer risk among tissues can be explained by the number of stem cell divisions," *Science,* vol. 347, pp. 78-81, Jan 2 2015.

[4] D. M. Rubio, E. E. Schoenbaum, L. S. Lee, D. E. Schteingart, P. R. Marantz, K. E. Anderson*, et al.*, "Defining Translational Research: Implications for Training," *Academic Medicine,* vol. 85, pp. 470-475, Mar 2010.

[5] E. P. Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature,* vol. 489, pp. 57-74, Sep 6 2012.

[6] R. D. Hawkins, G. C. Hon, and B. Ren, "Next-generation genomics: an integrative approach," *Nat Rev Genet,* vol. 11, pp. 476-86, Jul 2010.

[7] J. S. Reis-Filho, "Next-generation sequencing," *Breast Cancer Research,* vol. 11, 2009.

[8] J. Zhang, R. Chiodini, A. Badr, and G. F. Zhang, "The impact of next-generation sequencing on genomics," *Journal of Genetics and Genomics,* vol. 38, pp. 95-109, Mar 2011.

[9] B. Rabbani, M. Tekin, and N. Mahdieh, "The promise of whole-exome sequencing in medical genetics," *Journal of Human Genetics,* vol. 59, pp. 5-15, Jan 2014.

[10] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics,* vol. 10, pp. 57-63, Jan 2009.

[11] L. J. Core, J. J. Waterfall, and J. T. Lis, "Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters," *Science,* vol. 322, pp. 1845-1848, Dec 19 2008.

[12] S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou*, et al.*, "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia," *Genome Research,* vol. 22, pp. 1813-1831, Sep 2012.

[13] J. Ule, K. B. Jensen, M. Ruggiu, A. Mele, A. Ule, and R. B. Darnell, "CLIP identifies Nova-regulated RNA networks in the brain," *Science,* vol. 302, pp. 1212-1215, Nov 14 2003.

[14] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. Bin Mohamed*, et al.*, "An oestrogen-receptor-alpha-bound human chromatin interactome," *Nature,* vol. 462, pp. 58-64, Nov 5 2009.

[15] S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou*, et al.*, "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia," *Genome Res,* vol. 22, pp. 1813-31, Sep 2012.

[16] T. Juven-Gershon and J. T. Kadonaga, "Regulation of gene expression via the core promoter and the basal transcriptional machinery," *Developmental Biology,* vol. 339, pp. 225-229, Mar 15 2010.

[17] R. Stadhouders, A. van den Heuvel, P. Kolovos, R. Jorna, K. Leslie, F. Grosveld*, et al.*, "Transcription regulation by distal enhancers: who's in the loop?," *Transcription,* vol. 3, pp. 181-6, Jul-Aug 2012.

[18] W. Li, D. Notani, Q. Ma, B. Tanasa, E. Nunez, A. Y. Chen, *et al.*, "Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation," *Nature,* vol. 498, pp. 516-20, Jun 27 2013.

[19] L. Yao, H. Shen, P. W. Laird, P. J. Farnham, and B. P. Berman, "Inferring regulatory element landscapes and transcription factor networks from cancer methylomes," *Genome Biol,* vol. 16, p. 105, 2015.

[20] E. P. Consortium, B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, *et al.*, "An integrated encyclopedia of DNA elements in the human genome," *Nature,* vol. 489, pp. 57-74, Sep 6 2012.

[21] M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan, C. Cheng, *et al.*, "Architecture of the human regulatory network derived from ENCODE data," *Nature,* vol. 489, pp. 91-100, Sep 6 2012.

[22] J. C. Liao, R. Boscolo, Y. L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury, "Network component analysis: Reconstruction of regulatory signals in biological systems," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 100, pp. 15522-15527, Dec 23 2003.

[23] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat Rev Genet,* vol. 10, pp. 57-63, Jan 2009.

[24] P. Cahan, Y. Li, M. Izumi, and T. A. Graubert, "The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells," *Nature Genetics,* vol. 41, pp. 430-437, Apr 2009.

[25] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer, "The Database of Genomic Variants: a curated collection of structural variation in the human genome," *Nucleic Acids Research,* vol. 42, pp. D986-D992, Jan 2014.

[26] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, *et al.*, "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation," *Nature Methods,* vol. 6, pp. 677-U76, Sep 2009.

[27] S. S. Sindi, S. Onal, L. K. C. Peng, H. T. Wu, and B. J. Raphael, "An integrative probabilistic model for identification of structural variation in sequencing data," *Genome Biology,* vol. 13, 2012.

[28] T. LaFramboise, "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances," *Nucleic Acids Research,* vol. 37, pp. 4181-4193, Jul 2009.

[29] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, *et al.*, "An integrated map of structural variation in 2,504 human genomes," *Nature,* vol. 526, pp. 75-+, Oct 1 2015.

[30] J. M. Wang, C. G. Mullighan, J. Easton, S. Roberts, S. L. Heatley, J. Ma, *et al.*, "CREST maps somatic structural variation in cancer genomes with base-pair resolution," *Nature Methods,* vol. 8, pp. 652-U69, Aug 2011.

[31] A. Malhotra, M. Lindberg, G. G. Faust, M. L. Leibowitz, R. A. Clark, R. M. Layer, *et al.*, "Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms," *Genome Research,* vol. 23, pp. 762-776, May 2013.

[32] C. Cheng, R. Q. Min, and M. Gerstein, "TIP: A probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles," *Bioinformatics,* vol. 27, pp. 3221-3227, Dec 1 2011.

[33] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein*, et al.*, "Model-based analysis of ChIP-Seq (MACS)," *Genome Biol,* vol. 9, p. R137, 2008.

[34] J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson*, et al.*, "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls," *Nat Biotechnol,* vol. 27, pp. 66-75, Jan 2009.

[35] H. Xing, Y. Mo, W. Liao, and M. Q. Zhang, "Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data," *PLoS Comput Biol,* vol. 8, p. e1002613, 2012.

[36] V. Kumar, M. Muratani, N. A. Rayan, P. Kraus, T. Lufkin, H. H. Ng*, et al.*, "Uniform, optimal signal processing of mapped deep-sequencing data," *Nat Biotechnol,* vol. 31, pp. 615-22, Jul 2013.

[37] P. F. Kuan, D. J. Chung, G. J. Pan, J. A. Thomson, R. Stewart, and S. Keles, "A Statistical Framework for the Analysis of ChIP-Seq Data," *Journal of the American Statistical Association,* vol. 106, pp. 891-903, Sep 2011.

[38] D. Shlyueva, G. Stampfel, and A. Stark, "Transcriptional enhancers: from properties to genome-wide predictions," *Nature Reviews Genetics,* vol. 15, pp. 272-286, Apr 2014.

[39] C. Sabatti and G. M. James, "Bayesian sparse hidden components analysis for transcription regulation networks," *Bioinformatics,* vol. 22, pp. 739-746, Mar 15 2006.

[40] G. Chen, S. T. Jensen, and C. J. Stoeckert, "Clustering of genes into regulons using integrated modeling-COGRIM," *Genome Biology,* vol. 8, 2007.

[41] C. Angelini and V. Costa, "Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems," *Front Cell Dev Biol,* vol. 2, p. 51, 2014.

[42] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nat Rev Genet,* vol. 7, pp. 85-97, Feb 2006.

[43] J. L. Freeman, G. H. Perry, L. Feuk, R. Redon, S. A. McCarroll, D. M. Altshuler*, et al.*, "Copy number variation: new insights in genome diversity," *Genome Res,* vol. 16, pp. 949-61, Aug 2006.

[44] L. Yang, L. J. Luquette, N. Gehlenborg, R. Xi, P. S. Haseley, C. H. Hsieh*, et al.*, "Diverse mechanisms of somatic structural variations in human cancer genomes," *Cell,* vol. 153, pp. 919-29, May 2013.

[45] R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan*, et al.*, "Mapping copy number variation by population-scale genome sequencing," *Nature,* vol. 470, pp. 59-65, Feb 2011.

[46] D. C. Koboldt, R. S. Fulton, M. D. McLellan, H. Schmidt, J. Kalicki-Veizer, J. F. McMichael*, et al.*, "Comprehensive molecular portraits of human breast tumours," *Nature,* vol. 490, pp. 61-70, Oct 4 2012.

[47] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl*, et al.*, "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation," *Nat Methods,* vol. 6, pp. 677-81, Sep 2009.

[48] S. S. Sindi, S. Onal, L. C. Peng, H. T. Wu, and B. J. Raphael, "An integrative probabilistic model for identification of structural variation in sequencing data," *Genome Biol,* vol. 13, p. R22, 2012.

[49] L. Oesper, A. Mahmoody, and B. J. Raphael, "THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data," *Genome Biology,* vol. 14, 2013.

[50] J. Wang, V. V. Lunyak, and I. K. Jordan, "BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets," *Bioinformatics,* vol. 29, pp. 492-3, Feb 15 2013.
[51] A. Visel, E. M. Rubin, and L. A. Pennacchio, "Genomic views of distant-acting enhancers," *Nature,* vol. 461, pp. 199-205, Sep 10 2009.
[52] N. Heidari, D. H. Phanstiel, C. He, F. Grubert, F. Jahanbani, M. Kasowski*, et al.*, "Genome-wide map of regulatory interactions in the human genome," *Genome Res,* vol. 24, pp. 1905-17, Dec 2014.
[53] C. Cheng, R. Min, and M. Gerstein, "TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles," *Bioinformatics,* vol. 27, pp. 3221-7, Dec 1 2011.
[54] E. G. Giannopoulou and O. Elemento, "Inferring chromatin-bound protein complexes from genome-wide binding assays," *Genome Res,* vol. 23, pp. 1295-306, Aug 2013.
[55] W. Sikora-Wohlfeld, M. Ackermann, E. G. Christodoulou, K. Singaravelu, and A. Beyer, "Assessing computational methods for transcription factor target gene identification based on ChIP-seq data," *PLoS Comput Biol,* vol. 9, p. e1003342, 2013.
[56] H. Xu, L. Handoko, X. Wei, C. Ye, J. Sheng, C. L. Wei*, et al.*, "A signal-noise model for significance analysis of ChIP-seq with negative control," *Bioinformatics,* vol. 26, pp. 1199-204, May 1 2010.
[57] D. M. Budden, D. G. Hurley, and E. J. Crampin, "Predictive modelling of gene expression from transcriptional regulatory elements," *Brief Bioinform,* Sep 16 2014.
[58] M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine*, et al.*, "Histone H3K27ac separates active from poised enhancers and predicts developmental state," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 107, pp. 21931-21936, Dec 14 2010.
[59] Z. Ouyang, Q. Zhou, and W. H. Wong, "ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells," *Proc Natl Acad Sci U S A,* vol. 106, pp. 21521-6, Dec 22 2009.
[60] M. Mokry, P. Hatzis, J. Schuijers, N. Lansu, F. P. Ruzius, H. Clevers*, et al.*, "Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes," *Nucleic Acids Res,* vol. 40, pp. 148-58, Jan 2012.
[61] C. Cheng and M. Gerstein, "Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells," *Nucleic Acids Res,* vol. 40, pp. 553-68, Jan 2012.
[62] K. C. Wong, Y. Li, C. Peng, and Z. Zhang, "SignalSpider: probabilistic pattern discovery on multiple normalized ChIP-Seq signal profiles," *Bioinformatics,* vol. 31, pp. 17-24, Jan 1 2015.
[63] N. Sun, R. J. Carroll, and H. Zhao, "Bayesian error analysis model for reconstructing transcriptional regulatory networks," *Proc Natl Acad Sci U S A,* vol. 103, pp. 7988-93, May 23 2006.
[64] J. K. Pickrell, D. J. Gaffney, Y. Gilad, and J. K. Pritchard, "False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions," *Bioinformatics,* vol. 27, pp. 2144-6, Aug 1 2011.
[65] Z. D. Zhang, J. Rozowsky, M. Snyder, J. Chang, and M. Gerstein, "Modeling ChIP sequencing in silico with applications," *PLoS Comput Biol,* vol. 4, p. e1000158, 2008.

[66] N. Yamaguchi, T. Oyama, E. Ito, H. Satoh, S. Azuma, M. Hayashi*, et al.*, "NOTCH3 signaling pathway plays crucial roles in the proliferation of ErbB2-negative human breast cancer cells," *Cancer Res,* vol. 68, pp. 1881-8, Mar 15 2008.

[67] J. T. Park, M. Shih Ie, and T. L. Wang, "Identification of Pbx1, a potential oncogene, as a Notch3 target gene in ovarian cancer," *Cancer Res,* vol. 68, pp. 8852-60, Nov 1 2008.

[68] L. Magnani, A. Stoeck, X. Zhang, A. Lanczky, A. C. Mirabella, T. L. Wang*, et al.*, "Genome-wide reprogramming of the chromatin landscape underlies endocrine therapy resistance in breast cancer," *Proc Natl Acad Sci U S A,* vol. 110, pp. E1490-9, Apr 16 2013.

[69] D. A. Cusanovich, B. Pavlovic, J. K. Pritchard, and Y. Gilad, "The functional consequences of variation in transcription factor binding," *PLoS Genet,* vol. 10, p. e1004226, Mar 2014.

[70] P. Hayward, T. Kalmar, and A. M. Arias, "Wnt/Notch signalling and information processing during development," *Development,* vol. 135, pp. 411-24, Feb 2008.

[71] G. M. Collu and K. Brennan, "Cooperation between Wnt and Notch signalling in human breast cancer," *Breast Cancer Res,* vol. 9, p. 105, 2007.

[72] X. Chen, A. Stoeck, S. J. Lee, M. Shih Ie, M. M. Wang, and T. L. Wang, "Jagged1 expression regulated by Notch3 and Wnt/beta-catenin signaling pathways in ovarian cancer," *Oncotarget,* vol. 1, pp. 210-8, Jul 2010.

[73] N. Takebe, P. J. Harris, R. Q. Warren, and S. P. Ivy, "Targeting cancer stem cells by inhibiting Wnt, Notch, and Hedgehog pathways," *Nat Rev Clin Oncol,* vol. 8, pp. 97-106, Feb 2011.

[74] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo*, et al.*, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities," *Mol Cell,* vol. 38, pp. 576-89, May 28 2010.

[75] N. Hah, C. G. Danko, L. Core, J. J. Waterfall, A. Siepel, J. T. Lis*, et al.*, "A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells," *Cell,* vol. 145, pp. 622-34, May 13 2011.

[76] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics,* vol. 26, pp. 139-40, Jan 1 2010.

[77] M. C. Peitsch and D. de Graaf, "A decade of Systems Biology: where are we and where are we going to?," *Drug Discov Today,* vol. 19, pp. 105-7, Feb 2014.

[78] C. Sabatti and G. M. James, "Bayesian sparse hidden components analysis for transcription regulation networks," *Bioinformatics,* vol. 22, pp. 739-46, Mar 15 2006.

[79] G. Chen, S. T. Jensen, and C. J. Stoeckert, Jr., "Clustering of genes into regulons using integrated modeling-COGRIM," *Genome Biol,* vol. 8, p. R4, 2007.

[80] S. Wang, H. Sun, J. Ma, C. Zang, C. Wang, J. Wang*, et al.*, "Target analysis by integration of transcriptome and ChIP-seq data with BETA," *Nat Protoc,* vol. 8, pp. 2502-15, Dec 2013.

[81] J. Qin, Y. Hu, F. Xu, H. K. Yalamanchili, and J. Wang, "Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods," *Methods,* vol. 67, pp. 294-303, Jun 1 2014.

[82] A. Stone, E. Zotenko, W. J. Locke, D. Korbie, E. K. Millar, R. Pidsley*, et al.*, "DNA methylation of oestrogen-regulated enhancers defines endocrine sensitivity in breast cancer," *Nat Commun,* vol. 6, p. 7758, 2015.

[83] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera*, et al.*, "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics,* vol. 7 Suppl 1, p. S7, 2006.

[84] X. Zhang, K. Liu, Z. P. Liu, B. Duval, J. M. Richer, X. M. Zhao*, et al.*, "NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference," *Bioinformatics,* vol. 29, pp. 106-13, Jan 1 2013.

[85] S. Pillai and S. P. Chellappan, "ChIP on Chip and ChIP-Seq Assays: Genome-Wide Analysis of Transcription Factor Binding and Histone Modifications," *Chromatin Protocols, 3rd Edition,* vol. 1288, pp. 447-472, 2015.

[86] J. C. Liao, R. Boscolo, Y. L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury, "Network component analysis: reconstruction of regulatory signals in biological systems," *Proc Natl Acad Sci U S A,* vol. 100, pp. 15522-7, Dec 23 2003.

[87] R. C. Hardison and J. Taylor, "Genomic approaches towards finding cis-regulatory modules in animals," *Nat Rev Genet,* vol. 13, pp. 469-83, Jul 2012.

[88] Z. Liu, D. Merkurjev, F. Yang, W. Li, S. Oh, M. J. Friedman*, et al.*, "Enhancer activation requires trans-recruitment of a mega transcription factor complex," *Cell,* vol. 159, pp. 358-73, Oct 9 2014.

[89] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein*, et al.*, "Mapping and analysis of chromatin state dynamics in nine human cell types," *Nature,* vol. 473, pp. 43-9, May 5 2011.

[90] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology,* vol. 15, 2014.

[91] W. B. Li, D. Notani, Q. Ma, B. Tanasa, E. Nunez, A. Y. Chen*, et al.*, "Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation," *Nature,* vol. 498, pp. 516-+, Jun 27 2013.

[92] M. Kulis and M. Esteller, "DNA Methylation and Cancer," *Epigenetics and Cancer, Pt A,* vol. 70, pp. 27-56, 2010.

[93] B. He, C. Y. Chen, L. Teng, and K. Tan, "Global view of enhancer-promoter interactome in human cells," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 111, pp. E2191-E2199, May 27 2014.

[94] J. Y. Zhang, H. M. Poh, S. Q. Peh, Y. Y. Sia, G. L. Li, F. H. Mulawadi*, et al.*, "ChIA-PET analysis of transcriptional chromatin interactions," *Methods,* vol. 58, pp. 289-299, Nov 2012.

[95] J. Konig, K. Zarnack, N. M. Luscombe, and J. Ule, "Protein-RNA interactions: new genomic technologies and perspectives," *Nat Rev Genet,* vol. 13, pp. 77-83, Feb 2011.

[96] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature,* vol. 458, pp. 719-724, Apr 9 2009.

[97] P. Medvedev, M. Stanciu, and M. Brudno, "Computational methods for discovering structural variation with next-generation sequencing," *Nat Methods,* vol. 6, pp. S13-20, Nov 2009.

[98] C. T. Saunders, W. S. Wong, S. Swamy, J. Becq, L. J. Murray, and R. K. Cheetham, "Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs," *Bioinformatics,* vol. 28, pp. 1811-7, Jul 15 2012.

[99] A. Christoforides, J. D. Carpten, G. J. Weiss, M. J. Demeure, D. D. Von Hoff, and D. W. Craig, "Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs," *BMC Genomics,* vol. 14, p. 302, 2013.

[100] G. Klambauer, K. Schwarzbauer, A. Mayr, D. A. Clevert, A. Mitterecker, U. Bodenhofer*, et al.*, "cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate," *Nucleic Acids Res,* vol. 40, p. e69, May 2012.

[101] N. Cancer Genome Atlas, "Comprehensive molecular portraits of human breast tumours," *Nature,* vol. 490, pp. 61-70, Oct 4 2012.

[102] L. Di Croce and K. Helin, "Transcriptional regulation by Polycomb group proteins," *Nat Struct Mol Biol,* vol. 20, pp. 1147-55, Oct 2013.

[103] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky*, et al.*, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res,* vol. 20, pp. 1297-303, Sep 2010.

[104] G. Klambauer, K. Schwarzbauer, A. Mayr, D. A. Clevert, A. Mitterecker, U. Bodenhofer*, et al.*, "cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate," *Nucleic Acids Research,* vol. 40, May 2012.

[105] C. Bartenhagen and M. Dugas, "RSVSim: an R/Bioconductor package for the simulation of structural variations," *Bioinformatics,* vol. 29, pp. 1679-81, Jul 1 2013.

[106] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer*, et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics,* vol. 25, pp. 2078-9, Aug 15 2009.

[107] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows-Wheeler transform," *Bioinformatics,* vol. 26, pp. 589-95, Mar 1 2010.

[108] D. J. Araten, D. W. Golde, R. H. Zhang, H. T. Thaler, L. Gargiulo, R. Notaro*, et al.*, "A quantitative measurement of the human somatic mutation rate," *Cancer Res,* vol. 65, pp. 8111-7, Sep 2005.

[109] W. Huang da, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res,* vol. 37, pp. 1-13, Jan 2009.

[110] J. M. Giltnane and J. M. Balko, "Rationale for Targeting the Ras/MAPK Pathway in Triple-Negative Breast Cancer," *Discovery Medicine,* vol. 95, pp. 275-283, May 2014.

[111] A. Jene-Sanz, R. Váraljai, A. V. Vilkova, G. F. Khramtsova, A. I. Khramtsov, O. I. Olopade*, et al.*, "Expression of polycomb targets predicts breast cancer prognosis," *Mol Cell Biol,* vol. 33, pp. 3951-61, Oct 2013.

[112] d. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat Protoc,* vol. 4, pp. 44-57, 2009.

[113] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols,* vol. 4, pp. 44-57, 2009.

[114] U. Wazir, W. G. Jiang, A. K. Sharma, and K. Mokbel, "The mRNA Expression of DAP3 in Human Breast Cancer: Correlation with Clinicopathological Parameters," *Anticancer Research,* vol. 32, pp. 671-674, Feb 2012.

[115] P. K. Wu, S. K. Hong, S. Veeranki, M. Karkhanis, D. Starenki, J. A. Plaza*, et al.*, "A Mortalin/HSPA9-Mediated Switch in Tumor-Suppressive Signaling of Raf/MEK/Extracellular Signal-Regulated Kinase," *Molecular and Cellular Biology,* vol. 33, pp. 4051-4067, Oct 2013.

[116] T. Mukohara, "PI3K mutations in breast cancer: prognostic and therapeutic implications," *Breast Cancer-Targets and Therapy,* vol. 7, pp. 111-123, 2015.

[117] G. Kerdivel, A. Boudot, D. Habauzit, F. Percevault, F. Demay, F. Pakdel*, et al.*, "Activation of the MKL1/actin signaling pathway induces hormonal escape in estrogen-responsive breast cancer cell lines," *Molecular and Cellular Endocrinology,* vol. 390, pp. 34-44, Jun 5 2014.

[118] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein*, et al.*, "Mapping and analysis of chromatin state dynamics in nine human cell types," *Nature,* vol. 473, pp. 43-U52, May 5 2011.

[119] D. A. Cusanovich, B. Pavlovic, J. K. Pritchard, and Y. Gilad, "The Functional Consequences of Variation in Transcription Factor Binding," *Plos Genetics,* vol. 10, Mar 2014.

[120] C. S. Ross-Innes, R. Stark, A. E. Teschendorff, K. A. Holmes, H. R. Ali, M. J. Dunning*, et al.*, "Differential oestrogen receptor binding is associated with clinical outcome in breast cancer," *Nature,* vol. 481, pp. 389-U177, Jan 19 2012.

[121] L. Song and G. E. Crawford, "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells," *Cold Spring Harb Protoc,* vol. 2010, p. pdb prot5384, Feb 2010.

[122] J. Zhang, H. M. Poh, S. Q. Peh, Y. Y. Sia, G. Li, F. H. Mulawadi*, et al.*, "ChIA-PET analysis of transcriptional chromatin interactions," *Methods,* vol. 58, pp. 289-99, Nov 2012.

[123] I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen*, et al.*, "Accumulation of driver and passenger mutations during tumor progression," *Proc Natl Acad Sci U S A,* vol. 107, pp. 18545-50, Oct 26 2010.