STATISTICAL METHODS FOR THE ANALYSIS OF EPIGENOMIC DATA

Pedro L. Baldoni

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2020

Approved by:

Naim U. Rashid

Joseph G. Ibrahim

Michael I. Love

Yun Li

Douglas Phanstiel

ABSTRACT

Pedro L. Baldoni: Statistical Methods for the Analysis of Epigenomic Data
(Under the direction of Naim U. Rashid and Joseph G. Ibrahim)

Epigenomics, the study of the human genome and its interactions with proteins and other cellular elements, has become of significant interest in the past decade. Several landmark studies have shown that these interactions regulate essential cellular processes (gene transcription, gene silencing, etc.) and are associated with multiple complex disorders such as cancer incidence, cardiovascular disease, etc. Chromatin immunoprecipitation followed by massively-parallel sequencing (ChIP-seq) is one of several techniques used to (1) detect protein-DNA interaction sites, (2) classify differential epigenomic activity across conditions, and (3) characterize subpopulations of single-cells in heterogeneous samples. In this dissertation, we present statistical methods to tackle problems (1-3) in contexts where protein-DNA interaction sites expand across broad genomic domains.

First, we present a statistical model that integrates data from multiple epigenomic assays and detects protein-DNA interaction sites in consensus across multiple replicates. We introduce a class of zero-inflated mixed-effects hidden Markov models (HMMs) to account for the excess of observed zeros, the latent sample-specific differences, and the local dependency of sequencing read counts. By integrating multiple samples into a statistical model tailored for broad epigenomic marks, our model shows high sensitivity and specificity in both simulated and real datasets. Second, we present an efficient framework for the detection and classification of regions exhibiting differential epigenomic activity in multi-sample multi-condition designs. The presented model utilizes a finite mixture model embedded into a HMM to classify patterns of broad and short differential epigenomic activity across conditions. We utilize a fast rejection-controlled EM algorithm that makes our implementation among the fastest algorithms available, while showing improvement in performance in data from broad epigenomic marks. Lastly, we analyze data from single-cell ChIP-seq assays and present a statistical model that allows the simultaneous clustering and characterization of single-cell subpopulations. The presented framework is robust for the often observed sparsity in single-cell epigenomic data and accounts for the local dependency of counts. We introduce an initialization scheme for the initialization of the EM

algorithm as well as the identification of the number of single-cell subpopulations in the data, a common task in current single-cell epigenomic algorithms.

To Caro, Jacy, and Pedro, those who taught me the meaning of love.

ACKNOWLEDGEMENTS

I would like to thank my advisors Dr. Naim U. Rashid and Dr. Joseph G. Ibrahim for their extensive support and encouragement throughout the years of graduate education. I am deeply grateful for all the opportunities they have given me to grow as a researcher and as a person. I truly appreciate their time and commitment to this work. The completion of this dissertation would not be possible without them.

I also would like to thank Dr. Yun Li, Dr. Michael Love, and Dr. Douglas Phanstiel for serving on my dissertation committee and for providing constructive comments on my research. I am grateful for their dedication to this work and for their support.

A special thanks goes to Dr. Daniela Sotres-Alvarez, Dr. Jianwen Cai, and Dr. Michael G. Hudgens, with whom I worked alongside since the early years of graduate school. I have learnt so much from them and from all the opportunities they have given me.

To all the staff and personnel of the Department of Biostatistics who were always encouraging and willing to help, thank you. A special thanks goes to Melissa Hobgood, Terry Link, and Nana Abreu.

I would like to thank all of my friends from the Department of Biostatistics that made the difficult years of graduate school better. I will forever remember the good times and experiences that made Chapel Hill my second home during the last six years.

Lastly, I would like to thank my parents that were always with me and will always be, despite the distance. I would not have completed this journey without your support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AMI	Adjusted Mutual Information
ARI	Adjusted Rand Index
BIC	Bayesian Information Criterion
BP	Base Pairs
DPC	Differential Peak Caller
EM	Expectation-Maximization
FDR	False Discovery Rate
FPR	False Positive Rate
GLM	Generalized Linear Model
HMM	Hidden Markov Model
IP	Immunoprecipitation
KBP	Kilobase Pairs
LFC	Log Fold Change
LTPM	Log-transformed Transcripts Per Million
MBP	Megabase Pairs
MHMM	Mixture of Hidden Markov Models
NB	Negative Binomial
PCA	Principal Components Analysis
PDX	Patient-derived Xenograft
RCEM	Rejection-controlled Expectation-Maximization
SNR	Signal-to-noise Ratio
TF	Transcription Factor
TPM	Transcripts Per Million
TPR	True Positive Rate
UMAP	Uniform Manifold Approximation and Projection
ZINB	Zero-inflated Negative Binomial
ZIMHMM	Zero-inflated Mixed Effects Hidden Markov Model

CHAPTER 1: INTRODUCTION

Epigenomics, the study of the human genome and its interactions with proteins and other cellular elements, has become of significant interest in recent years. Such interactions have been shown to regulate essential cellular functions such as gene expression and DNA packaging (Kim et al., 2018), resulting in downstream phenotypic impact. The interrogation of how these interactions occur and how they may change across conditions, such as cell types or treatments, is of marked interest in biomedical research. In cancer research, for instance, certain types of protein-DNA interactions have been shown to play important roles in prostate carcinogenesis and progression (Pfister et al., 2015). Several landmark studies have identified specific genomic regions of changing (differential) epigenomic activity between conditions as drivers of cell differentiation (Creyghton et al., 2010), cancer progression (Varambally et al., 2002), and a number of human diseases (Portela and Esteller, 2010).

To quantify local epigenomic activity, a common high-throughput assay is chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq). ChIP-seq experiments begin with cross-linking DNA and proteins within chromatin structures, which are then fragmented by sonication in a particular sample. DNA fragments bound to the protein of interest are isolated by chromatin immunoprecipitation, which are then sequenced via massively parallel high-throughput sequencing to generate short sequencing reads pertaining to the original fragments. Sequences are then mapped onto a reference genome through sequence alignment to determine their likely locations of origin. Genomic coordinates containing a high density of mapped reads, often referred to as enrichment regions (peaks), indicate likely locations of protein-DNA interaction sites, and all other regions are referred to as background regions.

The detection of enrichment regions in ChIP-seq experiments is challenging for several reasons. Namely, the diversity of enrichment profiles, the presence of serial correlation in the data, sample-specific characteristics such as the signal-to-noise ratio, and an excessive number of zeros in the distribution of read counts. Hence, peak callers need to be tailored accordingly to capture the specific signal of the protein of interest. In differential peak detection, several other challenges affect the ability of existing methods to accurately detect

regions of differential activity from the wide range of ChIP-seq experiments (Section 3.2). First, differential regions may be both short or broad in length, causing difficulty for methods optimized for a particular type of signal profile (Stark and Brown, 2011; Chen et al., 2015). Second, methods that pool experimental replicates together (Song and Smith, 2011) often exhibit more false positive calls compared to methods that jointly model replicates from each condition (Steinhauser et al., 2016). Third, the analysis of ChIP-seq data is often subject to complex biases that may vary across the genome, as differences in local read enrichment may depend on the total read abundance in a given region. It is the purpose of this dissertation, divided into three main chapters, to fill the existing gaps in the literature and present novel approaches to integrate data from multiple experiments for the detection of consensus and differential protein-DNA binding sites.

In Chapter 2, an integrative approach for the detection of broad regions of enrichment in consensus across multiple ChIP-seq experiments is proposed. Through a class of zero-inflated mixed effects hidden Markov models (HMM), the presented model accounts for the main characteristics of broad and diffuse ChIP-seq data and provides better spatial resolution than current available methods. By including sample-specific random effects, we show that this novel framework applied to ChIP-seq data integration is able to account for the long-range correlation present in the data and potential biases due to the different library sizes. Lastly, we demonstrate that the integration of multiple replicates to call peaks in consensus improves the detection of protein-DNA interaction sites.

In Chapter 3, we present a statistical model that integrates data from multiple ChIP-seq experiments (with replicates) and detects broad and short differential regions of enrichment between multiple conditions. The presented model seeks to detect differential enrichment regions by embedding a mixture of negative binomial regression models into a three-component HMM. The HMM component with the embedded mixture model accounts for all possible combinatorial patterns of differential enrichment and background between conditions. As in Chapter 2, we show that the proposed model shows exceptional performance in detecting broad and diffuse differential regions of enrichment and that integrating data from multiple broad ChIP-seq experiments improves the spatial resolution of differential peak calls.

Finally, in Chapter 4 we present a comparative study of existing methods for epigenomic analysis on data sets generated from single-cell ChIP-seq assays. We show that under realistic scenarios, current methods have difficulties in characterizing single-cell sub populations due to the sparsity of the data, an issue that becomes critical in data sets with broad regions of enrichment for sequencing reads. We propose the use of a model-based approach to cluster single-cells into similar sub populations that share similar structural

characteristics. In addition, we present an algorithm for the determination of the existing number of sub populations in a heterogeneous samples, a necessary task in the analysis using current single-cell epigenomic algorithms. The presented approach accounts for the local dependency of counts and is able to analyze single-cell epigenomic data in high genomic resolution, without relying on a set of candidate peaks.

1.1 Literature Review

In this section, a literature review of this dissertation is presented. First, we present in Section 1.1.1 a review of the biological and technical aspects of the ChIP-seq assay. This review will provide the basis for the proposed statistical models to be presented in subsequent chapters. Then, Section 1.1.2 gives an overview of the literature available related to the detection of protein-DNA interaction regions in consensus from multiple ChIP-experiments. We discuss early approaches used for such purposes and how current methods tackle the common problems faced in ChIP-seq data analysis. Next, Section 1.1.3 reviews the literature available for the problem of detecting differential binding sites between multiple conditions. We review the ad hoc strategies used in early stages of the technology as well as existing gaps not fulfilled by current methods.

1.1.1 Introduction to Epigenomics and the ChIP-seq Assay

The interaction between proteins and DNA is a key event that plays a major role in almost all aspects of the cellular processes of living organisms. This phenomenon is known to the scientific community for decades and has gain more relevance in recent years given the reduction of sequencing costs and more data availability (Bulyk et al., 1999; Park, 2009). These interactions are known to regulate gene expression and packaging of DNA into condensed units called nucleosomes, influencing biological processes and phenotypes such as complex human disorders (Jones et al., 2016). Biological relevance of the study of such interactions include their effect on cell differentiation and how they are affected under different treatment conditions. The study of such events may reveal critical mechanisms such that these changes lead to treatment effects for a particular disease of interest.

Histones are proteins found in eukaryotic cells and comprise structural units called nucleosomes which aid in the packaging of DNA. When these proteins are enzimatically modified by either methylation, ADP-ribosylation, phosphorylation, glycosylation, or acetylation, their electric charge and shape are affected along with the structural and functional properties of the chromatin. Consequently, these modifications directly affect transcription, DNA repair, replication and recombination (Nelson et al., 2008; Bannister and Kouzarides, 2011). Histones that are directly associated with gene transcription are of particular interest

given its critical effect on the life of the cell. They can be classified into those associated with gene activation and those associated with gene repression. An example of the former is the trimethylation of histone H3 at lysine 36 (H3K36me3), which associates with RNA polymerase II and gene transcription (Li et al., 2002; Chantalat et al., 2011). On the other hand, the trimethylation of histone H3 at lysine 27 is an example of a known marker that binds to additional proteins to employ a repressive effect on genes (Cao et al., 2002; Liu et al., 2016). As all the functionalities of these proteins remain unknown, they have been the focus of studies in clinical investigation. In cancer research, for instance, the epigenomic mark H3K27me3 has been shown to play an important role in prostate carcinogenesis and progression while H3K36me3-deficient cancer cells are acutely sensitive to gene WEE1 inhibition and can be selectively killed by dNTP starvation (Ngollo et al., 2014; Pfister et al., 2015).

Transcription factors (TFs) form another class of proteins that interacts with the genetic material and mediates the transcription of information from DNA to messenger RNA (Latchman, 1997). These proteins bind to enhancer or promoter regions of the genome and controls the transcription of genes next to them by either blocking or stabilizing the binding of RNA polymerase to the DNA (Gill, 2001). In essence, TFs exert a critical role in the life of the cell by controlling its cycle and responding to internal and external signals (Wheaton et al., 1996). Examples of TFs largely studied include the transcription repressor CTCF, the activating transcription factor 4 (ATF4), and the RE1-silencing transcription factor (REST). These TFs can act by either down or up regulating the gene production and have been linked to several vital processes and diseases. As an example, REST is a known repressor that has been linked to colon and lung cancer, Huntington Disease and other illnesses (Westbrook et al., 2005). Therefore, the study of such class of proteins with respect to their interaction with the genetic material and their downstream phenotypic impact is an open area of research for investigators given its biological relevance.

The chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) is a commonly used technique to detect genome-wide regions of protein-DNA interaction, such as TF binding sites or regions containing histone modifications (Robertson et al., 2007). Results from ChIP-seq experiments have been successfully used to understand epigenomic mechanisms in which transcription factors and histone modifications play an important role (Barski et al., 2007; Robertson et al., 2007). Such mechanisms are hypothesized to explain heterogeneity at both the molecular level (gene expression, gene silencing, DNA replication etc.) and on an individual level (cancer incidence, cardiovascular disease, obesity, etc.). In cancer research, for instance, histone modifications have been shown to play an important role in carcinogenesis,

progression, and tumor suppression (Ngollo et al., 2014; Lu et al., 2016; Huang et al., 2017). Hence, ChIP-seq experiments provide a useful way for investigators to explore epigenomic modifications that might lead to downstream phenotypic impact.

ChIP-seq experiments begin with cross-linking DNA and proteins on chromatin structures followed by sonication-induced fragmentation. DNA fragments bound to the protein of interest are then isolated by a technique called Chromatin Immunoprecipitation (ChIP). Finally, the associated fragments are sequenced via massively parallel sequencing to generate short sequencing reads pertaining to the original fragments. These sequences are then mapped back to a reference genome through sequence alignment to determine their likely genomic locations of origin. Genomic coordinates containing a high density of mapped reads, referred to as *enriched* regions, are then identified through statistical analysis. These coordinates indicate likely locations where the protein of interest was bound to the DNA. Here, we refer to all other genomic positions pertaining to non-enriched regions as *background*. The majority of methods available for the detection of enrichment regions compare the distribution of the read counts (signal) in ChIP-seq experiments to similar regions in matched input control experiments. This comparison helps to more accurately define regions of enrichment in ChIP-seq samples by accounting for the technical variation in local read density in the input control sample. These methods calculate the experimental signal by first tiling the genome with non-overlapping windows (Rashid et al., 2011; Ibrahim et al., 2014; Cuscò and Filion, 2016) or sliding windows (Zhang et al., 2008), and then computing the number of reads mapped into each window.

The detection of enrichment regions in ChIP-seq experiments is challenging due to several reasons, including the diversity of enrichment profiles, the presence of serial correlation in the distribution of window read counts, and sample-specific characteristics such as the signal-to-noise ratio and the input control effect. In addition, under-sequenced experiments are characterized by an excessive number of observed zeros in the distribution of read counts, which imposes additional challenges when detecting regions of enrichment in diffuse data. Therefore, methods to detect regions of enrichment from ChIP-seq experiments need to be tailored accordingly to capture the specific signal profile of the protein of interest. For instance, TF binding sites are usually characterized by sharp and punctate read profiles while histone modifications show broad and diffuse regions of enrichment spanning tens of thousands of base pairs across the genome.

In addition, ChIP-seq experiments usually differ with respect to the total number of mappable reads, here referred to as the sequencing depth or library size. The heterogeneity across experiments has been shown to significantly influence the results from ChIP-seq data analysis, whose effect is more pronounced when

detecting differential protein-DNA interaction sites across multiple experiments (Chen et al., 2012). In this scenario, samples with higher library sizes tend to dominate the analysis, leading to an increased number of false positives. This bias has also been seen in other types of next generation sequencing (NGS) data, such as RNA-seq (Robinson and Oshlack, 2010). In ChIP-seq data, Jung et al. (2014) suggested a practical lower bound of 40-50 million reads for most of the marks from human cells in order to ensure robust conclusions from results derived from peak-calling algorithms. However, publicly available ChIP-seq data quite often do not meet this suggested minimum number of reads and show high variation with respect to their library sizes. Several methods have been introduced in the literature for data normalization to account for such differences, such as the trimmed mean of M-values (Robinson et al., 2010), the normalization via gene expression levels of housekeeping genes (Allhoff et al., 2016), and loess-based normalization for trended biases (Lun and Smyth, 2015) to name a few. As most of these methods rely on strong assumptions about the data, there is still no consensus in the literature on how to account for such differences, and the problem of ChIP-seq data normalization is still an active area of research.

In this dissertation, we aim to present statistical models to detect protein-DNA interaction sites while tackling the main challenges associated with ChIP-seq data analysis. In particular, the models presented in Chapters 2.7 and 3.6 address the issues of the excess of zeros found in broad and diffuse data as well as potential biases due to the different library sizes between experiments.

1.1.2 Statistical Approaches for the Detection of Consensus Peaks

Before the introduction of methods focused on the integration of multiple ChIP-seq experiments, protein-DNA interaction sites in consensus across multiple experiments were detected by means of ad hoc rules to combine peaks called independently from multiple samples (Valouev et al., 2008; Bottomly et al., 2010). The majority of these approaches consisted in using single sample peak callers, such as MACS2, HOMER, and ZINBA (Zhang et al., 2008; Heinz et al., 2010; Rashid et al., 2011), to detect regions of enrichment independently across samples and combine the results accordingly. The final set of regions of enrichment in consensus would be formed by those that were detected across all samples or in the majority of samples. Yang et al. (2014) presented guidelines on how to combine results from independent peak calls, which were often subject to the decision in downstream analyses. Specifically, the authors have shown that leveraging replicates from multiple experiments improved the detection of enrichment sites and recommended calling a candidate peak as common if it was in consensus in at least two out of three potential ChIP-seq experiments.

Due to the lack of sound statistical methodologies that integrate multiple ChIP-seq datasets, investigators often relied on these rules to find peaks from multiple samples.

As well pointed out by others (Ibrahim et al., 2014; Lun and Smyth, 2015; Cuscò and Filion, 2016; Allhoff et al., 2016), combining results from individual peak calls has several potential pitfalls. As previously stated, ChIP-seq experiments are characterized by distinct enrichment profiles, show distinct signal to noise ratios (SNRs), and are sometimes diffuse in signal. In these situations, the spatial resolution of peak calls are often compromised as the difference in signal intensity across samples is not properly taken into account, leading to results that are not biologically meaningful. Secondly, as pointed out by Lun and Smyth (2015), calling individual peaks and combining results using ad hoc rules might lead to a final set of consensus regions of enrichment that does not contain regions with low signal but are consistently seen in all the datasets, leading to a increase in the observed proportion of false negatives. Thirdly, combining peaks from several experiments often leads to a final set of narrow and discontiguous regions of enrichment, as we show in Chapter 2.7 of this dissertation. This effect is particularly pronounced when analyzing data from broad and diffuse ChIP-seq experiments, given that enriched domains usually expand thousands of base pairs with changes in the enrichment profile across the genome.

Alternatively, another approach used in early years and before the introduction of methods to integrate multiple ChIP-seq experiments was the pooled type of analysis (Niu et al., 2011; Young et al., 2011). In such an approach, reads from multiple experiments would be pooled together and analyzed as a single experiment. Then, any single sample peak caller could be potentially used to analyze and call peaks from the combined data. If using a window-based strategy to compute the ChIP-seq signal, one would sum up the read counts assigned to genomic windows across all experiments and call peaks as if they were originally generated from a single experiment.

Even though pooling aligned reads from technical replicates is one of recommended guidelines from the ENCODE consortium (Dunham et al., 2012), authors have noted that this strategy makes differences in the enrichment profiles across experiments indistinguishable (Ibrahim et al., 2014). This effect is particularly pronounced in broad and diffuse data, such that the spacial resolution of single-experiment peak callers become unrealistic. In this dissertation, we assessed these claims using simulated and biological data. We observed that the results under this approach tend to show an increased false positive rate as more replicates are combined (see Chapter 2.7). This is justified by the fact that ChIP-seq experiments are not expected to be entirely reproducible per se, given the inherent technical variation present in the protocol, even with technical

or biological replication. For this reason, pooled reads from several experiments tend to lead peak callers to call broader regions of enrichment than the actual one in consensus.

In the scenario of single sample ChIP-seq data analysis, methods for detecting TF binding sites and broad enrichment regions from histone modifications have been successfully presented in the literature (Zhang et al., 2008; Machanick and Bailey, 2011; Xing et al., 2012; Bardet et al., 2013; Wu and Ji, 2014; Rashid et al., 2014). However, methods focused on peak calling from multiple samples are of growing interest given the reduction of sequencing costs and higher data availability. Besides leveraging information from multiple experiments, these methods provide an output that is easier to interpret rather than attempting to integrate results from individual peak calls. To the best of our knowledge, the literature provides only two established methods for the detection of consensus regions of enrichment from multiple experiments. From our experience, these methods perform well in situations with either sharp, punctate, or non-diffuse datasets, but fail to call broad regions of enrichment from marks such as H3K27me3.

In JAMM, Ibrahim et al. (2014) integrate multiple technical replicates and fit a multivariate Gaussian mixture model to cluster genomic windows and call regions of consensus. In the presented model, the extended read counts are mapped back to a reference genome that is divided into narrow and non-overlapping bins. First, JAMM uses preprocessing rules to select and merge candidate enriched bins into larger and non-overlapping enriched windows. Secondly, to find consensus peaks across experiments, the presented model fits either a two- or three- component Gaussian mixture model (based on a priori knowledge of the data) on the smoothed extended read counts in each window separately. Bins within windows are then clustered according to the posterior probabilities calculated from an EM algorithm. For a given window, the mixture component with the largest mean is assumed to be the enrichment cluster and bins assigned to to this cluster are taken to be enriched and merged if neighboring. If multiple experiments are available, all replicates must agree with respect to the mixture component assignment in order for a peak to be considered to be in consensus. JAMM works on normalized counts that are computed as follows. First the geometric mean of ChIP signal is calculated for each bin across all replicates. Then, JAMM subtracts from it the background signal and calculate the peak-based average signal. Finally, JAMM executes a Mann-Whitney U test to compare the enrichment and background ChIP signal followed by a correction of p-values (Benjamini and Hochberg, 1995).

The key points of JAMM are its improved spatial resolution of peak calls, the universality of the method to analyze several types of datasets, and the robust peak scoring and sorting. The two-step approach used

by JAMM leads to a better spatial resolution than the benchmarked competitors when analyzing sharp and punctate data. The presented framework decides whether bins are enriched over background, merges those neighboring ones forming genomic windows, and then separately clusters bins within widows to make the final set of consensus regions of enrichment. This approach ensures that JAMM resolves neighboring punctate sites and avoid that peaks located nearby are not called as a single enrichment region. Ibrahim et al. (2014) showed good results from JAMM when analyzing data from transcription factors CTCF, NRSF, MAX, and SRF, and histone modification H3K4me3. This indicates that JAMM had good performance when calling peaks from data with distinct (although sharp) signal properties, making it a peak caller that that is robust for sharp data. JAMM provides a large number of peaks and robustly score them based on the background-normalized mean signal of peaks and uses the Benjamini-Hochberg corrected p-values.

In general, data from broad and diffuse ChIP-seq experiments are characterized by large domains of enriched regions. The good spatial resolution of JAMM when analyzing sharp and punctate datasets is not supported under broad data. Because diffuse enrichment regions usually contain a heterogeneous signal pattern within their domains, JAMM tends to detect narrow and discontiguous regions of enrichment when in fact broad regions should be called. We present these results in Chapter 2.7 of this dissertation. One reason for this is because JAMM does not account for the long-range correlation of read-counts that is characteristic of diffuse data. In addition, JAMM tends to call an excessively larger number of peaks than the benchmarked competitors studied in their paper. This leads JAMM to have a limited performance in diffuse data, given that it calls numerous narrow regions of enrichment that do not correspond to the entire consensus peak.

In Cuscò and Filion (2016), the authors present Zerone, a three-state hidden Markov model (HMM) with Zero-Inflated Negative Multinomial (ZINM) emission distributions to identify regions of enrichment in consensus across experiments. Zerone uses a two-step approach to produce the final set of consensus peaks by first discretizing the ChIP-seq signal with the HMM and then checking the results using a built in quality control tool to detect low quality and/or non reliable peaks. Their method uses ZINM distributions on the window level and conditions the read counts from replicates on the total number of mapped reads. The choice of using a three-component HMM is based on the claim that the baseline signal of ChIP-seq experiments can show low amplitude and expands through large domains, which would make a two-component HMM to falsely detect such regions as enriched. Therefore, their first two HMM components are dedicated to call baseline regions while the third component is aimed to detect the consensus region of interest. The Baum-Welsch algorithm (Rabiner, 1989) is used to obtain parameter estimates of the model and the Viterbi

algorithm (Viterbi, 1967) is used to compute the most likely segmentation. The built-in quality control was based on a trained SVM with 91 datasets with a successful discretization and 91 negative cases obtained by discretizing controls without immunoprecipitation, all datasets from the ENCODE project. It is important to note that the success of discretization was subjective, given that there is no gold standard for protein biding.

The key aspects of Zerone include its built-in quality control tool, its efficient running time and the HMM-based approach to call regions in consensus across replicates. In their paper, Zerone was the fastest method when benchmarking with BayesPeak (Spyrou et al., 2009), JAMM, and MACS2. In addition, Zerone showed a smaller memory footprint than the competitor MACS2. The reason for its high efficiency is likely due to its estimation process during Baum-Welsch cycles, which assumes that provided input controls captures systematic biases such as batch effects. The built-in quality control provides an alternative tool to the recommended IDR method recommended by Encode Consortium, as it can analyze more than two replicates, a key limitation of the IDR. This quality control tool does not assume any signal profile to the data and can be applied to any number of replicates. In addition, because it uses a HMM approach, Zerone is able to account for the long range correlation present in broad data, a characteristic that is lacking in JAMM. This puts Zerone as an alternative approach to call regions in consensus from broad data.

However, in the work presented by Cuscò and Filion (2016), the authors do not present any analysis of broad and diffuse data from histone modification H3K27me3. They analyzed data from transcription factor CTCF and histone modification H3K36me3. Even though this histone modification is known to be characterized by broad peaks, the data resulting from this mark are not as diffuse as the ones from H3K27me3. In this dissertation, we assessed the performance of Zerone when calling peaks from this mark. However, Zerone showed limited performance as its called regions were, in general, narrow and discontiguous (see Chapter 2.7). We believe that the reason for this fact is that the HMM used by Zerone assigns regions with low enrichment profile that expand across large domains to background, and only those with an elevated number of read mapped onto are considered to be as enriched. For a histone modification like H3K27me3, it is critical to classify these low profile regions as enriched. We believe that such regions are still of interest and should be considered as consensus if the profile is consistent across datasets. In addition, the 'all-or-none' strategy used by Zerone in its quality control tool is not ideal, as it might miss domains with low signal in under-sequenced experiments such as those from H3K27me3. Also, when integrating data from multiple experiments, Zerone combines all available input controls. Because ChIP-seq experiments are not entirely

reproducible, even in scenarios of biological and technical replicates, it would be ideal to have sample-specific input controls taken into account when detecting peaks in consensus.

In general, current approaches that integrate multiple ChIP-seq experiments (Ibrahim et al., 2014; Cuscò and Filion, 2016) have limited performance with respect to the spatial resolution of their called regions when analyzing broad and diffuse data, even after leveraging additional data. We observed that the low read density profile of diffuse histone modifications, such as H3K27me3 and H3K36me3, led the aforementioned methods to fragment broad regions of enrichment into narrower and discontiguous peak calls. In this dissertation, we aim to tackle these issues and present in Chapter 2.7 a Zero-Inflated Mixed Effects Hidden Markov Model (ZIMHMM) to analyze data from multiple ChIP-seq experiments. Our model is tailored to detect broad consensus regions of enrichment across multiple experiments. ZIMHMM accounts for the excess of zeros, common to broad and diffuse histone modifications, as well as sample-specific library sizes and ChIP-control relationship via random effects. To the best of our knowledge, there is no work published in the literature that proposes a random effects model for joint analysis of multiple ChIP-seq experiments while accounting for zero inflation and sample-specific effects. Using publicly available ChIP-seq data from both H3K27me3 and H3K36me3 marks from the ENCODE Consortium (Dunham et al. 2012), we compared the performance of our method to the current peak callers JAMM, Zerone, and MACS2 (under both independent and pooled approaches). Based on real data analyses, we show that ZIMHMM outperforms the existing methods for detection of broad consensus regions of enrichment from multiple ChIP-seq experiments.

1.1.3 Statistical Approaches for the Detection of Differential Peaks

Investigators are often interested in comparing results from data of multiple ChIP-seq experiments in order to detect differences in binding of a given protein of interest under different conditions. Such conditions could be different treatments, cell lines, mutated and wild type cells, to name a few (Feng et al., 2014; Koues et al., 2015; Clouaire et al., 2014). In the early stages of the development of the ChIP-seq technology, ad hoc methods were used for such purpose. For instance, a common and straightforward approach was to compare peaks called independently from ChIP-seq experiments to find those that were in consensus or unique across datasets. Often, a common practice was to use Venn diagrams to represent all possible configurations of peaks and find those that were differential (Chen et al., 2008).

Several caveats exist in this approach. First, it completely ignores the differences in ChIP signal intensities across experiments. Quite often ChIP-seq datasets differ with respect to the library size, here defined as the

total number of reads mapped onto the reference genome. In such cases, it is expected that peaks found independently in the datasets will show a consistent difference in signal that will not be accounted when naively comparing the sets of peak calls, leading to an elevated number of false negatives (Chen et al., 2015). Other sources of bias due to the multi-stage steps of the ChIP-seq protocol might as well lead to the same problem. Lun and Smyth (2014) have shown that such an approach leads to loss of error rate when comparing peaks for differential binding. Secondly, the spatial resolution of the differential regions of enrichment could be compromised for scenarios of sharp and broad domains. Post comparison of inaccurate peak calls might completely miss sharp events, such as those from TF, or even segment broad and diffuse enrichment regions into narrow and discontiguous peaks that do not correspond to the actual data. Thirdly, because the set of differential peaks is restricted to those previously found independently in each sample, this naive approach cannot detect changes within the broad differential differential domains, an issue that is particularly relevant for proteins that expand large regions of the genome (Allhoff et al., 2014).

To overcome these issues, joint peak callers for the detection of differential binding sites were proposed in the literature (Zhang et al., 2014; Shen et al., 2013; Lun and Smyth, 2015; Allhoff et al., 2016). The majority of these methods adopted a window-based approach in which the ChIP signal is calculated across the genome into non-overlapping or sliding windows. The main goal of these approaches was to reduce the systematic bias and improve the low spatial resolution from the two-stage strategies that were used to compare peaks across experiments. Most of the methods that integrate data from multiple experiments for differential peak call were tailored for particular scenarios. As discussed in this section, however, these methods have been tailored to analyze data under certain scenarios, such as to detect differential enrichment profiles from sharp events. Additional benefits of the current methods include the post-processing FDR control of differential events on the region level, the integration of technical or biological replicates, and the introduction of novel normalization methods, to name a few. However, we believe that the problem of differential peak call is still an open problem, as current methods still have limitations regarding the number of possible experiments in comparison, the detection of broad differential binding sites, and subjective approaches that rely on ad hoc rules for the classification of regions that make the final results difficult to interpret.

In Zhang et al. (2014), the authors present PePr, a peak caller that identifies consistent or differential binding sites in ChIP-seq experiments with replicates. The authors use an sliding window approach to model read counts across replicates and conditions using a local negative binomial model. In their paper, the authors benchmark PePr with other methods using transcription factor data to find differential peaks across groups.

PePr is peak caller able to analyze and compare ChIP-seq data from only two groups. This limitation leaves to the investigator the decision on how to analyze data from multiple groups. In addition, the sliding window approach used by PePr can be highly sensitive to the window chosen. While large windows are not ideal to detect small changes in domains of broad data, narrow windows lead to discontiguous peaks that are difficult to interpret. Another downsides of PePr include its tendency to call peaks larger than other tools and the observed histone changes, as noted by others (Allhoff et al., 2014), and its inability to account for input controls to correct for systematic biases due to the library preparation.

In Allhoff et al. (2016), the authors presented THOR, a peak caller for the detection of differential binding sites. It uses a Hidden Markov Model approach and is able to analyze pairs of biological conditions with replicates. A negative binomial model used to fit the read counts and account for the overdispersion. It provides the trimmed mean of M values (TMM) as a normalization method, a method often used when analyzing gene expression data (Robinson et al., 2010), as well as a novel normalization method based on housekeeping genes for activating histone marks. As well pointed out by the authors, the normalization method is crucial for heterogeneous data, specially in cases with distinct signal to noise ratio. However, quite often the normalization methods used in ChIP-seq data analysis rely on strong assumptions that are not realistic. For instance, TMM assumes that the counts assigned to enriched domains are constant across the genome, a not realistic assumption in ChIP-seq data given that different conditions might show distinct amounts of binding proteins. In their package, the authors provide the complete preprocessing steps necessary for the data analysis of multiple ChIP-seq experiments, namely the fragment size estimation, GC-content correction, scaling based on input control, and signal normalization.

One of the main disadvantages of THOR is its ability of analyzing only two groups at a time. In addition, we observed that differential peaks detected by THOR tend to include those in consensus across both groups. This fact is possibly explained by the fact that the presented HMM includes only three components in their model and is not able to separate peaks is in consensus across groups from those that are differential. It is unclear from their paper how this case is handled in the proposed HMM. In addition, the authors uses a moment estimator to obtain parameter estimates for the model mean without taking a regression-based approach that would potentially allow the control for additional covariates of interest. Moreover, the authors estimate the dispersion parameter of the model using a two step approach. Under this strategy, the authors assume that the model variance is a quadratic function of the mean, which is previously estimated from the data, and then the dispersion parameter is estimated using the properties of the negative binomial model.

The authors do not mention any lower bound constraints in their estimation process to ensure that estimated dispersion parameter would assume positive values. Additionally, the normalization method presented by the authors of THOR assumes that the protein of interest being analyzed is positively associated with gene expression. This leaves the case of normalization for repressive marks as an open problem.

In Lun and Smyth (2015), the authors present csaw, a joint peak caller aimed for differential binding analysis. To the best of our knowledge, csaw and ChIPComp (Chen et al., 2015) are the only methods available in the literature able to simultaneously compare more than two groups. It can handle complex experimental designs with biological replicates and allow quantitative comparisons between DNA samples or experimental conditions. Csaw addresses the main issues of one- and two-stage differential peak callers that do not control for the FDR on the region level Ross-Innes et al. (2012); Liang and Keleş (2011); Allhoff et al. (2016). Because results from ChIP-seq data analyses are often interpreted on domains created after merging neighboring differential windows, it is key to control for FDR on the regions level. In addition, csaw provides several tools for batch effect removal and ANOVA-like testing approaches.

Csaw requires replicates in at least one of the analyzed conditions. In their paper, the authors only address the issues of broad and diffuse data under simulation studies. In our investigations, we observed that csaw had suboptimal performance when calling peaks from diffuse data, as most of the peaks were discontiguous and did not cover the entire differential regions of enrichment across groups (see Chapter 3.6). Moreover, the authors do not present in their paper a genome-wide quantitative comparison of their method with competitors, leaving the final conclusions based on visual inspections of the results. For the results from diffuse data, the authors states that csaw was not able to detect the differential regions of enrichment and argue that a possible solution would be to increase the window size, if such events are of interest. However, we believe that this introduces an additional subjective level in their framework given that it is left for the investigator the choice of the window size. The paper do not discuss the trade-off between the increase the window size and the loss of spatial resolution of broad domains. In their analyses of TF data with sharp peaks, csaw did not exhibit any power advantage over the competitor Diffbind in their simulation study. In this dissertation, we evaluate the performance of csaw under publicly available data and observed a limited performance when analyzing diffuse datasets (see Chapter 3.6)

1.1.4 Current approaches for the analysis of single-cell ChIP-seq data

In recent years, several methods designed for the analysis of single-cell epigenomic data have proposed in the literature (González-Blas et al., 2019; Fang et al., 2019; Cusanovich et al., 2018; Baker et al., 2019). Utilizing data from the single-cell Assay for Transposase Accessible Chromatin using sequencing (scATAC-seq), the ultimate goal of these methods is the clustering (followed by the characterization) of single cells into homogeneous sub populations of cells exhibiting similar epigenomic profiles regarding the accessibility of their chromatin landscape. The analysis of scATAC-seq data allows biomedical researchers to study a number of chromatin-accessibility signatures on the single-cell level that include the binding of transcription factors that control the expression of nearby genes (Schep et al., 2017). In cancer research, for instance, the information from the epigenome of individual cells can explain parts of the biological variation found in treatment responses that are shown to be cell type dependent (Kagohara et al., 2020). The recent advances in single-cell epigenomic technologies as well as in methods focused on the analysis of these types of datasets allow researchers to understand much of the biological heterogeneity that was often unexplained in previous years of bulk sequencing assays.

A recent study compared the performance of such methods in an extensive analysis utilizing simulated and real data and provide guidelines on their use (Chen et al., 2019). A few characteristics are shared across nearly all scATAC-seq methods benchmarked by Chen et al. 2019. First, current methods require as input a set of pre-specified genomic coordinates that are thought to differentiate the sub population of cells well. Second, methods do not account for or explicitly model the local dependency of single-cell counts in their analytic framework. Third, all methods rely on a two-step procedure for clustering and subsequent peak calling within sub populations of cells regarding the epigenomic activity of interest. To this end, the optimal number of existing sub populations of cells is estimated from the data in an early step of the analysis (Xiong et al., 2019).

In contrast to scATAC-seq data, data from scChIP-seq experiments pose challenges to these methods. Since candidate peaks are often specified using bulk data, the choice of the peak calling algorithm and its parametrization can highly influence the final set of peaks, specially for broad marks, as we show in the Chapter 2 of this dissertation. Moreover, the low sequencing depth and relatively high noise of scChIP-seq experiments, in addition to the broadness of regions of activity from certain epigenomic marks, may cause these methods to have limited performance in the analysis of scChIP-seq data (see Section 4.4). Due to these

issues and the lack of statistical methods for the analysis of scChIP-data, current analyses of scChIP-seq data often use ad hoc approaches that are tailored for the particular problem at hand (Grosselin et al., 2019).

In this dissertation, we present a comparative study of scATAC-seq method on simulated data for scChIP-seq experiments and propose the use of an initialization algorithm for the selection of candidate differential regions of enrichment from single-cell data. Using scChIP-seq simulated data, candidate differential regions are shown to better distinguish sub populations of single-cells and improve the performance of current scATAC-seq method in scChIP-seq data. Existing scATAC-seq methods rely on sets of candidate peaks detected from aggregated single-cell data. We show in this chapter that such a strategy may compromise the analysis of scChIP-seq data sets, which often exhibits broad regions of enrichment that, once aggregated, leads to candidate peaks that mask differences among sub populations of single-cells. In addition, we present an algorithm for the determination of the existing number of sub populations in a heterogeneous samples, a necessary task in the analysis using current single-cell epigenomic algorithms. The presented approach accounts for the local dependency of counts and is able to analyze single-cell epigenomic data in high genomic resolution, without relying on a set of candidate peaks.

CHAPTER 2: IMPROVED DETECTION OF EPIGENOMIC MARKS WITH MIXED EFFECTS HIDDEN MARKOV MODELS

2.1 Introduction

Chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) is a technique to detect genome-wide regions of protein-DNA interaction, such as transcription factor (TF) binding sites or regions containing histone modifications (Robertson et al., 2007). These interactions may regulate gene expression and influence biological processes (Jones et al., 2016). ChIP-seq experiments have been used to understand epigenomic mechanisms in which TFs and histone modifications play an important role (Barski et al., 2007). Such mechanisms are hypothesized to explain heterogeneity at both the molecular level (gene expression, gene silencing, etc.) and on an individual level (cancer incidence, cardiovascular disease, etc.). In cancer research, histone modifications have been shown to play an important role in carcinogenesis, progression, and tumor suppression (Huang et al., 2017).

ChIP-seq experiments begin with cross-linking DNA and proteins on chromatin structures followed by sonication-induced fragmentation. DNA fragments bound to the protein of interest are then isolated by chromatin immunoprecipitation. The associated fragments are sequenced via massively parallel sequencing to generate short sequencing reads pertaining to the original fragments. These sequences are then mapped back to a reference genome through sequence alignment to determine their likely genomic locations of origin. Genomic coordinates containing a high density of mapped reads, referred to as *enriched* regions, are then identified through statistical analysis. These coordinates indicate likely locations where the protein of interest was bound to the DNA. Here, we refer to all other genomic positions pertaining to non-enriched regions as *background*. Methods available for the detection of enrichment regions calculate the distribution of the read counts (signal) in ChIP-seq experiments by first tiling the genome with non-overlapping windows (Rashid et al., 2011) or sliding windows (Zhang et al., 2008), and then computing the number of reads mapped into each window.

The detection of enrichment regions (peaks) in ChIP-seq experiments is challenging for several reasons. Namely, the diversity of enrichment profiles, the presence of serial correlation in the data, sample-specific characteristics such as the signal-to-noise ratio, and an excessive number of zeros in the distribution of read counts. Hence, peak callers need to be tailored accordingly to capture the specific signal of the protein of interest. Although single sample ChIP-seq peak callers have been successfully presented in the literature (Zhang et al., 2008; Xu et al., 2010; Kuan et al., 2011; Song and Smith, 2011; Xing et al., 2012; Rashid et al., 2014), multi-sample peak callers are of growing interest given the reduction of sequencing costs. Leveraging additional data into a joint framework leads to a significant improvement when detecting consensus peaks across samples (Yang et al., 2014). However, current approaches that integrate multiple ChIP-seq replicates (Ibrahim et al., 2014; Cuscò and Filion, 2016) show poor spatial resolution of peak calls when analyzing diffuse data due to the low signal profile of broad histone modifications. Under these scenarios, we observed that broad regions of enrichment are fragmented into narrow and discontiguous peak calls by the aforementioned methods (see Sections 2.2 and 2.6).

To tackle these challenges, we present a Zero-Inflated Mixed effects Hidden Markov Model (ZIMHMM) to analyze data and detect broad peaks in consensus across multiple ChIP-seq technical or biological replicates. The ZIMHMM accounts for the excess of zeros as well as sample-specific sequencing depth and ChIP-control relationship via random effects. Using data from H3K27me3 and H3K36me3 ChIP-seq experiments on human cells from the ENCODE Consortium and the Roadmap Epigenomics Project (Dunham et al. 2012; Bernstein et al. 2010; see Appendix A for details), we compared the performance of the ZIMHMM to the current multi-sample peak callers JAMM and Zerone, as well as the single-sample methods BCP, CCAT, MACS2, MOSAiCS, and RSEG. Based on real data analyses, we show that the ZIMHMM outperforms the existing approaches for detection of broad consensus regions of enrichment from multiple ChIP-seq experiments (see Section 2.6).

2.2 Background

Histones are proteins found in eukaryotic cells and comprise structural units called nucleosomes which aid in the DNA packaging. When these proteins are enzimatically modified by either methylation, ADP-ribosylation, phosphorylation, glycosylation, or acetylation, their electric charge and shape are affected along with the structural and functional properties of the chromatin. Consequently, these modifications directly affect transcription, DNA repair, replication and recombination (Bannister and Kouzarides, 2011). From

all the variant forms of histones, the trimethylation of histone H3 at lysines 36 and 27 (H3K36me3 and H3K27me3) are of particular interest due to their association to actively transcribed genes and gene repression, respectively (Liu et al., 2016). In cancer research, for instance, the epigenomic mark H3K27me3 has been shown to play an important role in prostate carcinogenesis and progression while H3K36me3-deficient cancer cells are acutely sensitive to gene WEE1 inhibition and can be selectively killed by dNTP starvation (Pfister et al., 2015).

ChIP-seq experiments usually differ with respect to the number of mappable reads, referred to as the sequencing depth. Jung et al. (2014) suggested a practical lower bound of 40-50 million reads for most of the marks from human cells in order to ensure robust conclusions from results derived from peak-calling algorithms. In general, publicly available ChIP-seq data do not meet this suggested minimum number of reads and show high variation regarding their sequencing depths. We observed that this variation mediates the effect of the input control on the distribution of ChIP signal across different experiments and regions of the genome (see Section 2.6 and Figure A.1 in Baldoni et al. 2019b). While the input controls might well explain the technical variation in ChIP read counts on enrichment regions from highly sequenced experiments, their effect is not pronounced in under-sequenced data.

When analyzing diffuse or under-sequenced data, we observed that current multi-sample peak callers fail to call sufficiently broad regions of enrichment. In general, such methods call narrow and discontiguous peaks that do not correspond to the entire range of protein-DNA binding site. Under the pooling type of analysis, methods tended to call the union of individual peaks as the ChIP-seq signals were combined by merging reads from multiple samples (Ibrahim et al., 2014). In addition, these data are characterized by a low read density profile and an excess of zeros that one would not expect to observe if modeling the signal with either a Poisson or Negative Binomial (NB) distribution. In this scenario, we find that the Zero-Inflated Negative Binomial (ZINB) model appears to accurately capture the excess of zeros present in background regions of the genome (Figure 2.1).

Before the introduction of methods focused on the integration of multiple ChIP-seq experiments, consensus regions of enrichment were detected by using ad hoc rules to combine peaks called independently from different samples (Valouev et al., 2008). Alternatively, aligned reads from all experiments available could be pooled and analyzed by single-sample procedures (Young et al., 2011). However, we observed that peak calls from pooled samples usually correspond to the union of individual enrichment regions (see Figure D.1 in Baldoni et al. 2019b). In recent years, a few methods have focused on the joint analysis of ChIP-seq

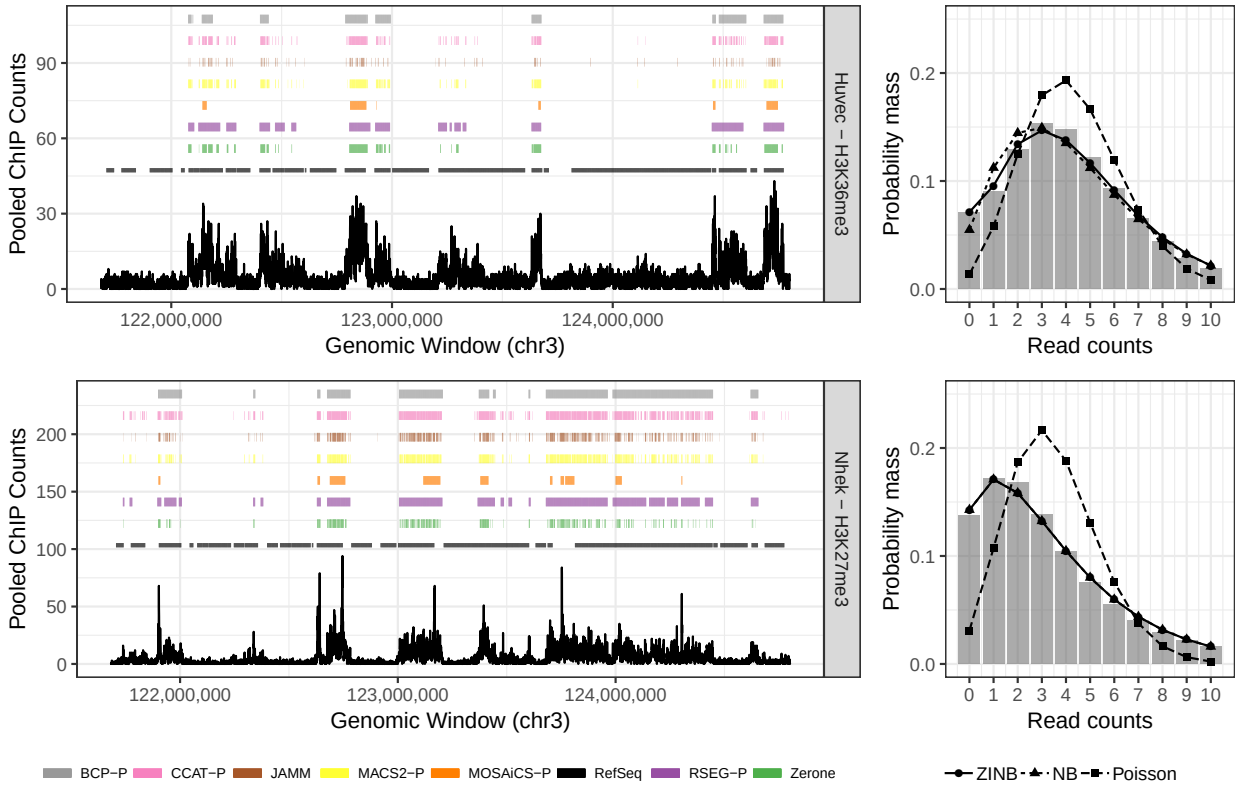

Figure 2.1: Low and broad signal profile of histone modifications H3K36me3 (top panels) and H3K27me3 (bottom panels). On the left, pooled read counts of technical replicates of diffuse histone marks ChIP-seq on human Huvec and Nhek cells, respectively, and peaks called by some of the current methods. On the right, bar plots of the observed count distribution of ENCODE background regions on these cells and expected proportions under the Poisson, NB, and ZINB models. This figure appears in color in the electronic version of this article.

data to call consensus peaks. JAMM integrates multiple technical replicates and fits a multivariate Gaussian mixture model to cluster genomic windows and call regions of consensus. Zerone fits a three-state HMM with Zero-Inflated Negative Multinomial emission distributions to identify regions of enrichment. As shown in Figure 2.1, these methods do not perform well when capturing broad regions of enrichment in consensus across multiple samples.

2.3 Methods

Here, we first introduce an immediate extension of the single-sample HMM proposed by Rashid et al. (2014) in Section 2.3.1. Such an extension is aimed to call consensus regions of enrichment from multiple ChIP-seq experiments by fitting a two-state fixed effects multivariate Zero-Inflated HMM. In Section 2.3.2, we present the ZIMHMM, a mixed effects version of the extended model motivated by the work from Altman (2007). Both models capture the excess of background zeros from diffuse data and, in addition, the ZIMHMM accounts for sample-specific differences via random effects.

2.3.1 Multi-sample Zero-Inflated HMM

From here onwards, all models will be presented under a two-state HMM with ZINB and NB emission distributions associated with the background and enrichment states, respectively. For genomic window j of experiment i, $j = 1, \ldots, M$ and $i = 1, \ldots, N$, let Y_{ij} and X_{ij} denote the random variables pertaining to the ChIP and log-transformed input control read counts, respectively. Here, y_{ij} and x_{ij} denote the observed values of Y_{ij} and X_{ij}, respectively. For multiple experiments sharing the same input control, we have $X_{ij} = X_{i'j}$ for all $i \neq i'$. We assume a single latent discrete time stationary Markov chain $\mathbf{Z} = \{Z_j\}_{j=1}^{M}$, $Z_j \in \{1, 2\}$, with state-to-state transition probabilities $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22})'$ and initial probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2)'$. Conditionally upon Z_j, we observe the vectors of independent counts $\mathbf{Y}_{\cdot j} = (Y_{1j}, \ldots, Y_{Nj})'$ and $\mathbf{X}_{\cdot j} = (X_{1j}, \ldots, X_{Nj})'$, for all windows $j = 1, \ldots, M$ and across all N replicated experiments.

Let $\boldsymbol{\psi}_{z_j}$ denote the vector of state-specific parameters, f_1 and f_2 denote the emission distributions corresponding to the hidden states, and x_{ij} denote the predictor of $\mu_{z_j,ij}$, the state-specific mean read count of Y_{ij}. Under this set-up, the observed data $\mathbf{Y} = \{\mathbf{Y}_{\cdot j}\}_{j=1}^{M}$ follow a multi-sample HMM whose likelihood

function is

$$f(\mathbf{y}|\mathbf{x};\boldsymbol{\Psi}) = \sum_{\mathbf{Z}\in\mathcal{Z}} \left\{ \prod_{k=1}^{2} \left(\pi_k \prod_{i=1}^{N} f_k\left(y_{i1}|x_{i1};\boldsymbol{\psi}_k\right) \right)^{I(Z_1=k)} \times \right.$$
$$\left. \times \prod_{j=2}^{M} \prod_{k=1}^{2} \left(\prod_{i=1}^{N} f_k\left(y_{ij}|x_{ij};\boldsymbol{\psi}_k\right)^{I(Z_j=k)} \right) \prod_{l=1}^{2} \gamma_{lk}^{I(Z_{j-1}=l,Z_j=k)} \right\}, \tag{2.1}$$

where the emission distributions f_1 and f_2 are defined as

$$f_1(y_{ij}|x_{ij};\boldsymbol{\psi}_1) = \Pr(Y_{ij}=y_{ij}|Z_j=1,X_{ij}=x_{ij};\boldsymbol{\psi}_1) = \theta_{ij}\mathrm{I}(y_{ij}=0) + (1-\theta_{ij})\mathrm{NB}\left(y_{ij}|\mu_{1ij},\phi_1\right),$$
$$f_2(y_{ij}|x_{ij};\boldsymbol{\psi}_2) = \Pr(Y_{ij}=y_{ij}|Z_j=2,X_{ij}=x_{ij};\boldsymbol{\psi}_2) = \mathrm{NB}\left(y_{ij}|\mu_{2ij},\phi_2\right), \quad y_{ij}\geq 0. \tag{2.2}$$

Here, $I(\cdot)$ is an indicator function, $\boldsymbol{\Psi} = (\boldsymbol{\pi}^{'},\boldsymbol{\gamma}^{'},\boldsymbol{\psi}^{'})^{'}$, and $\boldsymbol{\psi} = (\boldsymbol{\psi}_1^{'},\boldsymbol{\psi}_2^{'})^{'}$. In addition, $\mathrm{NB}(y_{ij}|\mu_{z_j,ij},\phi_{z_j})$ indicates the NB probability mass function with mean $\mu_{z_j,ij}$ and dispersion ϕ_{z_j} such that $\mathrm{Var}(Y_{ij}) = \mu_{z_j,ij}(1+\mu_{z_j,ij}/\phi_{z_j})$, $\log(\mu_{z_j,ij}) = \beta_{z_j,1} + \beta_{z_j,2}x_{ij}$, $z_j \in \{1,2\}$, and $\log(\theta_{ij}/1-\theta_{ij}) = \lambda_1 + \lambda_2 x_{ij}$. For ChIP-seq experiments with a single input control, we allow the probabilities θ_{ij}, $i = 1,\ldots,N$, to differ across replicates by including the log-transformed total number of ChIP read counts as an offset in the model. This is particularly important as replicates with different amount of mapped reads are likely to have different distributions of observed zeros in the background regions. We describe the EM algorithm to obtain parameter estimates from Equation 2.1 in Section 2.4.

2.3.2 Multi-sample Zero-Inflated Mixed Effects HMM

Here we present the ZIMHMM, an immediate mixed effects extension of the model presented in Section 2.3.1 and a special case of the model proposed by Altman (2007), as it assumes a single sequence of hidden states common to all experiments to ensure the detection of consensus peaks. Let the latent random vector $\mathbf{B} = (B_1,\ldots,B_N)^{'}$ be an N-dimensional vector of sample-specific scalar random effects to be included in the linear model. We will assume that $\mathbf{B} \sim \mathcal{N}_N(\mathbf{0},\sigma^2\mathbf{I})$, where $\mathbf{I}$ denotes an $N \times N$ identity matrix. For better computational stability and efficiency, we will make use of the change of variables $\mathbf{B}$ to random effects $\mathbf{U}$ following the ideas presented by Bates et al. (2014). Define the linear transformation from a N-dimensional spherical random vector, $\mathbf{U}$, to $\mathbf{B}$ as $\mathbf{B} = \sigma\mathbf{U}$, $\mathbf{U} \sim \mathcal{N}_N(\mathbf{0},\mathbf{I}_N)$. We will assume that, conditional on the random effects $\mathbf{U}$, and the Markov chain $\mathbf{Z}$, the observed data $\mathbf{Y} = \{\mathbf{Y}_{\cdot j}\}_{j=1}^{M}$ follow a HMM, and observations from different experiments are independent. In addition, conditionally upon the

unobserved realization u_i of U_i, we model Y_{ij} according to ZINB and NB emission distributions associated with background and enriched states, respectively.

Let r_{ij} denote the design variable associated with the random effects indicating whether the model has either sample-specific random intercept ($r_{ij} = 1$) or random slope ($r_{ij} = x_{ij}$). In addition, let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^{'}$, $\boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})^{'}$, $\boldsymbol{\phi} = (\phi_1, \phi_2)^{'}$, and $\boldsymbol{\Psi} = (\boldsymbol{\pi}^{'}, \boldsymbol{\gamma}^{'}, \boldsymbol{\lambda}^{'}, \boldsymbol{\beta}^{'}, \boldsymbol{\phi}^{'}, \sigma)^{'}$ denote the vectors of all model parameters. The likelihood function of the ZIMHMM is

$$f(\mathbf{y}|\mathbf{r}, \mathbf{x}; \boldsymbol{\Psi}) = \int_{\mathbf{u}\in\mathbb{R}^N} \sum_{\mathbf{z}\in\mathcal{Z}} \left\{ \prod_{k=1}^{2} \pi_k^{I(Z_1=k)} \times \prod_{j=1}^{M} \prod_{k=1}^{2} \prod_{i=1}^{N} f_k \left(y_{ij}|u_i, r_{ij}, x_{ij}; \boldsymbol{\psi}_k, \sigma\right)^{I(Z_j=k)} \times \right.$$
$$\left. \times \prod_{j=2}^{M} \prod_{k=1}^{2} \prod_{l=1}^{2} \gamma_{lk}^{I(Z_{j-1}=l, Z_j=k)} \right\} \times f(\mathbf{u}) d\mathbf{u}, \tag{2.3}$$

where $\boldsymbol{\psi}_1 = (\boldsymbol{\lambda}^{'}, \boldsymbol{\beta}_1^{'}, \phi_1)^{'}$, $\boldsymbol{\psi}_2 = (\boldsymbol{\beta}_2^{'}, \phi_2)^{'}$, and $\boldsymbol{\beta}_{z_j} = (\beta_{z_j,1}, \beta_{z_j,2})^{'}$, for $z_j \in \{1, 2\}$. Here, f_1 and f_2 are defined as in Equation 2.2 with $\log(\mu_{z_j,ij}) = \beta_{z_j,1} + \beta_{z_j,2} x_{ij} + \sigma u_i$, $z_j \in \{1, 2\}$. In ChIP-seq peak calling, a model with random intercepts would account for differences in the sequencing depth of replicates by modeling sample-specific random shifts in the mean model of read counts. Conversely, a random slope model would be particularly interesting when modelling experiments with input controls having differential relationships with the distribution of read counts. Different datasets might exhibit different ChIP-control relationships due to differences in immunoprecipitation (IP) efficiency across experiments (Chen et al., 2015; Lun and Smyth, 2015). While efficient IP shows strong peaks in read coverage at binding sites and a mild control effect (in adjusting for technical variability in enrichment regions), inefficient IP will result in weaker peaks and a larger control effect in enrichment regions, as it is harder to separate technical variability from the true signal in such cases.

Under this model setup, the inclusion of random effects has a critical impact on the marginal covariance structure of read counts. Specifically, it is possible to show that $\text{Cov}(Y_{ij}, Y_{ij'}) \to \kappa > 0$, as $|j - j^{'}| \to \infty$ (Altman 2007; see Appendix A for technical derivations). For the fixed effects model presented in Section 2.3.1, however, such a long-range positive dependence decays to zero. We propose an EM algorithm to estimate the model parameters from the likelihood function in Equation 2.3, which is presented in Section 2.4.

2.4 Estimation

Besides the unknown parameters $\boldsymbol{\Psi} = (\boldsymbol{\pi}', \boldsymbol{\gamma}', \boldsymbol{\lambda}', \boldsymbol{\beta}', \boldsymbol{\phi}', \sigma)'$, the likelihood in Equation 2.3 contains two unobserved quantities: the M-dimensional vector of the state path $\mathbf{Z} \in \mathcal{Z}$, $\mathcal{Z} = \{1,2\}^M$, and the N-dimensional vector of sample-specific random effects $\mathbf{U} \in \mathbb{R}^N$. In the s^{th} step of the EM algorithm, the Q function of the complete data log-likelihood can be written as

$$Q\left(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(s)}\right) = \int_{\mathbf{u}\in\mathbb{R}^N} E\left(\log\left(f(\mathbf{y},\mathbf{z},\mathbf{u}|\mathbf{r},\mathbf{x};\boldsymbol{\Psi})\right)|\mathbf{y},\mathbf{u},\mathbf{r},\mathbf{x};\boldsymbol{\Psi}^{(s)}\right) f\left(\mathbf{u}|\mathbf{y},\mathbf{r},\mathbf{x};\boldsymbol{\Psi}^{(s)}\right) d\mathbf{u}.$$

We make use of the Laplace's approximation to maximize the Q function with respect to $\boldsymbol{\Psi}$. Following the notation presented in Altman (2007), the Q function can be rewritten as (see Appendix A for technical derivations)

$$\begin{aligned} Q\left(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(s)}\right) = \int_{\mathbf{u}\in\mathbb{R}^N} \Bigg\{ & \sum_{k=1}^{2} P\left(Z_1 = k|\mathbf{y},\mathbf{u},\mathbf{r},\mathbf{x};\boldsymbol{\Psi}^{(s)}\right)\log(\pi_k) + \\ & + \sum_{j=1}^{M}\sum_{k=1}^{2}\sum_{i=1}^{N} P\left(Z_j = k|\mathbf{y},\mathbf{u},\mathbf{r},\mathbf{x};\boldsymbol{\Psi}^{(s)}\right)\log\left(f_k(y_{ij}|u_i,r_{ij},x_{ij};\boldsymbol{\psi}_k,\sigma)\right) + \\ & + \sum_{j=2}^{M}\sum_{k=1}^{2}\sum_{l=1}^{2} P\left(Z_{j-1} = l, Z_j = k|\mathbf{y},\mathbf{u},\mathbf{r},\mathbf{x};\boldsymbol{\Psi}^{(s)}\right)\log(\gamma_{lk}) + \log(f(\mathbf{u}))\Bigg\} \times \\ & \times \frac{\left(\mathbf{A}^{(s)}\prod_{j=2}^{M}\mathbf{C}_j^{(s)}\mathbb{I}\right) f(\mathbf{u})}{\int_{\mathbf{u}\in\mathbb{R}^N}\left(\mathbf{A}^{(s)}\prod_{j=2}^{M}\mathbf{C}_j^{(s)}\mathbb{I}\right) f(\mathbf{u})d\mathbf{u}} d\mathbf{u} = \int_{\mathbf{u}\in\mathbb{R}^N} h\left(\mathbf{u};\boldsymbol{\Psi},\boldsymbol{\Psi}^{(s)}\right)\times g\left(\mathbf{u};\boldsymbol{\Psi}^{(s)}\right)d\mathbf{u}, \end{aligned} \tag{2.4}$$

where $\mathbf{A}^{(s)} = \left(A_1^{(s)}, A_2^{(s)}\right)$, $A_k^{(s)} = \pi_k^{(s)} f_k\left(\mathbf{y}_{\cdot 1}|\mathbf{u},\mathbf{r}_{\cdot 1},\mathbf{x}_{\cdot 1};\boldsymbol{\psi}_k^{(s)},\sigma^{(s)}\right)$, $\mathbf{C}_j^{(s)}$ is a 2×2 matrix with elements $C_{j,lk}^{(s)} = \gamma_{lk}^{(s)} f_k\left(\mathbf{y}_{\cdot j}|\mathbf{u},\mathbf{r}_{\cdot j},\mathbf{x}_{\cdot j};\boldsymbol{\psi}_k^{(s)},\sigma^{(s)}\right)$ for all l and k in $\{1,2\}$, and $\mathbb{I}$ is a 2-dimensional vector of ones. The integral in Equation 2.4 is approximated by its integrand evaluated at $\mathbf{u} = \widehat{\mathbf{u}}$ such that $\mathbf{J}_g|_{\mathbf{u}=\widehat{\mathbf{u}}} = \mathbf{0}$. Here, $\mathbf{J}_g|_{\mathbf{u}=\widehat{\mathbf{u}}}$ denotes the Jacobian of the function g evaluated at $\mathbf{u} = \widehat{\mathbf{u}}$. Note that neither g nor its Hessian matrix depends on $\boldsymbol{\Psi}$.

In the E-step, we compute $\widehat{\mathbf{u}}$ via numerical optimization of g using the BOBYQA algorithm (Powell, 2009). The posterior probabilities from Equation 2.4 can be calculated by a standard Forward-Backward algorithm (Rashid et al., 2014). In the M-step, the Q function is maximized with respect to the unknown

parameters $\mathbf{\Psi}$. It is possible to show (see Appendix A) that one can approximate the Q function as

$$
\begin{aligned}
Q\left(\mathbf{\Psi}|\mathbf{\Psi}^{(s)}\right) \approx & \sum_{k=1}^{2} P\left(Z_1 = k|\mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \mathbf{\Psi}^{(s)}\right) \log(\pi_k) + \\
& + \sum_{j=1}^{M}\sum_{k=1}^{2}\sum_{i=1}^{N} P\left(Z_j = k|\mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \mathbf{\Psi}^{(s)}\right) f_k(y_{ij}|\widehat{u}_i, r_{ij}, x_{ij}; \boldsymbol{\psi}_k, \sigma) + \\
& + \sum_{j=2}^{M}\sum_{k=1}^{2}\sum_{l=1}^{2} P\left(Z_{j-1} = l, Z_j = k|\mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \mathbf{\Psi}^{(s)}\right) \log(\gamma_{lk}). \qquad (2.5)
\end{aligned}
$$

In this setting, one can obtain closed forms for the estimates of the initial and transition probabilities. We perform conditional maximizations to compute estimates of $(\boldsymbol{\psi}_1', \boldsymbol{\psi}_2')'$, and σ using the BFGS algorithm (Fletcher, 2013). The EM algorithm iterates until the maximum absolute relative change in the parameter estimates three iterations apart is less than 10^{-3} for three consecutive iterations. For better efficiency, we use a rejection-controlled EM (RCEM; Ma et al. 2006) with threshold 0.05 and a weighted maximization approach on aggregated data. The final set of posterior probabilities can be used to determine the hidden path of the states $\mathbf{Z}$ and segment the genome into either enriched or background windows. By denoting $p_j = P\left(Z_j = 1|\mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \mathbf{\Psi}^{(s)}\right)$ the probability of window j belonging to background, one could classify window j to be enriched if $p_j \leq \alpha$, where α is chosen such that the total false discovery rate (FDR) is $\sum_{j=1}^{M} p_j I(p_j \leq \alpha) / \sum_{j=1}^{M} I(p_j \leq \alpha)$ (Efron et al., 2001). Alternatively, the Viterbi algorithm (Viterbi, 1967) can be used to determine the most likely sequence of background and enrichment windows without the need of a subjective choice of an FDR threshold. Finally, regions of enrichment are created by merging adjacent windows either meeting a cutoff α or belonging to the same Viterbi's predicted state.

2.5 Simulation Study

In this study, we evaluated the performance of the ZIMHMM under a set of different scenarios where experimental replicates differed with respect to sequencing depth and ChIP-input control relationship. We compared the ZIMHMM to its fixed-effects version (ZIHMM) and to a naïve multi-sample HMM that does not account for zero-inflation (HMM). For each scenario, we simulated a hundred ChIP-seq multi-sample data under random intercept and random slope models mimicking the main characteristics of H3K27me3 ChIP-seq data. First, we generated a sequence of hidden states with length $M = 25,000$ from a first-order Markov chain with two states and transition probabilities $\gamma_{11} = \gamma_{22} = 0.95$ to ensure broad background and enrichment regions. Secondly, for a given path of states, a set of N input control read counts was independently simulated

following a NB distribution with parameters $(\mu, \phi)^{'} = (9, 2.5)^{'}$. Thirdly, N sequences of ChIP read counts with length M was simulated as a function of the log-transformed input control counts following a mixture of random effects ZINB and NB distributions. Here, we simulate data under scenarios with $N = \{2, 3, 6, 9\}$ ChIP-seq replicates and explored scenarios with low, medium, and high levels of heterogeneity across the N simulated ChIP replicates. These levels of heterogeneity are represented by different values of the variance component σ^2 for both the random intercept and random slope models (see Figure C.1 in Baldoni et al. 2019b).

2.5.1 Simulation Results

Table 2.1 shows the true values, the sample median, 25^{th}, and 75^{th} percentiles of the parameter estimates from simulated data relative to scenarios with low level of heterogeneity and random intercept model. The median values of the estimates associated with the parameters from the mean model $(\boldsymbol{\lambda}^{'}, \boldsymbol{\beta}_1^{'}, \boldsymbol{\beta}_2^{'})^{'}$ appeared to be symmetric and centered at the true values, suggesting that the proposed Laplace approximation works relatively well even for a small number of replicates. The estimates of the variance component were close to the true values in all simulated scenarios. We present the median observed true and false positive rates (TPR and FPR, respectively) based on the sequence of predicted states by the Viterbi algorithm. Regardless of the number of replicates, the ZIMHMM performed well when predicting the path of hidden states. We observed that its classification performance improved in scenarios with higher number of replicates, as expected. This is particularly important as a common practice in the analysis of multiple ChIP-seq data is to call peaks utilizing two replicates only. In the analyzed scenario, integrating data from additional replicates improved the detection of enrichment regions in consensus.

The simulation results indicated that estimates associated with dispersion parameters $(\phi_1, \phi_2)^{'}$ were biased even for scenarios with a high number of ChIP replicates. An extensive statistical literature makes reference to biased estimates of the dispersion parameter in the NB regression model and proposes possible corrections to it (Robinson and Smyth, 2007). Here, given the good classification performance of the ZIMHMM regarding the TPR and FPR across all different simulated scenarios, we did not explore alternative solutions to the estimation of the dispersion parameter as this investigation would be beyond the scope of this work. Nonetheless, we believe that such a correction would lead to better precision for the parameter estimates. Thresholding posterior probabilities with different FDR levels allowed us to compare the performance of the ZIMHMM with the ZIHMM and HMM. The ZIMHMM had a better classification performance than the

Table 2.1: Median (first, and third quantiles) of parameter estimates under random intercept models (low heterogeneity).

Parameter	True value	Two rep.	Three rep.	Six rep.	Nine rep.
β_{11}	1.50	1.63 (1.31, 1.97)	1.46 (1.30, 1.69)	1.50 (1.40, 1.66)	1.51 (1.40, 1.67)
β_{12}	0.75	0.75 (0.74, 0.75)	0.75 (0.75, 0.75)	0.75 (0.75, 0.75)	0.75 (0.75, 0.75)
β_{21}	2.50	2.66 (2.34, 3.02)	2.47 (2.31, 2.71)	2.50 (2.40, 2.67)	2.52 (2.41, 2.67)
β_{22}	0.50	0.50 (0.50, 0.50)	0.50 (0.50, 0.50)	0.50 (0.50, 0.50)	0.50 (0.50, 0.50)
ϕ_1	5.00	4.95 (4.81, 5.01)	4.71 (4.34, 4.93)	4.38 (4.02, 4.66)	4.22 (3.90, 4.47)
ϕ_2	2.50	2.49 (2.44, 2.50)	2.43 (2.33, 2.48)	2.34 (2.24, 2.42)	2.30 (2.21, 2.37)
λ_1	-0.75	-0.75 (-0.78, -0.72)	-0.75 (-0.78, -0.71)	-0.75 (-0.77, -0.73)	-0.75 (-0.77, -0.74)
λ_2	-0.60	-0.60 (-0.61, -0.58)	-0.60 (-0.61, -0.59)	-0.60 (-0.61, -0.59)	-0.60 (-0.60, -0.59)
σ^2	0.10	0.12 (0.00, 1.47)	0.09 (0.02, 0.28)	0.11 (0.05, 0.23)	0.12 (0.06, 0.18)
TPR		0.94 (0.94, 0.95)	0.96 (0.96, 0.97)	0.98 (0.98, 0.99)	0.99 (0.99, 0.99)
FPR		0.07 (0.07, 0.08)	0.05 (0.04, 0.05)	0.02 (0.02, 0.02)	0.01 (0.01, 0.01)

misspecified models ZIHMM and HMM in all the scenarios (see Figure 2.2). However, we observed a higher relative performance of the ZIMHMM over the ZIHMM and HMM when a low number of replicates was analyzed. In the context of heterogeneous replicates, these results suggest that accounting for sample-specific biases boosts the detection of consensus regions of enrichment and its improvement is particularly significant when only a few replicates are available.

2.6 Data Applications

We applied the ZIMHMM with sample-specific random intercepts to detect consensus regions of enrichment from multiple ChIP-seq experiments of H3K27me3 and H3K36me3 marks from the ENCODE Consortium and the Roadmap Epigenomics Project. Data were analyzed in two different scenarios. In Section 2.6.1, we report results from the analysis of technical replicates from H3K36me3 and H3K27me3 experiments of Huvec and Nhek cell lines, respectively. In this standard scenario of multi-sample ChIP-seq peak calling, technical replicates are expected to show low spatial heterogeneity regarding the signal profile across the genome. In Section 2.6.2, we present results of the analysis of H3K36me3 and H3K27me3 experiments from white blood cell lines CD4 memory, CD4 naïve, CD8 naïve, and CD34 primary cells. In this scenario, white blood cell lines are assumed to be similar but show a certain level of heterogeneity regarding the signal profile and genomic locations of protein-DNA binding sites.

We sought to benchmark the genome-wide performance of the ZIMHMM to the multi-sample peak callers JAMM and Zerone, as well as single-sample peak callers under the pooling approach BCP-P, CCAT-P, MACS2-P, MOSAiCS-P, and RSEG-P. We compared methods regarding peak accuracy, broadness, coverage

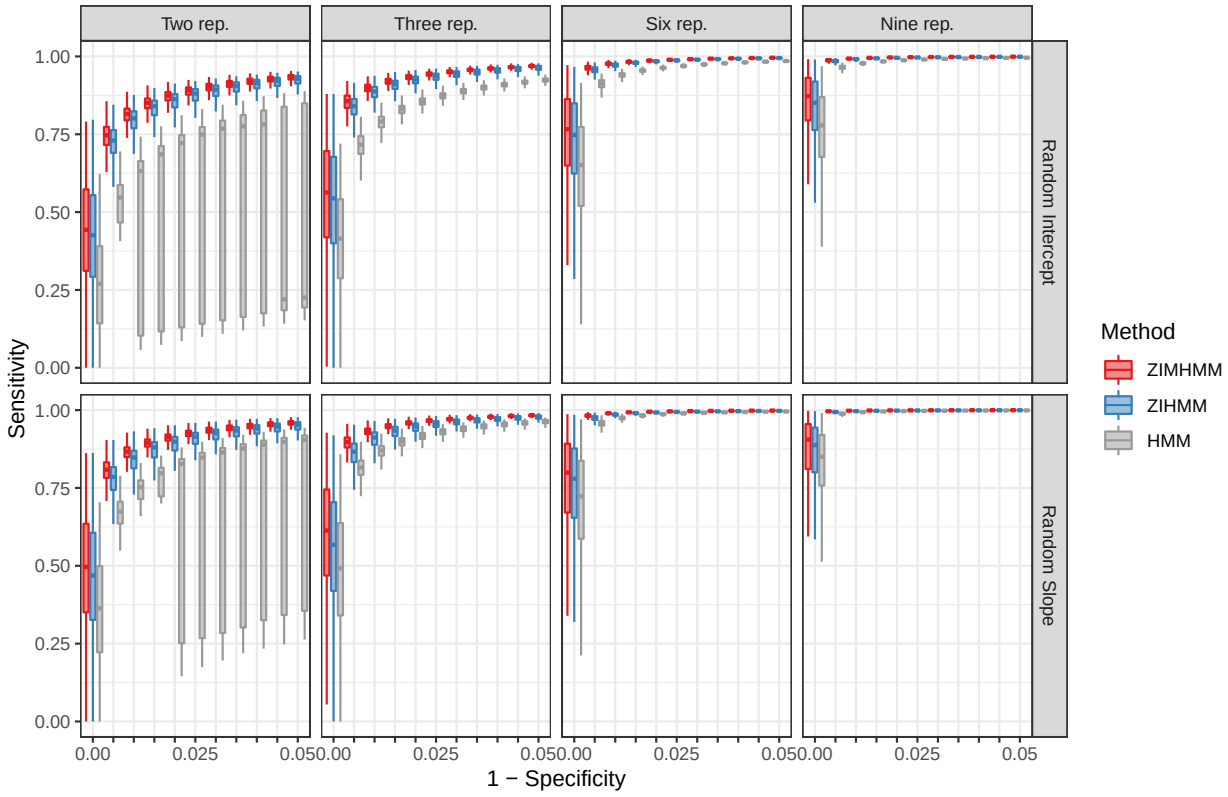

Figure 2.2: Classification performance of the proposed models on simulated random effects data (top: intercepts, bottom: slopes) for two, three, six, and nine ChIP-seq experimental replicates assuming a low level of heterogeneity across experiments.

of the observed read density from both analyzed marks, coverage of active and inactive genomic regions, and running time. To assess the benefits of the random effects approach, results from the fixed effects model ZIHMM presented in Section 2.3.1 are shown. Read counts were computed using non-overlapping windows of 500bp in both scenarios. For the ZIMHMM and the ZIHMM, enrichment regions were defined by merging neighboring predicted enriched windows using the Viterbi algorithm. A discussion about the choice of the window size and a comparison between the Viterbi algorithm and the FDR thresholding approach is presented in Section 2.6.1.

2.6.1 Analysis of ChIP-seq Data From Technical Replicates

For benchmarking purposes, we created a set of measures and associations that were first introduced by Xing et al. (2012) (Table 2.2). First, we calculated the median size of called peaks (in kbp) by each method. For both analyzed marks, we observed that the ZIMHMM called substantially broader regions of enrichment than the multi-sample peak callers JAMM and Zerone, but narrower than the regions of the single-sample peak callers BCP-P and RSEG-P. Next, we defined the read coverage as the proportion of reads from the analyzed mark mapped on called peaks out of the total number of mapped reads. Read counts were previously normalized by the median log-ratios of each sample over the geometric mean (after adding 1 pseudo count to avoid undefined ratios in windows with zero counts). Results showed that the ZIMHMM covered most of the mapped reads while still maintaining a low size of peak calls. While RSEG-P had a reasonable coverage of counts, its called peaks were often excessively large and did not capture minor changes in the signal profile (Figure 2.3). This is a known characteristic of the pooling type of analysis of single-sample peak callers and the results highlight the improved accuracy of the peaks called by the multi-sample peak caller the ZIMHMM. Here, the ZIHMM was fitted using the total sum of read counts as an offset to attempt the correction of differences in sequencing depth across replicates. However, the inclusion of replicate-specific random effects led to a better coverage of read counts across the genome.

To asses whether the high sensitivity of the ZIMHMM was indeed due to an improved segmentation, we computed empirical TPRs and FPRs based on the coverage of actively transcribed genes and reads of the reverse mark (see Figure 2.5). Histones H3K36me3 and H3K27me3 are known to be associated with gene transcription and repression, respectively. For the former (latter), enrichment regions are usually deposited on genes with high (low) expression and are nearly mutually exclusive, although the activity of H3K27me3 can also be seen in genomic regions without any gene bodies (Xing et al. 2012; Figure 2.1).

Table 2.2: Genome-wide peak calls and common associations for ChIP-seq data of H3K36me3 and H3K27me3 marks from three technical replicates of Huvec and Nhek cells, respectively. The running time of each method is shown in hours.

				Coverage			
Mark	Method	Peaks	Median Size	Reads	Active Regions	Inactive Regions	Time
H3K36me3	BCP-P	6852	29.298	0.400	0.497	0.027	1.618
	CCAT-P	94181	1.026	0.345	0.345	0.015	17.642
	JAMM	66751	0.300	0.123	0.105	0.007	5.376
	MOSAiCS-P	3626	17.704	0.184	0.178	0.004	0.512
	MACS2-P	53950	1.616	0.353	0.356	0.018	0.132
	RSEG-P	8259	33.204	0.470	0.623	0.043	1.659
	Zerone	16913	7.322	0.336	0.346	0.016	0.024
	ZIHMM	14867	18.064	0.508	0.682	0.049	0.324
	ZIMHMM	12574	22.948	0.517	0.709	0.055	6.336
H3K27me3	BCP-P	6618	16.114	0.412	0.032	0.147	1.335
	CCAT-P	193893	0.978	0.504	0.034	0.165	30.758
	JAMM	109855	0.303	0.253	0.012	0.058	6.925
	MOSAiCS-P	4726	4.090	0.159	0.004	0.024	1.829
	MACS2-P	89258	1.147	0.394	0.019	0.100	0.148
	RSEG-P	12801	20.997	0.564	0.047	0.246	0.981
	Zerone	34397	1.465	0.240	0.008	0.040	0.027
	ZIHMM	51276	5.859	0.622	0.053	0.262	0.642
	ZIMHMM	54994	5.845	0.634	0.056	0.272	12.307

Using RNA-seq data, we determined cell line-specific actively transcribed genes and computed the coverage of active and inactive regions. Specifically, we used RNA-seq experimental data from the ENCODE Consortium on Nhek and Huvec human cells to define sets of actively transcribed genes in each cell as follows. First, we used Salmon (Patro et al., 2017) to quantify transcript expression from cell-specific RNA-seq experiments. We then calculated, using the R package $tximport$ (Soneson et al., 2015), estimated counts using abundance estimates (transcripts per million, TPM) scaled up to the average transcript length over samples and library size. This step ensures that counts computed from Salmon are not correlated with the average transcript length. Secondly, we defined the set of actively transcribed genes in each cell by fitting a two-components Gaussian mixture model on the log-transformed TPM (LTPM) and selecting genes with an LTPM greater than the estimated mean of the upper Gaussian component. Finally, using the genomic ranges of the actively transcribed genes, we calculate the coverage of active and inactive genomic regions as empirical measure of TPR and FPR, respectively, for H3K36me3 peaks. For H3K27me3 peaks, the coverage of active and inactive regions are taken to be empirical measures of FPR and TPR respectively. Here, we

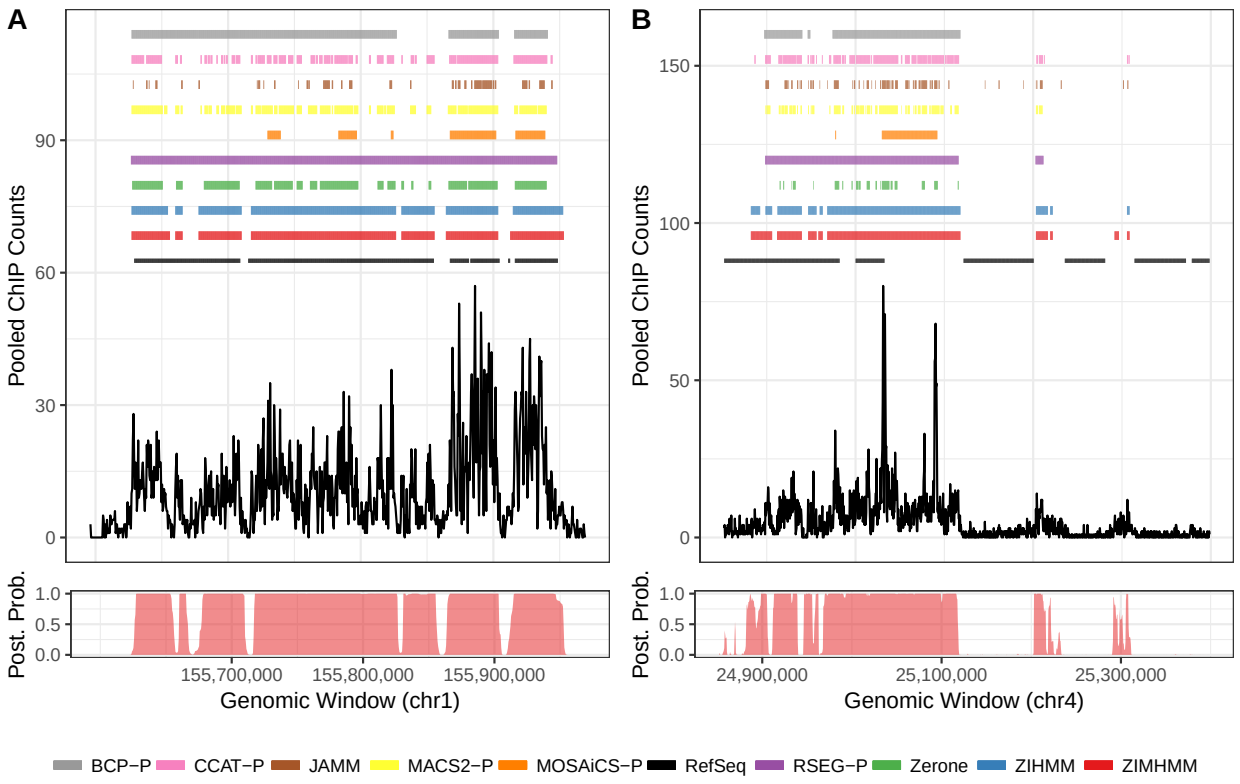

Figure 2.3: Pooled read counts of three technical replicates of histone modifications H3K36me3 (A) and H3K27me3 (B) on human cells Huvec and Nhek, respectively. At the top, called peaks from benchmarked methods. At the bottom, posterior probabilities of enrichment from ZIMHMM, which calls broad peaks in consensus that better associate with the read counts profile from the analyzed diffuse marks. This figure appears in color in the electronic version of this article.

define an inactive region to be any genomic region not overlapping an actively transcribed gene, which includes intergenic regions and inactive genes.

Here, we define an inactive region to be any genomic region not overlapping an actively transcribed gene, which includes intergenic regions and inactive genes. We observed that the ZIMHMM had the highest coverage among all methods. Its called peaks for H3K36me3 (H3K27me3) covered most of the active (inactive) locations, respectively, while still maintaining low false positives. Both multi-sample peak callers JAMM and Zerone performed poorly under this scenario regarding these metrics (Figure 2.3). Here, single-sample peak callers had mixed performances and called peaks that were either excessively large and expanded multiple actively transcribed gene bodies (BCP-P and RSEG-P) or overly segmented (CCAT-P, MACS2-P, and MOSAiCS-P).

We observed that peak callers varied substantially regarding their computational time. MACS2-P, Zerone, and ZIHMM were among the fastest methods under comparison taking no longer than an hour to analyze the entire human genome. Conversely, CCAT-P, JAMM, and ZIMHMM were the peak callers that took longer to complete the analysis. The approximate running time of the ZIMHMM was six and twelve hours to analyze three replicates of H3K36me3 and H3K27me3, respectively. Conversely, CCAT-P had an approximate running time of 18 and 28 hours for these marks, respectively. It is worth noting that single-sample peak callers such as BCP-P, MOSAiCS-P, RSEG-P, and MACS2 are in general faster than multi-sample peak callers simply by the fact that technical replicates are pooled together and analyzed as a single experiment. We believe that the performance of the ZIMHMM can be further improved and will be left as a project in a future implementation of the model.

The performance of peak callers under different choices of window sizes was investigated. Results were consistent across windows of 250bp, 500bp, 750bp, and 1000bp, although peaks from ZIMHMM became larger for wider window sizes (see Tables D.1-D.3 in Baldoni et al. 2019b). In Ibrahim et al. (2014), the authors propose the use of a cost function to select the window size. Here, we choose to report results based on the window size of 500bp calculated as a function of the average fragment length, an approach also used by MACS2. Moreover, we compared peaks called by the ZIMHMM via both the Viterbi algorithm and FDR thresholding. The Viterbi peaks were similar regarding the metrics used in this paper to peaks based on a FDR cutoff of 0.05. An increasing (decreasing) trend in sensitivity (specificity) across the different thresholds was observed (see Table 2.3 and Tables D.4-D.7 in Baldoni et al. 2019b).

Table 2.3: Genome-wide performance of ZIMHMM with Viterbi and FDR thresholding methods (Window size 500bp).

				Coverage			
Mark	Method	Peaks	Median Size	Reads	Active Regions	Inactive Regions	Time
H3K36me3	Viterbi	12574	22.948	0.517	0.735	0.066	0
	FDR = 0.01	15162	17.089	0.508	0.711	0.061	0
	FDR = 0.05	14611	19.528	0.527	0.751	0.071	0
	FDR = 0.10	15315	19.041	0.540	0.780	0.080	0
	FDR = 0.20	23775	7.812	0.566	0.823	0.101	0
H3K27me3	Viterbi	54994	5.845	0.634	0.053	0.269	0
	FDR = 0.01	70362	3.905	0.590	0.043	0.225	0
	FDR = 0.05	67243	4.395	0.634	0.053	0.266	0
	FDR = 0.10	69724	4.395	0.660	0.060	0.292	0
	FDR = 0.20	81648	4.102	0.699	0.074	0.336	0

We compared the performance of the ZIMHMM under the whole-genome analysis presented in this paper with peaks called chromosome-wise. We observed a better sensitivity/specificity of the whole-genome analysis over the chromosome-wise analysis for small chromosomes (see Figure 2.4 and Figures D.3 and D.4 in Baldoni et al. 2019b). A possible explanation for the increase in performance is that small chromosomes may have less data to better resolve peak regions. In addition, chromosomes with less gene activity are likely to have fewer enrichment regions for certain marks. The whole-genome analysis could be a workaround for a potential convergence issues in a chromosome where most of the reads are coming from background.

2.6.2 Analysis of ChIP-seq Data From Multiple Cell Lines

We analyzed data from CD4 memory, CD4 naive, CD8 naive, and CD34 mobilized primary cell lines from the Roadmap Project. We expected these cell lines to be heterogeneous regarding the enrichment profile of read counts and, therefore, served as a basis for a sensitivity analysis for the benchmarked consensus peak callers. The measures presented in Section 2.6.1 were also used in this scenario. Using RNA-seq data, genes were considered to be actively transcribed in consensus across cell lines if they were simultaneously active in all white blood cells. Specifically, we downloaded RNA-seq experiments from the Roadmap Project on human white blood cells CD4 memory, CD4 naive, CD8 naive, and CD34 mobilized primary cells, and quantified the transcript expression using Salmon. Then, using abundance estimates adjusted for transcript length and library size, we measured the log-transformed TPM (LTPM) and fitted a two-component Gaussian mixture regression model on the LTPM values of the set of Ensembl genes (Zerbino et al., 2017) to define the set of actively transcribed genes. A two-component Gaussian mixture model was fitted on the gene-level LTPM values of all four distinct human cell lines and genes were classified to be actively transcribed if their cell-specific LTPM values were uniformly above the larger estimated mean across all four cell lines. Results are presented in Table 2.4.

In this analyzed scenario, we observed that peak callers performed similarly for the H3K36me3 mark regarding the coverage of read counts, although BCP-P and MOSAiCS-P had a slightly higher coverage of actively transcribed gene bodies than the ZIMHMM. However, regions called by these two methods were consistently larger than actual gene bodies and did not show a reasonable spatial resolution when detecting minor changes in enrichment profile across cells (see Figures D.1 and D.2 Baldoni et al. 2019b). As noted, these are known characteristics of the pooling type of analysis from single-sample peak callers. For H3K27me3, we observed a significant improvement of the ZIMHMM over current multi- and single-sample

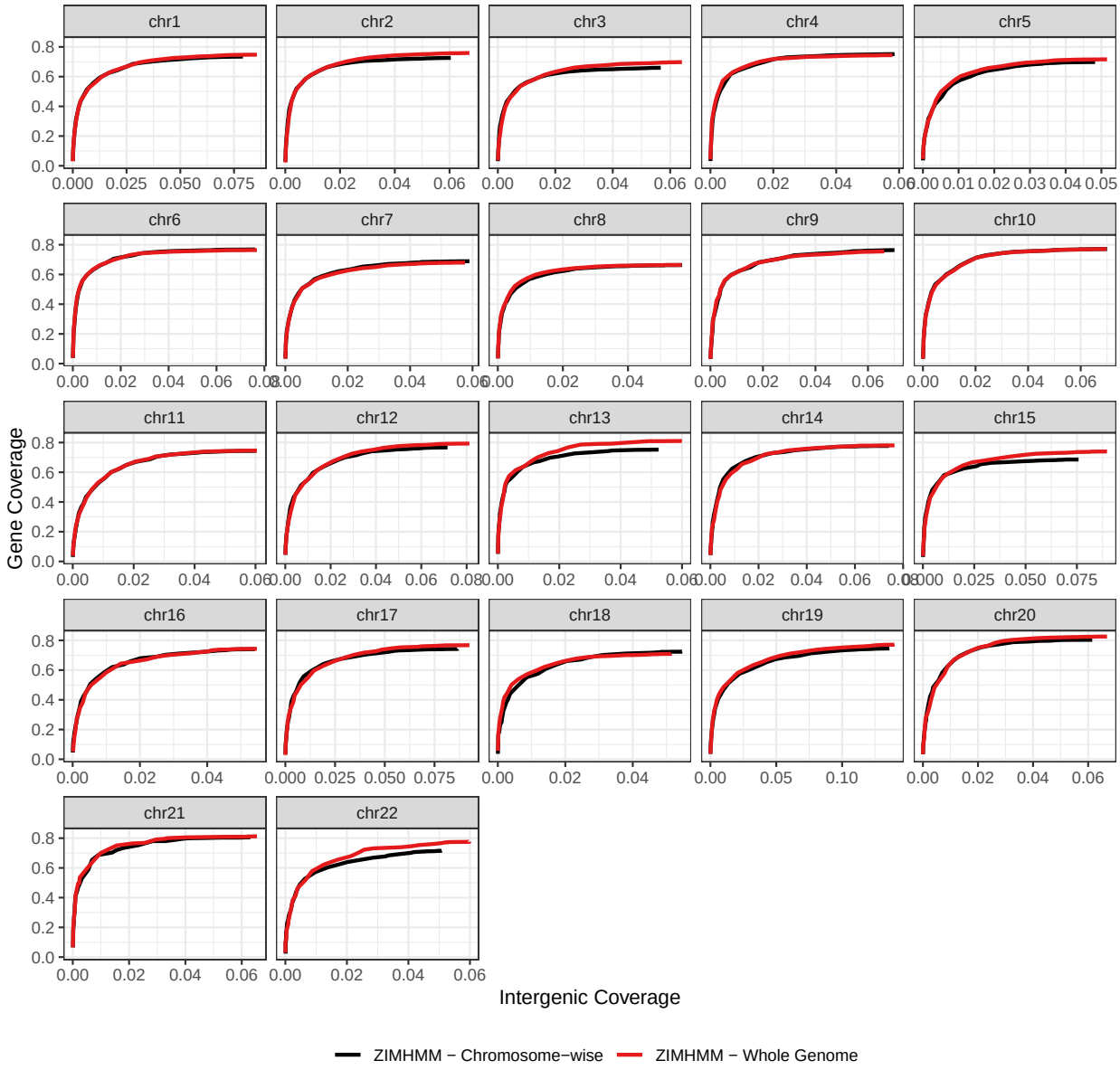

Figure 2.4: Comparative Performance of Whole Genome Analysis and Chromosome-Wise Analysis of ZIMHMM Peak Calls From H3K36me3 of Huvec Cells (Window Size 500bp)

Table 2.4: Genome-wide peak calls and common associations for ChIP-seq data of H3K36me3 and H3K27me3 marks from CD4 memory primary, CD4 naive primary, CD8 naive primary, and CD34 mobilized primary cell lines. The running time of each method is shown in hours.

				Coverage			
Mark	Method	Peaks	Median Size	Reads	Active Regions	Inactive Regions	Time
H3K36me3	BCP-P	8572	26.368	0.331	0.595	0.030	3.150
	CCAT-P	38735	1.075	0.131	0.150	0.003	11.247
	JAMM	72470	0.571	0.219	0.317	0.012	5.955
	MOSAiCS-P	15941	12.573	0.334	0.579	0.029	1.586
	MACS2-P	64331	1.478	0.310	0.489	0.025	1.036
	RSEG-P	6936	27.833	0.280	0.478	0.021	4.033
	Zerone	28913	2.930	0.210	0.289	0.009	0.024
	ZIHMM	31852	5.370	0.345	0.578	0.032	0.588
	ZIMHMM	29747	5.371	0.328	0.538	0.028	20.183
H3K27me3	BCP-P	6872	12.208	0.191	0.015	0.120	2.379
	CCAT-P	8725	1.026	0.028	0.001	0.007	3.012
	JAMM	118528	0.295	0.106	0.009	0.054	8.123
	MOSAiCS-P	16630	8.508	0.190	0.015	0.107	1.503
	MACS2-P	91632	0.792	0.158	0.012	0.077	1.113
	RSEG-P	855	13.673	0.029	0.004	0.014	10.578
	Zerone	29304	2.929	0.118	0.010	0.061	0.038
	ZIHMM	58117	8.301	0.520	0.071	0.394	0.947
	ZIMHMM	51655	9.277	0.543	0.076	0.424	12.559

peak callers regarding the coverage of read counts and gene bodies. Specifically, benchmarked methods covered no more than 20% of the mapped reads and had a low genome-wide coverage of inactive regions. In this scenario, accounting for cell line-specific shifts in the signal profile of read counts significantly improved the detection of enrichment regions in consensus across cells. Here, the ZIMHMM was more time consuming than other approaches, specially single-sample peak callers that call peaks with pooled

2.6.3 Association of H3K36me3, H3K27me3, and Gene Expression

We further compared peak callers regarding the genome-wise association of peaks with gene expression data as well as the coverage of the reads from the opposite mark. Called peaks were sorted with respect to the number of mapped reads and the coverage of active and inactive regions by the top- and bottom-most peaks were calculated, respectively. Peaks were also sorted regarding their read counts and the coverage of H3K27me3 and H3K36me3 reads mapped onto the top- and bottom-most peaks, respectively, was calculated. These quantities provide measures of association between the two analyzed marks and their role on gene

activation and suppression. In all the scenarios, read counts were previously normalized by the median log-ratios as in Section 2.6.1. Results are presented in Figure 2.5.

Overall, top peaks called by the ZIMHMM had a superior performance than all other methods in most of the scenarios. The proposed model covered more of actively transcribed gene bodies and read counts for H3K36me3 and H3K27me3, respectively. We observed that the performance of all methods but CCAT-P, JAMM, and Zerone was homogeneous when calling H3K36me3 peaks from white blood cells. Both multi-sample peak callers performed poorly for the two analyzed diffuse marks in all the scenarios.

2.7 Discussion

Here, we presented the ZIMHMM, a statistical model tailored to call broad peaks in consensus across multiple ChIP-seq technical or biological replicates. The ZIMHMM models the excess of zeros of broad and diffuse marks and accounts for sample differences via random effects.

The ZIMHMM should be applied in multiple biological or technical ChIP-seq replicates with broad regions of signal, such as those pertaining to epigenomic marks. Methods focused on peak calling from multiple samples are of growing interest given the reduction of sequencing costs and higher data availability. Prior work from multi-sample peak callers has shown the benefits of data integration in ChIP-seq data analysis. However, there is no consensus in the literature on how to integrate results from multiple replicates and current approaches perform poorly in finding epigenomic marks with broad peaks. In this paper, we analyzed H3K36me3 and H3K27me3, marks that are associated with gene activation and gene suppression, respectively. For the former mark, in particular, enrichment regions detected by the ZIMHMM better associated with activated gene bodies than any other benchmarked peak caller. These results could trigger, for instance, new insights to investigators interested in detecting cell-specific activated genes, for instance. The ZIMHMM is comparable to most of the current peak callers in terms of computing time and has been implemented into an R package that is available for download (see Appendix A for details).

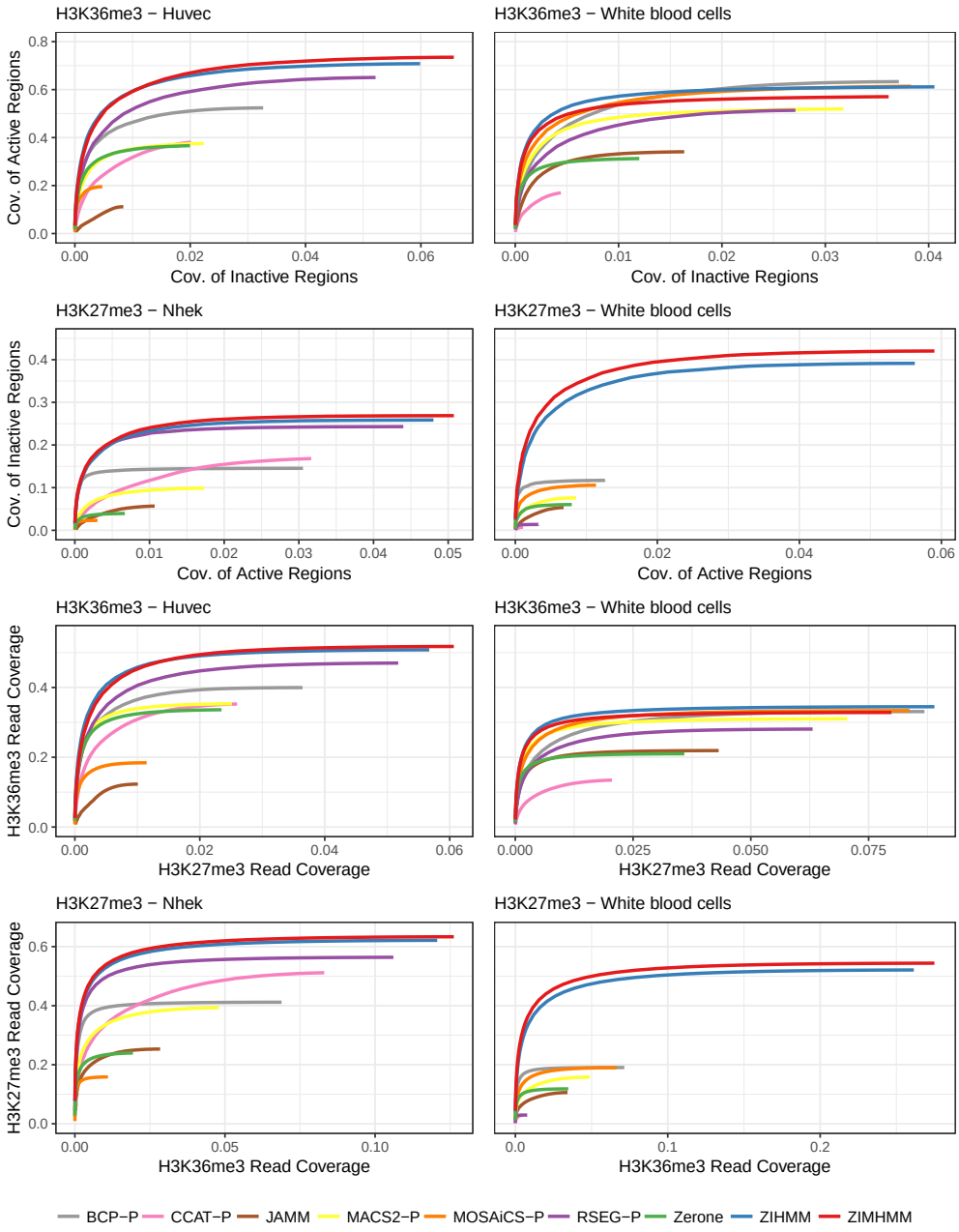

Figure 2.5: Genome-wide performance of ZIMHMM and other peak callers. We analyzed diffuse histone modifications H3K36me3 and H3K27me3 under scenarios of technical replicates and multiple cell lines. ZIMHMM showed superior performance in most of the scenarios, better associating with gene expression and read counts than other methods. Overall, peaks called by ZIMHMM showed a reasonably low number of false positives, here characterized by the coverage of inactive (active) regions by H3K36me3 (H3K27me3) peaks and coverage of reads from the other mark. This figure appears in color in the electronic version of this article.

CHAPTER 3: EFFICIENT DETECTION AND CLASSIFICATION OF EPIGENOMIC CHANGES UNDER MULTIPLE CONDITIONS

3.1 Introduction

Epigenomics, the study of the human genome and its interactions with proteins and other cellular elements, has become of significant interest in recent years. Such interactions have been shown to regulate essential cellular functions such as gene expression and DNA packaging (Kim et al., 2018), resulting in downstream phenotypic impact. Therefore, the interrogation of how these interactions may change across conditions, such as cell types or treatments, is of marked interest in biomedical research. Several landmark articles have identified specific genomic regions of changing (differential) epigenomic activity between conditions as drivers of cell differentiation (Creyghton et al., 2010), cancer progression (Varambally et al., 2002), and a number of human diseases (Portela and Esteller, 2010). Within differential regions, the delineation of specific patterns of change across conditions is also of interest, for example classifying the gain-of- or loss-of-activity in genomic loci due to treatment (Clouaire et al., 2014). The identification of specific combinations of processes acting locally may also be informative, such as for segmenting the genome into regulatory states (Kundaje et al., 2015).

To quantify local epigenomic activity, a common high-throughput assay is chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq). ChIP-seq experiments begin with cross-linking DNA and proteins within chromatin structures, which are then fragmented by sonication in a particular sample. DNA fragments bound to the protein of interest are isolated by chromatin immunoprecipitation, which are then sequenced via massively parallel high-throughput sequencing to generate short sequencing reads pertaining to the original fragments. Sequences are then mapped onto a reference genome through sequence alignment to determine their likely locations of origin. Genomic coordinates containing a high density of mapped reads, often referred to as enrichment regions (peaks), indicate likely locations of protein-DNA interaction sites, and all other regions are referred to as background regions. This local read density is often summarized by counting the number of reads mapped onto non-overlapping windows of fixed length tiling

the genome (window read counts), forming the basis for downstream analyses. Across multiple conditions, regions exhibiting enrichment in at least one condition, but not across all conditions, indicate the presence of differential activity pertaining to the protein-DNA interaction of interest.

To date, many differential peak callers (DPCs) have been proposed (Song and Smith, 2011; Stark and Brown, 2011; Shen et al., 2013; Chen et al., 2015; Lun and Smyth, 2015; Allhoff et al., 2016). However, several challenges affect their ability to accurately detect regions of differential activity from the wide range of ChIP-seq experiments (Section 3.2). First, differential regions may be both short or broad in length, causing difficulty for methods optimized for a particular type of signal profile (Stark and Brown, 2011; Chen et al., 2015). Second, methods that pool experimental replicates together (Song and Smith, 2011) often exhibit more false positive calls compared to methods that jointly model replicates from each condition (Steinhauser et al., 2016). Third, the analysis of ChIP-seq data is often subject to complex biases that may vary across the genome, as differences in local read enrichment may depend on the total read abundance in a given region. DPCs that solely rely on sample-specific global scaling factors or control subtraction methods (Stark and Brown, 2011; Shen et al., 2013; Chen et al., 2015; Allhoff et al., 2016) may be prone to detecting spurious differences due to the lack of non-linear normalization methods (Lun and Smyth, 2015). Reflecting these limitations, a recent comparison of DPCs demonstrated that current methods tend to detect either a large number of short peaks (low sensitivity) or exhibit a high number of false positive calls (low specificity) in ChIP-seq experiments with broad regions of enrichment (Steinhauser et al., 2016). Moreover, few methods are able to simultaneously test for differential activity across three or more conditions (Chen et al., 2015; Lun and Smyth, 2015), or can classify specific differential combinatorial patterns. Altogether, these limitations can impact the drawing of accurate insights from modern epigenomic studies.

Here, we propose an efficient and flexible statistical method to identify differential regions of enrichment from epigenomic experiments with diverse signal profiles and collected under common multi-replicate, multi-condition settings. Our method overcomes the limitations of current DPCs with three major features. First, it uses a hidden Markov model (HMM) to account for the diversity in differential enrichment profiles that may result from short and broad epigenomic ChIP-seq data sets. Second, it captures specific differential combinatorial patterns through a novel finite mixture model emission distribution within the HMM's differential state. Each mixture component pertains to a particular differential combinatorial pattern that is formed by the presence or absence of local enrichment across conditions, where a generalized linear model (GLM) is used to model the specific differential combinatorial pattern while accounting for sample-

and window-specific normalization factors via offsets. Third, it enables the simultaneous detection and classification of epigenomic changes under three or more conditions, a novelty not yet available in any other DPC algorithm. The presented method offers additional benefits over current HMM-based DPC algorithms (Song and Smith, 2011; Allhoff et al., 2016) that include a GLM-based framework with an embedded mixture model, which allows the modeling of covariates of interest as well as the inclusion of model offsets for non-linear normalization, and a fast and accurate parameter estimation scheme via rejection-controlled EM algorithm (RCEM).

3.2 Data

Histones are proteins that interact and condense DNA in eukaryotic cells into structural units called nucleosomes. Multiple types of enzymatic modifications may be applied to histones, resulting in changes in local DNA packaging and chromatin accessibility mediated by nucleosomes (Bannister and Kouzarides, 2011). In turn, cellular processes such as gene transcription, gene silencing, DNA repair, replication, and recombination are also affected. Proteins that interact with DNA and alter its functional properties are often referred to as epigenomic marks. For example, the trimethylation of histone H3 at lysines 36 and 27 (H3K36me3 and H3K27me3) are two types of histone modifications that tend to occur in genomic loci containing actively transcribed and repressed genes (Liu et al., 2016), respectively, and exhibit broad enrichment profiles. These marks have been investigated in cancer studies, where their absence is often observed in multiple cancer types (Wei et al., 2008). As a result, H3K36me3 and H3K27me3 are considered to be key prognostic indicators in patients with breast, ovarian, and pancreatic cancer. EZH2, a major component of the polycomb complex PRC2 that catalyzes the methylation of H3K27me3 (Margueron and Reinberg, 2011), is another example of a protein with experimental signal characterized by broad enrichment domains and co-occurs with the activity of H3K27me3.

Using ChIP-seq data pertaining to histone modifications H3K27me3, H3K36me3, and the enhancer EZH2 from the ENCODE Consortium, we find that current DPCs have difficulty in accurately detecting broad regions of differential enrichment between several common cell lines (Figure 3.6). In line with previous findings (Steinhauser et al., 2016), we observe that even current DPCs designed for broad data (Song and Smith, 2011; Allhoff et al., 2016) tend to detect either overly fragmented differential peaks or call regions exhibiting no difference in experimental signal between conditions as differential (Figure 3.6A). The low specificity and sensitivity of such methods may impair the biological interpretation of the resulting peak calls

in downstream analyses. Methods that rely on candidate peaks may also exhibit a compromised performance due to the limitations of single-sample peak callers in broad data (Stark and Brown, 2011; Chen et al., 2015). In addition, most current DPCs restrict their application to the analysis of two experimental conditions. For methods that are tailored for the analysis of three or more conditions, the classification of specific differential combinatorial patterns across conditions (or across various epigenomic processes) is still an open problem. The classification of such patterns would allow researchers to, for example, quantify treatment responses on the epigenomic level (Clouaire et al., 2014), or identify sets of processes working together to regulate local chromatin state. We find that the performance of such methods exhibit low sensitivity and specificity in calling differential regions in broad marks (Figure 3.6B).

We assessed the performance of our model on ChIP-seq experiments characterized by broad peaks (H3K36me3, H3K27me3, and EZH2) and short peaks (H3K27ac, H3K4me3, and the transcription factor CTCF). In simulations (Section 3.4) and in data sets from the ENCODE Consortium (Landt et al., 2012), we show that our model addresses the issues of the current peak callers in broad data (Section 3.5.1), while being flexible for short peaks (Section 3.5.2) and comparable to the fastest DPCs regarding the computation time. We show that our method can also be utilized for genomic regulatory state segmentation when studying multiple types of epigenomic processes from a single condition or cell line (Section 3.5.3). The Appendix B presents the data accession codes and the data pre-processing steps, respectively. Code implementing the method and to replicate the presented results are available in Appendix B.

3.3 Methods

3.3.1 Statistical Model

Let Y_{hij} denote the random variable pertaining to the ChIP read count for genomic window j from sample i of condition h, where $j = 1, \ldots, M$, $i = 1, \ldots, n_h$, $h = 1, \ldots, G$, and let y_{hij} be the observed count. Here, n_h is the number of samples in condition h and $N = \sum_{h=1}^{G} n_h$ is the total number of samples across the G conditions. At the j^{th} window, let $\mathbf{y}_{..j} = (y_{11j}, \ldots, y_{Gn_Gj})'$ denote the $N \times 1$ vector of ChIP window read counts across all samples and conditions, and let $\mathbf{y} = (\mathbf{y}'_{..1}, \ldots, \mathbf{y}'_{..M})'$ denote the corresponding $NM \times 1$ vector of window read counts spanning all windows, samples, and conditions. We assume that each window belongs to one of three possible hidden states: consensus background (state 1), differential (state 2), and consensus enrichment (state 3). Windows exhibiting low (high) enrichment across all conditions will be modeled by an emission distribution pertaining to the consensus background (enrichment) state. Windows

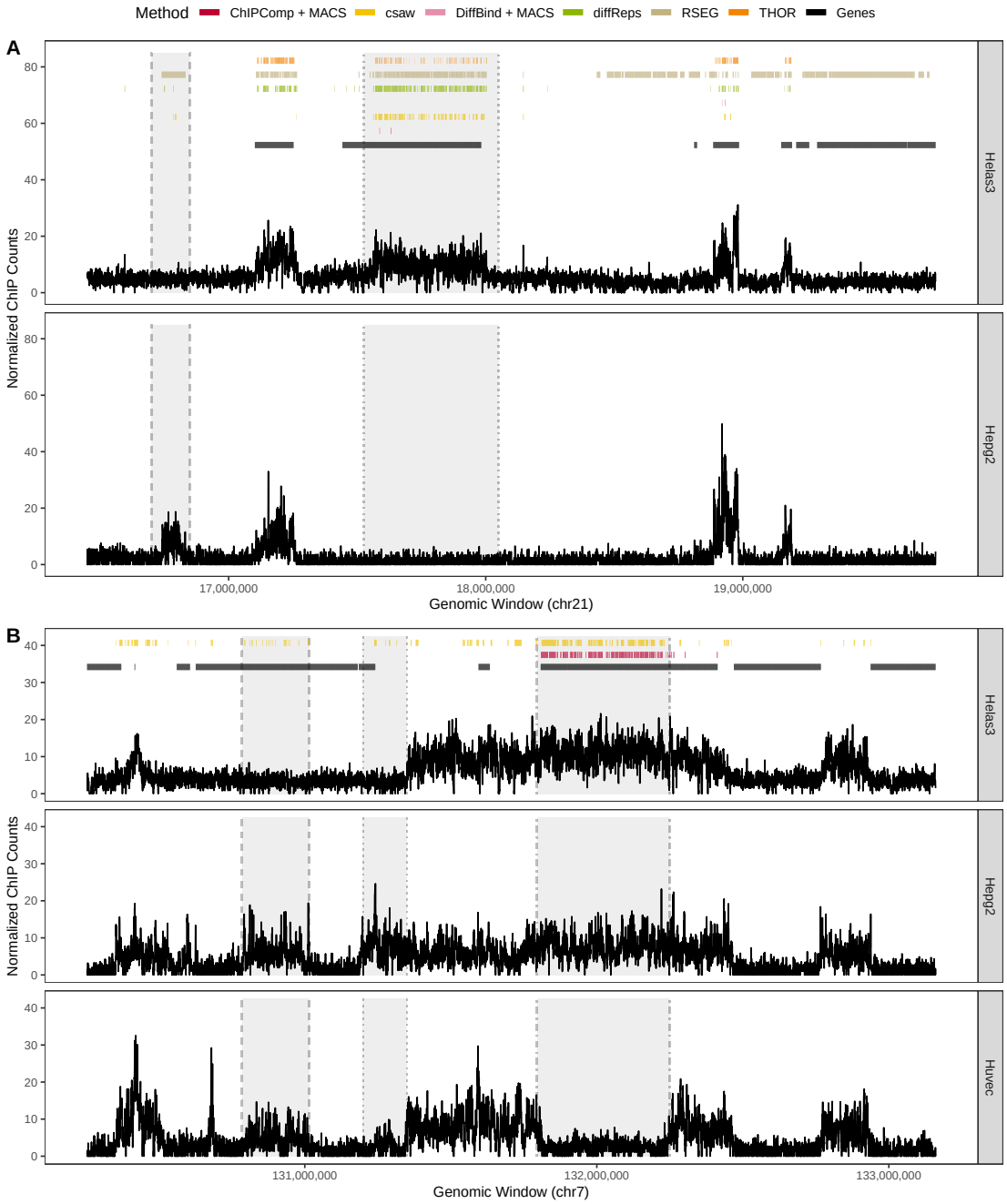

Figure 3.6: Performance of current DPC methods on calling differential enrichment regions in broad marks under a false discovery rate control of 0.05. (A): Differential peak calls between cell lines Helas3 and Hepg2 for the H3K36me3 histone modification. (B): Differential peak calls between cell lines Helas3, Hepg2, and Huvec for the H3K27me3 histone modification. Only ChIPComp, csaw, and DiffBind are designed for DPC under three or more conditions. Shaded regions indicate observed differential enrichment, and each vertical line type bordering each region represents a different combinatorial pattern of enrichment across cell lines. Optimal DPCs would call broad peaks inside shaded regions and no peaks outside them.

exhibiting enrichment under at least one condition, but not all conditions, will be modeled by an emission distribution pertaining to the differential state. If G conditions are of interest, there are $L = 2^G - 2$ possible differential combinatorial patterns of enrichment and background across conditions at a given window. The emission distribution pertaining to the differential state models all L possible differential combinatorial patterns via a mixture model with mixture proportions $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_L)'$, such that $\sum_{l=1}^{L} \delta_l = 1$ (see Figure B.1 in Baldoni et al. 2019a).

To model transitions between states, we assume a single latent discrete time stationary Markov chain $\mathbf{Z} = \{Z_j\}_{j=1}^{M}$, $Z_j \in \{1, 2, 3\}$, with state-to-state transition probabilities $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{33})'$ and initial probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)'$, such that $\sum_{s=1}^{3} \gamma_{rs} = 1$ and $\sum_{s=1}^{3} \pi_s = 1$ for $r \in \{1, 2, 3\}$. To facilitate the notation, let $f_r(\mathbf{y}_{..j}|\boldsymbol{\psi}_r)$ denote the emission distribution corresponding to the r^{th} hidden state, where $\boldsymbol{\Psi} = (\boldsymbol{\pi}', \boldsymbol{\gamma}', \boldsymbol{\delta}', \boldsymbol{\psi}')'$ denotes the vector of all model parameters, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1', \boldsymbol{\psi}_2', \boldsymbol{\psi}_3')'$ denotes each state's set of emission distribution-specific parameters, and $\mathcal{Z}$ denotes the set of 3^M possible state paths of length M. Then, the likelihood function pertaining to the proposed HMM may be written as

$$f(\mathbf{y}|\mathbf{x}; \boldsymbol{\Psi}) = \sum_{\mathbf{Z} \in \mathcal{Z}} \left\{ \prod_{r=1}^{3} \pi_r^{I(Z_1 = r)} \times \left(\prod_{j=2}^{M} \prod_{r=1}^{3} \prod_{s=1}^{3} \gamma_{rs}^{I(Z_{j-1} = r, Z_j = s)} \right) \times \right. \tag{3.6}$$
$$\left. \times \left(\prod_{j=1}^{M} f_1\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_1\right)^{I(Z_j=1)} f_2\left(\mathbf{y}_{..j}|\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\psi}_2\right)^{I(Z_j=2)} f_3\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_3\right)^{I(Z_j=3)} \right) \right\}.$$

Here, $\mathbf{x}$ is a fixed $G \times L$ design matrix enumerating each of the L possible differential combinatorial patterns in terms of the presence or absence of enrichment across each of the G conditions, only in the emission distribution of the differential state.

We assume that read counts pertaining to genomic windows from the consensus background ($r = 1$) and consensus enrichment ($r = 3$) states follow a Negative Binomial (NB) distribution with state-specific parameters $\boldsymbol{\psi}_r = (\mu_{(r,hij)}, \phi_r)'$, with mean $\mu_{(r,hij)}$ and variance $\mu_{(r,hij)}(1 + \mu_{(r,hij)}/\phi_r)$. Assuming independence of read counts across experiments and samples, conditional upon the HMM state, the emission distribution of the consensus background and consensus enrichment states, respectively, can be written as

$$f_r(\mathbf{y}_{..j}|\boldsymbol{\psi}_r) = \prod_{h=1}^{G}\prod_{i=1}^{n_h} \frac{\Gamma(y_{hij}+\phi_r)}{y_{hij}!\Gamma(\phi_r)} \left(\frac{\phi_r}{\mu_{(r,hij)}+\phi_r}\right)^{\phi_r} \left(\frac{\mu_{(r,hij)}}{\mu_{(r,hij)}+\phi_r}\right)^{y_{hij}}, \quad r \in \{1,3\}, \tag{3.7}$$

with $y_{hij} \in \{0,1,2,\ldots\}$, such that $\log(\mu_{(1,hij)}) = \beta_1 + u_{hij}$, $\log(\phi_1) = \lambda_1$, $\log(\mu_{(3,hij)}) = \beta_1 + \beta_3 + u_{hij}$, and $\log(\phi_3) = \lambda_1 + \lambda_3$. The offset u_{hij} adjusts for technical artifacts and allows the non-linear normalization of the signal profile across genomic windows, conditions, and samples (Appendix B). When $u_{hij} = 0$, β_1 and λ_1 represent the log-mean and log-dispersion, respectively, of read counts pertaining to consensus background state windows, whereas β_3 and λ_3 represent the difference in log-mean and log-dispersion of read counts from consensus enrichment state windows relative to consensus background state windows.

For windows belonging to the differential state ($r = 2$), we assume that the corresponding read counts are modeled by a L-component finite mixture model with mixture components that follow a Negative Binomial distribution, where each component corresponds to a particular differential combinatorial pattern. To define these patterns, let us consider the sets $S_1, \ldots, S_L$ that delineate the subset of the G conditions that are enriched in each of the L differential combinatorial patterns. For instance, if $G = 3$, the sets $S_1 = \{1\}$, $S_2 = \{2\}$, $S_3 = \{3\}$, $S_4 = \{1,2\}$, $S_5 = \{1,3\}$, and $S_6 = \{2,3\}$ define the six possible differential combinatorial patterns of enrichment and background across three conditions. That is, the set S_1 denotes enrichment in only the first condition and background in all others, whereas the set S_6 denotes enrichment in conditions 2 and 3 and background in condition 1. The presence or absence of enrichment in each of the L sets is encoded into each column of $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L)$, such that $\mathbf{x}_l = (x_{1l}, \ldots, x_{Gl})'$, and $x_{hl} = I(h \in S_l)$ for $l = 1, \ldots, L$ and $h = 1, \ldots, G$. That is, $\mathbf{x}_l$ is the $G \times 1$ vector of binary indicator variables denoting which subset of conditions are enriched in pattern (mixture component) l. A graphical illustration of our proposed model is provided in Figure B.1 in Baldoni et al. 2019a.

Let $\boldsymbol{\psi}_2$ denote the state-specific parameter vector pertaining to the differential state and let $\boldsymbol{\psi}_{(2,l)}$ denote the set of parameters pertaining to the l^{th} mixture component. Assuming independence of read counts across conditions and samples, conditional upon the differential HMM state, the finite mixture model emission distribution can be written as

$$f_2(\mathbf{y}_{..j}|\mathbf{x};\boldsymbol{\delta},\boldsymbol{\psi}_2) = \sum_{l=1}^{L} \delta_l \left[\prod_{h=1}^{G} \prod_{i=1}^{n_h} \frac{\Gamma\left(y_{hij} + \phi_{(2,l,h)}\right)}{y_{hij}!\Gamma\left(\phi_{(2,l,h)}\right)} \left(\frac{\phi_{(2,l,h)}}{\mu_{(2,l,hij)} + \phi_{(2,l,h)}} \right)^{\phi_{(2,l,h)}} \times \right.$$
$$\left. \times \left(\frac{\mu_{(2,l,hij)}}{\mu_{(2,l,hij)} + \phi_{(2,l,h)}} \right)^{y_{hij}} \right], \quad y_{hij} \in \{0, 1, 2, \ldots\}, \tag{3.8}$$

where $\mu_{(2,l,hij)}$ and $\phi_{(2,l,h)}$ are the mean and dispersion, respectively, pertaining to read counts originating from window j and sample i in condition h from the mixture component l. We assume that $\log(\mu_{(2,l,hij)}) = \beta_1 + \beta_3 x_{hl} + u_{hij}$ and $\log(\phi_{(2,l,h)}) = \lambda_1 + \lambda_3 x_{hl}$. That is, in the mixture component l, we utilize the same consensus background (consensus enriched) log-mean and log-dispersion from Equation 3.7 in all conditions that are specified by $\mathbf{x}_l$ to be background (enriched) in the l^{th} differential combinatorial pattern. There are several advantages to such a parametrization for the differential emission distribution. For example, it ensures that windows exhibiting differential enrichment across conditions share means and dispersions that are common between the consensus background and consensus enrichment states, a reasonable assumption that significantly increases computational efficiency. Utilizing a mixture model as the differential state emission distribution avoids the computational burden that would come from assuming separate hidden states for each of the L differential combinatorial patterns, particularly as G increases. We evaluate the strength of these assumptions through multiple simulations and a real data benchmarking analysis in Sections 3.4 and 3.5.

Two novel features result from our proposed approach that are relevant to the context of differential enrichment detection from ChIP-seq experiments. By using a modified version of the Expectation-Maximization (EM) algorithm to estimate the model parameters, we are able not only to detect differential enrichment regions across multiple conditions, but we can also classify various differential combinatorial patterns of enrichment within broad and short differential enrichment domains. With state-specific parameters, the current implementation of the method allows the direct modeling of continuous covariates (e.g. input controls; Appendix B), for which a state-level testing of their effects on the read count distribution could be performed. In a simulation study and in real data analyses, however, we did not observe a significant improvement in performance in differential peak detection after accounting for the effect of input controls (Appendix B), a fact that has also been observed by others (Lun and Smyth, 2015).

3.3.2 Estimation

To simplify the parameter estimation in Equation 3.9, we introduce another set of latent variables $\mathbf{W} = (\mathbf{W}_1', \ldots, \mathbf{W}_M')'$, such that $\mathbf{W}_j = (W_{j1}, \ldots, W_{jL})'$ for $j = 1, \ldots, M$. We assume that $\mathbf{W}$ is a sequence of independent random vectors such that $\mathbf{W}_j|(Z_j = 2) \sim$ Multinomial$(1, \boldsymbol{\delta})$ and $\mathbf{W}_j|(Z_j = r) = 0$ with probability 1 if $r = \{1, 3\}$. Under this setup, one may define the data generating mechanism when $Z_j = 2$ (differential state) and $W_{jl} = 1$ (l^{th} differential combinatorial pattern) such that read counts pertaining to genomic window j are sampled from $f_{(2,l)}$ given $\boldsymbol{\psi}_{(2,l)}$ and $\mathbf{x}_l$. Let $\mathcal{W}$ denote the set of L^M possible combinations of latent vectors $\mathbf{W}$. Hence, the likelihood function of the observed data (Equation 3.6) can be rewritten as

$$f(\mathbf{y}|\mathbf{x};\boldsymbol{\Psi}) = \sum_{\mathbf{Z}\in\mathcal{Z}}\sum_{\mathbf{W}\in\mathcal{W}}\left\{\left[\prod_{r=1}^{3}\pi_r^{I(Z_1=r)}\prod_{j=2}^{M}\prod_{r=1}^{3}\prod_{s=1}^{3}\gamma_{rs}^{I(Z_{j-1}=r,Z_j=s)}\right]\times\left[\prod_{j=1}^{M}\left(\prod_{l=1}^{L}\delta_l^{W_{jl}}\right)^{I(Z_j=2)}\right]\times\right.$$
$$\left.\times\left[\prod_{j=1}^{M}f_1\left(\mathbf{y}_{\cdot\cdot j}|\boldsymbol{\psi}_1\right)^{I(Z_j=1)}\left(\prod_{l=1}^{L}f_{(2,l)}(\mathbf{y}_{\cdot\cdot j}|\mathbf{x}_l;\boldsymbol{\psi}_{(2,l)})^{W_{jl}}\right)^{I(Z_j=2)}f_3\left(\mathbf{y}_{\cdot\cdot j}|\boldsymbol{\psi}_3\right)^{I(Z_j=3)}\right]\right\}, \tag{3.9}$$

where $f_{(2,l)}(\mathbf{y}_{\cdot\cdot j}|\mathbf{x}_l;\boldsymbol{\psi}_{(2,l)})$ is defined as in Equation 3.8. In the t^{th} step of the EM algorithm, the Q function of the complete data log-likelihood can be written as

$$\begin{aligned} Q\left(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(t)}\right) &= E_{\mathbf{Z}}\left(E_{\mathbf{W}|\mathbf{Z}}\left(\log\left(f(\mathbf{y},\mathbf{W},\mathbf{Z}|\mathbf{x};\boldsymbol{\Psi})\right)|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right), \\ &= Q_0(\boldsymbol{\pi},\boldsymbol{\gamma}|\boldsymbol{\Psi}^{(t)}) + Q_1(\boldsymbol{\psi}_1|\boldsymbol{\Psi}^{(t)}) + Q_2(\boldsymbol{\delta},\boldsymbol{\psi}_2|\boldsymbol{\Psi}^{(t)}) + Q_3(\boldsymbol{\psi}_3|\boldsymbol{\Psi}^{(t)}), \end{aligned} \tag{3.10}$$

where $Q_0(\boldsymbol{\pi},\boldsymbol{\gamma}|\boldsymbol{\Psi}^{(t)})$, $Q_1(\boldsymbol{\psi}_1|\boldsymbol{\Psi}^{(t)})$, $Q_2(\boldsymbol{\delta},\boldsymbol{\psi}_2|\boldsymbol{\Psi}^{(t)})$, and $Q_3(\boldsymbol{\psi}_3|\boldsymbol{\Psi}^{(t)})$ are defined in Appendix B. In the E-step of the EM algorithm, we compute the posterior probabilities from Equation 3.10. The quantities $Pr\left(Z_j = r|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)$ and $Pr\left(Z_{j-1} = r, Z_j = s|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)$, defined in Appendix B, can be calculated through the Forward-Backward algorithm (see Appendix B) and $Pr(W_{jl} = 1|Z_j = 2,\mathbf{y}_{\cdot\cdot j},\mathbf{x};\boldsymbol{\Psi}^{(t)}) = f_{(2,l)}(\mathbf{y}_{\cdot\cdot j}|\mathbf{x}_l;\boldsymbol{\psi}_{(2,l)}^{(t)})\delta_l^{(t)}/\sum_{k=1}^{L}f_{(2,k)}(\mathbf{y}_{\cdot\cdot j}|\mathbf{x}_k;\boldsymbol{\psi}_{(2,k)}^{(t)})\delta_k^{(t)}$ for $l = 1, \ldots, L$.

The Q function is maximized with respect to the parameters $\boldsymbol{\Psi} = (\boldsymbol{\pi}', \boldsymbol{\gamma}', \boldsymbol{\delta}', \beta_1, \beta_3, \lambda_1, \lambda_3)'$ during the M-step of the algorithm. Estimates for the initial and transition probabilities can be directly calculated as

$\hat{\pi}_r^{(t+1)} = Pr\left(Z_1 = r|\mathbf{y}, \mathbf{x}; \mathbf{\Psi}^{(t)}\right)$ and $\hat{\gamma}_{rs}^{(t+1)} = \sum_{j=2}^{M} Pr(Z_{j-1} = r, Z_j = s|\mathbf{y}, \mathbf{x}; \mathbf{\Psi}^{(t)}) / \sum_{j=2}^{M} Pr(Z_{j-1} = r|\mathbf{y}, \mathbf{x}; \mathbf{\Psi}^{(t)})$, respectively, restricted to $\sum_{r=1}^{3} \hat{\pi}_r^{(t+1)} = 1$ and $\sum_{s=1}^{3} \hat{\gamma}_{rs}^{(t+1)} = 1$, for $r \in \{1, 2, 3\}$. We perform conditional maximizations to compute estimates of the remaining model parameters $(\boldsymbol{\delta}', \beta_1, \beta_3, \lambda_1, \lambda_3)'$. First, mixture proportions can be estimated as $\hat{\delta}_l^{(t+1)} = \sum_{j=1}^{M} Pr(Z_j = 2|\mathbf{y}, \mathbf{x}; \mathbf{\Psi}^{(t)}) Pr(W_{jl} = 1|Z_j = 2, \mathbf{y}_{..j}, \mathbf{x}; \mathbf{\Psi}^{(t)}) / \sum_{j=1}^{M} Pr(Z_j = 2|\mathbf{y}, \mathbf{x}; \mathbf{\Psi}^{(t)})$. Estimating $(\beta_1, \beta_3, \lambda_1, \lambda_3)'$ from Equation 3.10 can be seen as obtaining parameter estimates from a series of weighted NB regression models with shared mean and dispersion parameters. We jointly estimate these quantities via the algorithm BFGS (Fletcher, 2013).

The estimation scheme is robust to situations where certain differential combinatorial patterns of enrichment are rare (Figure 3.13). This unique characteristic results from the fact that ChIP-seq experiments often provide enough data (usually $M > 10^7$ non-overlapping windows of 250 bp fixed size for the human reference genome) to estimate the parameters $(\beta_1, \beta_3, \lambda_1, \lambda_3)'$, which are shared across all L mixture components and HMM states. If pruning differential combinatorial patterns of the differential mixture component is of interest, the optimal number of mixture components L^*, $L^* < L$, can be selected via the Bayesian Information Criterion (BIC) for HMMs. We observed that selecting the optimal number of mixture components based on BIC agrees with the pruning of rare differential combinatorial patterns that we would not biologically expect to observe in real data (see Appendix B for a discussion).

To obtain the parameter estimates $\hat{\mathbf{\Psi}}$, the EM algorithm iterates until the maximum absolute relative change in the parameter estimates three iterations apart is less than 10^{-3} for three consecutive iterations. To reduce the computation time, we make use of a RCEM algorithm with threshold 0.05. Briefly, the RCEM algorithm substantially reduces the dimensionality of the data during the M-step by randomly assigning a zero posterior probability to genomic windows unlikely to belong to each of the HMM states. The current estimation set up allows genomic windows exhibiting equal distribution of read counts to have their posterior probability aggregated during the M-step of the algorithm. Often, the distribution of read counts along the genome is highly concentrated on a particular set of values, such as 0, 1, and 2 for instance. Genomic windows exhibiting a particular pattern of counts across samples and conditions can have their posterior probability aggregated during the M-step, which further reduces the dimensionality of the objective function during the numerical optimization and leads to a fast gradient-based optimization.

Once the algorithm reaches convergence, the final set of HMM posterior probabilities can be used to segment the genome into consensus background, differential, or consensus enrichment windows. Approaches

that control the total false discovery rate (FDR) via posterior probabilities (Efron et al., 2001) or that estimate the most likely sequence of hidden states (Viterbi, 1967) can be used for such purposes. Let $\hat{\rho}_{j2} = Pr\left(Z_j = 2|\mathbf{y}, \mathbf{x}; \hat{\mathbf{\Psi}}\right)$ denote the estimated posterior probability that the j^{th} genomic window belongs to the differential HMM state, $j = 1, \ldots, M$. For a cutoff of posterior probability α, the total FDR is $\sum_{j=1}^{M}(1 - \hat{\rho}_{j2})I(\hat{\rho}_{j2} \geq 1 - \alpha)/\sum_{j=1}^{M} I(\hat{\rho}_{j2} \geq 1 - \alpha)$, where $I(\cdot)$ is an indicator function. The posterior probability cutoff is then chosen by controlling the total FDR. Differential regions of enrichment are formed by merging adjacent windows that either meet a given FDR threshold level for the differential HMM state or belong to the same Viterbi's predicted state. Additional details of proposed EM algorithm and the implemented code are available in the Appendix B.

We evaluated the FDR approach using cutoffs 0.01, 0.05, 0.10, 0.15, and 0.20. We compared the results between the two approaches using window sizes of 250bp, 500bp, 750bp, and 1000bp. Overall, we observed that the Viterbi sequence of states led to similar results than the sequences based on FDR control cutoffs across all choices of window size. Specifically, we observed that the sensitivity and specificity of the sequence of Viterbi states were close to those from FDR control, in particular for FDR control 0.10. These results are shown in Figure 3.7 and in Baldoni et al. (2019a). These facts are also reflected by the length and number of called peaks. In Figure 3.8 and in Baldoni et al. (2019a) we show examples of peak calls from all FDR control cutoffs and the Viterbi sequence of states. Overall, we observed minor differences regarding the size of peak calls of the Viterbi and FDR control sequences across different choices of window sizes. These differences were mainly present in the data for H3K27me3, which is known to be a histone mark that expands through broader domains than H3K36me3. Finally, it is worth noting that the Viterbi algorithm gives us a way to call peaks that does not depend on the choice of the FDR cutoff.

3.4 Simulation Studies

We evaluate the presented model in two independent simulation studies of broad epigenomic marks. In the first study (Section 3.4.1), we simulated read count-based data to assess the precision of the parameter estimation scheme, the performance of differential peak detection, and the accuracy of the classification of specific differential combinatorial patterns of enrichment within differential peaks. In the second simulation study (Section 3.4.2), we utilize the simulation pipeline presented in (Lun and Smyth, 2015) to generate synthetic ChIP-seq reads from *in silico* experiments with broad differential peaks. The aim of the second

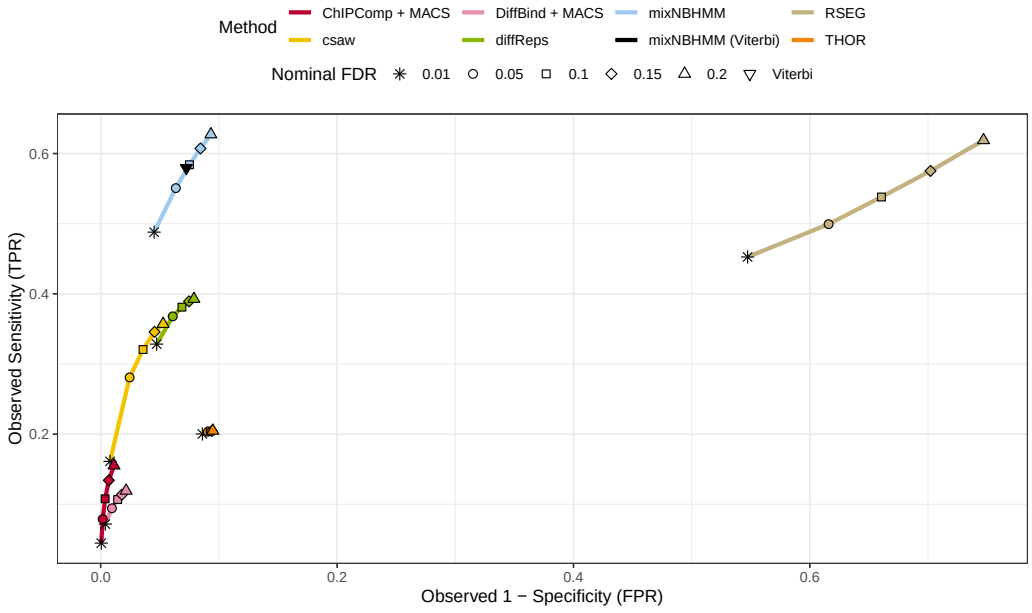

Figure 3.7: FDR-based results from broad marks (500bp) and Viterbi-based result from mixNBHMM.

simulation study was to compare our model with other DPCs in a more realistic scenario with broad peaks, while also avoiding the assumption of a parametric model for the data.

3.4.1 Read Count Simulation

Read counts were simulated under different scenarios that varied regarding the type of histone modification mark (H3K36me3 and H3K27me3), genome length (M, 10^5, 5×10^5, and 10^6 windows), number of conditions (G, 2, 3, and 4), and number of replicates per condition (n, 1, 2, and 4). We further assessed our model under different Signal-to-Noise Ratio (SNR) levels. We define the SNR as the ratio between the means of consensus enrichment and consensus background emission distributions. Mean and dispersion parameters used in this simulation study were estimated from ENCODE data and are presented in Table 3.5 and in Baldoni et al. (2019a) for all the remaining scenarios. Different SNR levels were defined by decreasing the ratio of the means in decrements of 10% while maintaining the mean-variance relationship. Read counts were assumed to follow a NB distribution and were simulated using a first-order Markov chain with 2^G states, representing every combination of background and enrichment across G conditions. We aimed to assess whether our model was able to assign all $2^G - 2$ simulated differential states to the differential HMM state, while maintaining a precise parameter estimation scheme and accurate classification of differential combinatorial patterns.

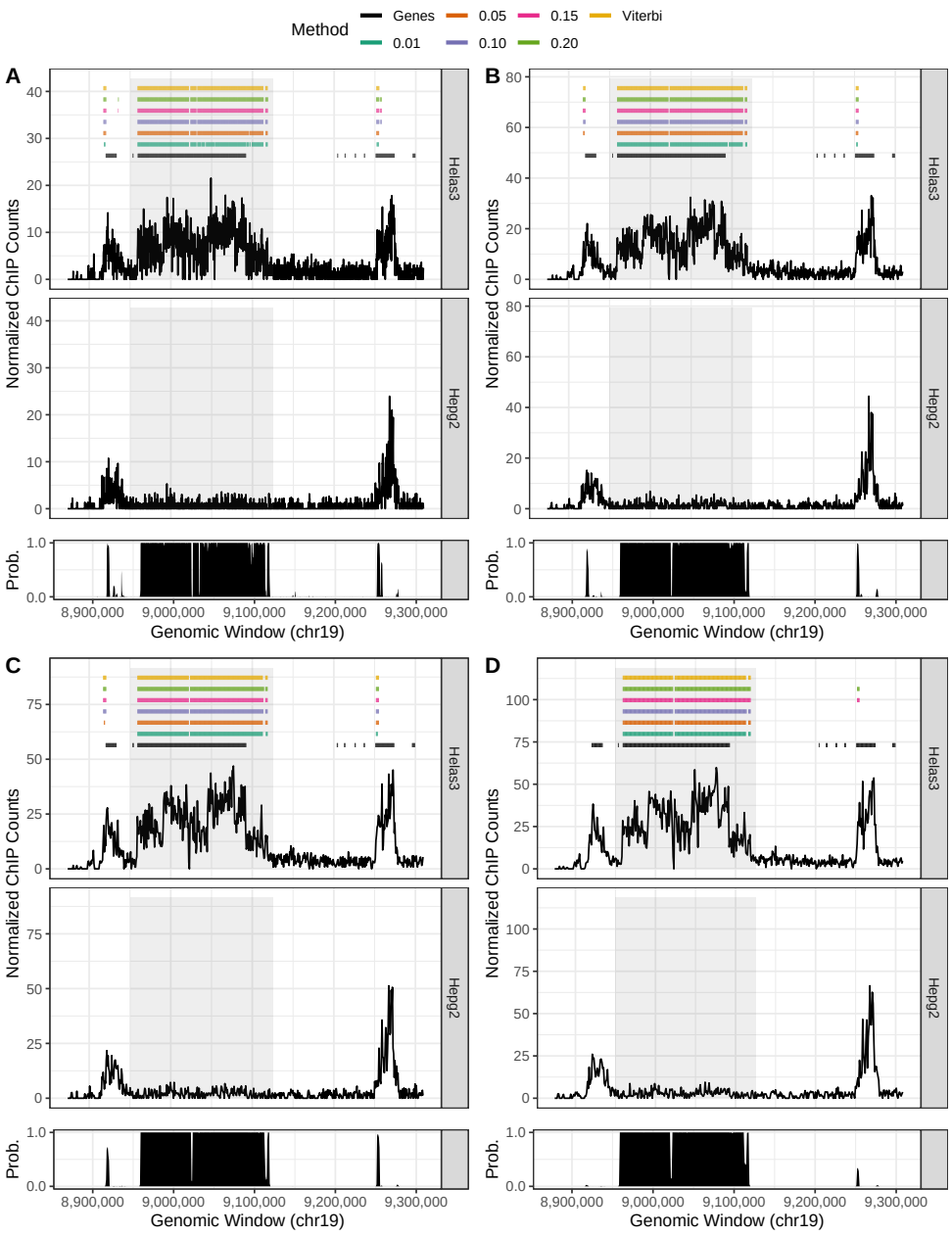

Figure 3.8: FDR- and Viterbi-based peak calls from H3K36me3 with 250bp (A), 500bp, (B), 750bp (C), and 1000bp (D).

3.4.1.1 Simulation Results

Table 3.5 shows the true values and the average relative bias of parameter estimates (and the 2.5^{th}, 97.5^{th} percentiles) from a hundred simulated data sets relative to the scenario of H3K27me3 with 10^5 genomic windows (see Baldoni et al. (2019a) for additional details). Results are shown for different levels of SNR, number of conditions, and number of replicates per condition. Overall, no significant differences regarding the relative bias of parameter estimates were observed across simulations under different genome lengths. Depending on the number of conditions, the observed relative bias and the range of the reported percentiles tended to decrease as more replicates were included in the analyses. This effect was particularly significant in scenarios with four conditions with respect to parameters β_3 and λ_3. In general, scenarios with higher SNR showed lower relative bias and variability of the parameter estimates in comparison to scenarios with lower SNR, regardless of the number of conditions or replicates per condition. In scenarios with lower SNR levels or higher number of conditions, these results also highlight the importance of experimental replicates to achieve precise parameter estimates. The proposed estimation approach via EM algorithm led to precise parameter estimates and was robust to a data generating mechanism that was different than the one assumed by the proposed model.

Next, we assessed the sensitivity of our method to detect simulated differential regions of enrichment. First, differential regions were defined from the HMM posterior probabilities pertaining to the differential state by controlling the total FDR as defined in Section 3.3.2. For different nominal FDR threshold levels, the model sensitivity was estimated as the proportion of windows correctly assigned as differential out of the total number of simulated differential windows. Additionally, the observed FDR was calculated as the proportion of genomic windows incorrectly called as differential out of the total number of called differential windows. Figure 3.9A shows the average observed true positive rate (y-axis) and the observed FDR (x-axis) for different nominal FDR levels across a hundred simulated data relative to the scenario of H3K27me3 with $M = 10^5$ genomic windows. Results are shown for different levels of SNR, number of conditions, and number of replicates per condition. Overall, we observed that the number of replicates per condition played a major role on the sensitivity levels of the model, in which scenarios with two and four replicates had the best results regardless of the number of conditions and SNR levels. For scenarios with either high number of conditions or low SNR levels, more replicates were needed to achieve higher sensitivity.

Table 3.5: Read count simulation. True values and average relative bias of parameter estimates (and 2.5^{th}, 97.5^{th} percentiles) across a hundred simulated data sets are shown for H3K27me3 with 10^5 genomic windows. Scenarios with observed SNR and 70% of observed SNR are shown.

				One Replicate		Two Replicates		Four Replicates	
SNR	Conditions	Parameter	True	R. Bias	($P_{2.5}$,$P_{97.5}$)	R. Bias	($P_{2.5}$,$P_{97.5}$)	R. Bias	($P_{2.5}$,$P_{97.5}$)
70% of	Two	β_1	1.116	0.000	(-0.004, 0.004)	0.000	(-0.002, 0.002)	0.000	(-0.002, 0.002)
Observed		β_3	0.808	-0.001	(-0.012, 0.008)	0.000	(-0.005, 0.006)	0.000	(-0.003, 0.004)
SNR		λ_1	1.281	0.000	(-0.016, 0.013)	0.000	(-0.010, 0.010)	0.000	(-0.008, 0.007)
		λ_3	-0.232	-0.001	(-0.109, 0.110)	0.000	(-0.078, 0.067)	0.000	(-0.058, 0.048)
	Three	β_1	1.116	0.002	(-0.005, 0.013)	-0.003	(-0.007, 0.001)	-0.001	(-0.002, 0.001)
		β_3	0.808	-0.156	(-0.233, -0.098)	-0.014	(-0.025, -0.005)	-0.001	(-0.004, 0.003)
		λ_1	1.281	0.005	(-0.033, 0.032)	0.004	(-0.006, 0.014)	0.001	(-0.006, 0.007)
		λ_3	-0.232	0.898	(0.679, 1.083)	0.130	(0.012, 0.242)	0.015	(-0.033, 0.061)
	Four	β_1	1.116	0.001	(-0.005, 0.020)	-0.012	(-0.017, -0.008)	-0.010	(-0.014, -0.007)
		β_3	0.808	-0.145	(-0.256, -0.120)	-0.087	(-0.098, -0.075)	-0.032	(-0.062, -0.015)
		λ_1	1.281	0.016	(-0.028, 0.033)	0.028	(0.016, 0.041)	0.017	(0.007, 0.028)
		λ_3	-0.232	0.947	(0.819, 1.083)	0.769	(0.672, 0.877)	0.366	(0.211, 0.613)
Observed	Two	β_1	1.116	0.000	(-0.004, 0.004)	0.000	(-0.003, 0.002)	0.000	(-0.002, 0.002)
SNR		β_3	1.165	0.000	(-0.005, 0.005)	0.000	(-0.003, 0.003)	0.000	(-0.002, 0.003)
		λ_1	1.281	0.000	(-0.015, 0.018)	-0.001	(-0.010, 0.008)	0.001	(-0.004, 0.007)
		λ_3	0.124	-0.012	(-0.231, 0.174)	0.008	(-0.108, 0.129)	-0.008	(-0.107, 0.085)
	Three	β_1	1.116	-0.004	(-0.008, 0.001)	-0.001	(-0.003, 0.002)	0.000	(-0.002, 0.002)
		β_3	1.165	-0.010	(-0.019, -0.002)	0.000	(-0.003, 0.003)	0.000	(-0.002, 0.002)
		λ_1	1.281	-0.001	(-0.012, 0.012)	0.000	(-0.008, 0.009)	0.000	(-0.006, 0.007)
		λ_3	0.124	-0.422	(-0.697, -0.068)	-0.026	(-0.171, 0.083)	-0.003	(-0.085, 0.081)
	Four	β_1	1.116	-0.013	(-0.019, -0.008)	-0.008	(-0.014, -0.003)	0.000	(-0.002, 0.001)
		β_3	1.165	-0.111	(-0.121, -0.101)	-0.005	(-0.010, -0.001)	0.000	(-0.002, 0.002)
		λ_1	1.281	-0.011	(-0.025, 0.001)	0.009	(-0.002, 0.018)	0.001	(-0.004, 0.006)
		λ_3	0.124	-3.526	(-3.782, -3.239)	-0.438	(-0.740, -0.217)	-0.009	(-0.084, 0.059)

Finally, we used the estimated mixture model posterior probabilities $Pr(W_{jl} = 1|Z_j = 2, \mathbf{y}_{..j}, \mathbf{x}; \hat{\mathbf{\Psi}})$, $j = 1, \ldots, M$, to classify the differential combinatorial patterns of enrichment of detected differential windows. To this end, we first calculated the maximum estimated mixture model posterior probability across all L components to determine the most likely differential combinatorial pattern from genomic windows assigned to be part of the differential state. Then, we compared the window-based classification with the true window-based simulated states from the Markov Chain (states $2, \ldots, G-1$). Figure 3.9B shows the confusion matrices of classified (x-axis) and simulated (y-axis) differential windows for a scenario with three conditions and data simulated from H3K27me3 with 10^5 genomic windows. Differential combinatorial patterns of enrichment are represented by the sequences of letters 'E' (enrichment) and 'B' (background), such that each letter corresponds to the status of a given condition. The number of windows (averaged over a all simulated data sets) is shown as entries of the matrices and represented by the color scale. Darker colors on the diagonal entries indicate better agreement between simulated and classified patterns. By utilizing the posterior probabilities from the mixture model, we observed a good performance when classifying the differential combinatorial pattern of enrichment from differential windows. Results were best under scenarios with higher number of replicates or SNR.

Overall, simulated scenarios with higher number of replicates or a higher SNR led to less biased and more precise parameter estimates, higher accuracy of differential peak detection, and best classification of the differential direction of enrichment. To the best of our knowledge, the classification capability of the proposed model in settings with more than two conditions is a novelty not yet available in any other method for the detection of differential protein-DNA binding sites. Although (Lun and Smyth, 2015) presented a DPC tailored for multiple conditions, its current implementation does not allow the classification of differential combinatorial patterns of enrichment under three or more conditions.

3.4.2 Sequencing Read Simulation

We performed a second simulation study aiming to compare the proposed model with the current DPCs ChIPComp, csaw, DiffBind, diffReps, RSEG, and THOR. We used the simulation pipeline presented by (Lun and Smyth, 2015) where data were generated in a more general scheme without a particular read count model assumption. Here, sequencing reads from broad ChIP-seq experiments were generated for two conditions and two replicates per condition. For the differential peaks callers ChIPComp and DiffBind that require sets of candidate regions, we followed the analyses presented by (Lun and Smyth, 2015) and called peaks in advance

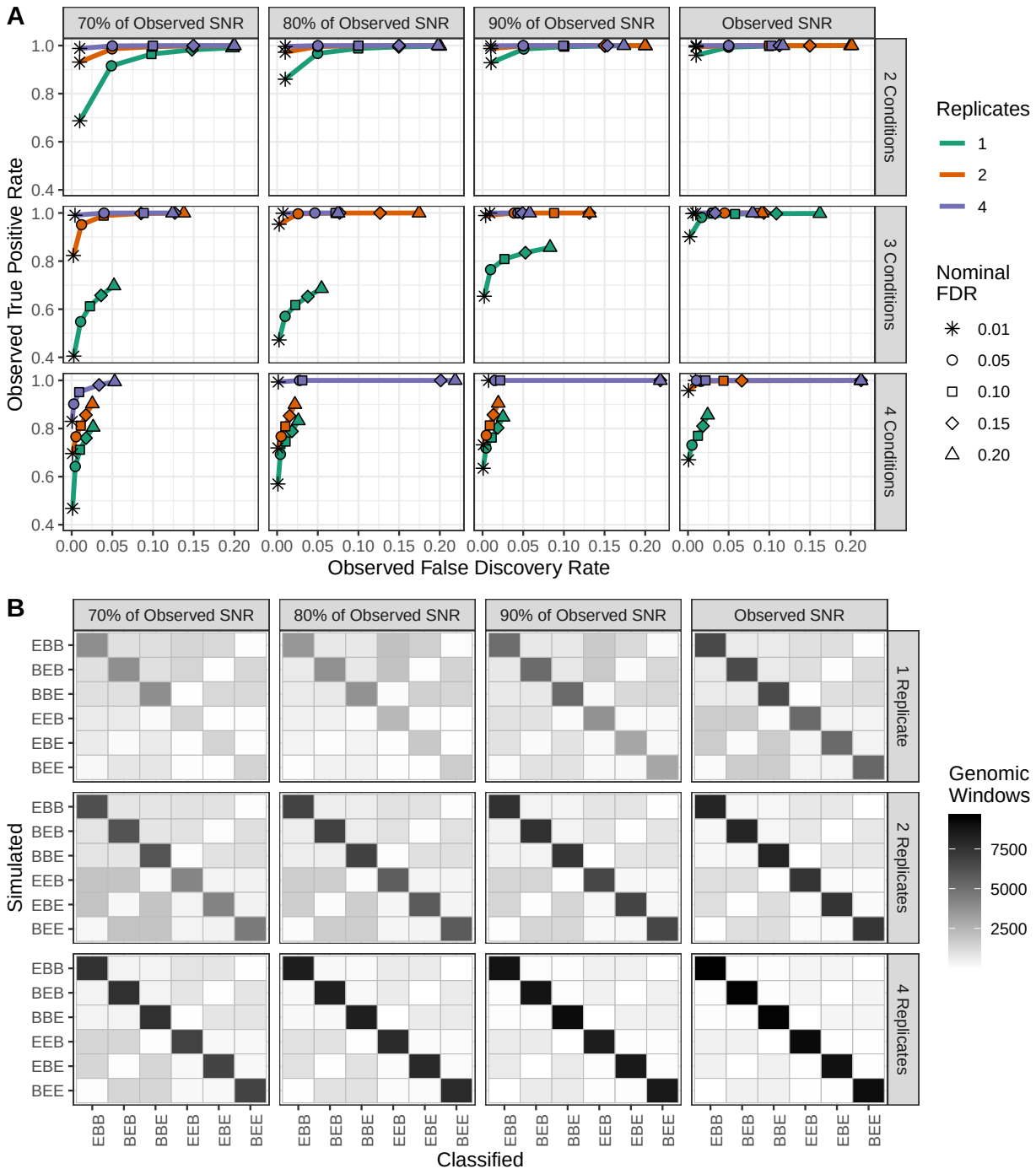

Figure 3.9: Read count simulation. (A): average observed TPR and FDR for different nominal FDR levels for simulated scenarios of H3K27me3 with 10^5 windows. (B): confusion matrices for the 3 conditions scenario. On x- and y-axes, the labels indicate the classified and simulated patterns, respectively (e.g., BEB denotes enrichment in conditions 2 only). Darker colors on the diagonal indicate better agreement.

using HOMER. Peaks were then used as input in the respective software for differential call. A hundred simulated data sets were generated and peaks were called by all the methods under multiple nominal FDR thresholds. For our method and RSEG, window-based posterior probabilities were used to control the total FDR as described in Section 3.3.2.

3.4.2.1 Simulation Results

Figure 3.10 shows the main results of our second simulation study. Out of 100 simulated data sets, RSEG either failed to analyze the data due to internal errors or called the entire genome as differential in 26 and in 3 instances, respectively. Similar issues have been previously reported in other studies (Starmer and Magnuson, 2016). We observed that our method showed the highest observed sensitivity among all DPCs, regardless of the nominal FDR thresholding level, while maintaining a moderate observed FDR (Figure 3.10A). Methods such as diffReps, RSEG, and THOR showed higher observed FDR levels than the nominal threshold due to the excessive number of differential peaks called outside true differential regions (shaded area in Figure 3.10D). While diffReps and THOR called an excessive number of short and discontiguous peaks, RSEG called regions that were usually wider than the observed differential enrichment regions. These results are further illustrated in Figure 3.10B, where we present the average ratio of the number of called and simulated peaks (y-axis) and the average number of called peaks intersecting true differential regions (x-axis). Regarding the computation time, the HMM-based algorithms RSEG and THOR appeared to be the most computationally intensive and required longer amounts of time to analyze the data. In Figure 3.10C, we present the box plots of computing time (in minutes) across a hundred simulated data sets for all benchmarked methods. While still being an HMM-based algorithm, our method was among the fastest tools for differential peak detection due to the implemented strategies to improve the computation time of the EM algorithm (Section 3.3.2). Figure 3.10D shows an example of a genomic region with simulated data and called peaks from various methods using nominal FDR threshold 0.05. As shown, our method was able to consistently cover most of true differential regions with broad peaks while exhibiting a limited number of false discoveries (see Baldoni et al. (2019a) for additional examples and results).

3.5 Application to ENCODE Data

We applied our method to ChIP-seq data from the ENCODE Consortium (Section 3.2) to detect differential regions of enrichment of several epigenomic marks across distinct cell lines. First, we analyzed broad data from the histone modifications H3K36me3 and H3K27me3 as well as data from the enhancer EZH2

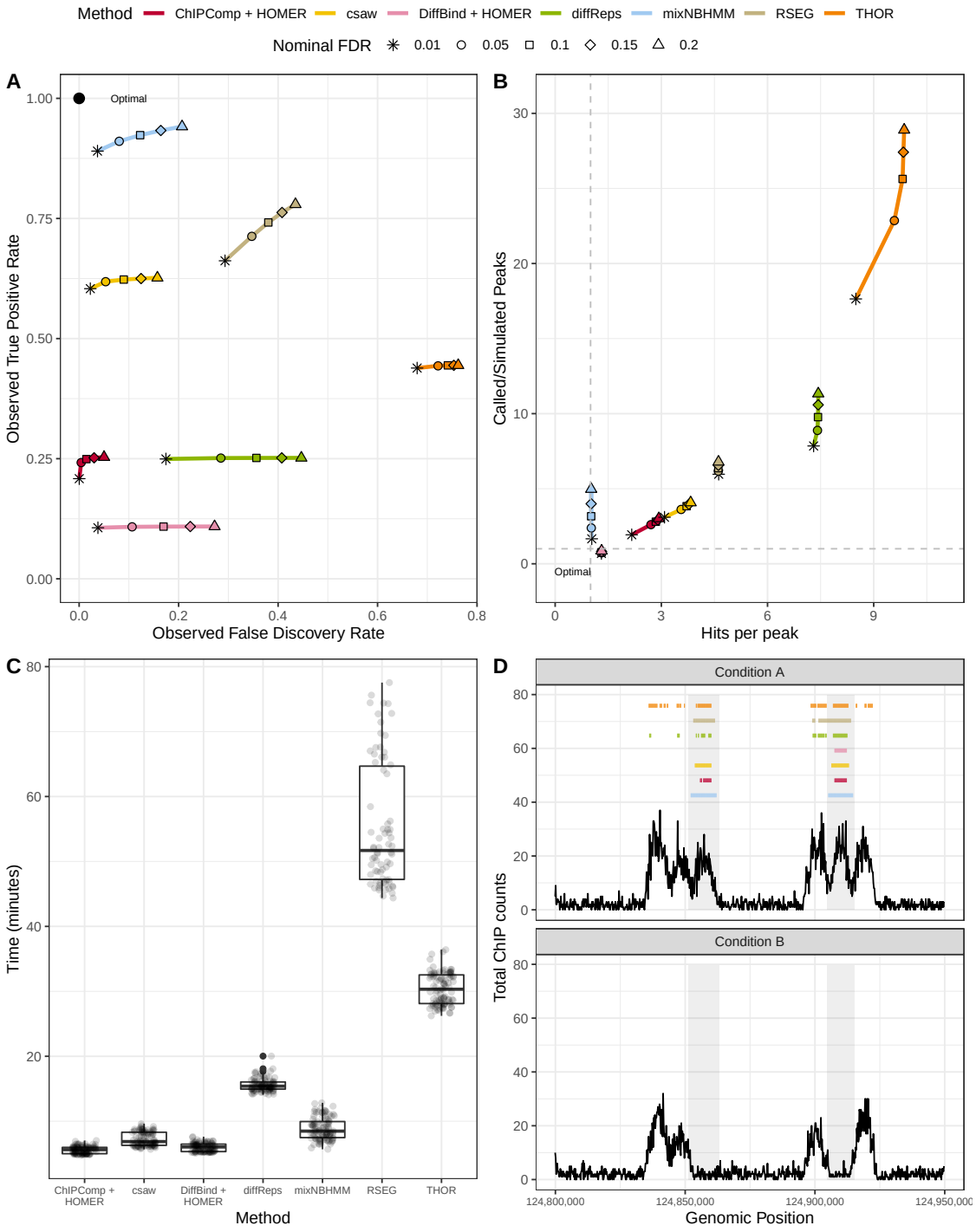

Figure 3.10: Sequencing read-based simulation from the csaw pipeline. (A): average observed sensitivity and FDR for various methods. (B): scatter plot of average ratio of called and simulated peaks (y-axis) and number of called peaks intersecting true differential regions (x-axis). (C): box plot of computing time (in minutes) for various algorithms. (D): an example of differential peak calls under a nominal FDR control of 0.05. Shaded areas indicate true differential peaks.

(Section 3.5.1). Secondly, we assessed the performance of the presented model on ChIP-seq experiments from the transcription factor CTCF and the histone modifications H3K27ac and H3K4me3 characterized by short peaks (Section 3.5.2). For H3K27ac and H3K4me3, enrichment peaks are usually deposited on the promoter regions of actively transcribed genes and several studies have associated their role with gene transcription (Creyghton et al., 2010; Lauberth et al., 2013). The transcription factor CTCF is a protein that binds to short DNA motifs and is responsible for several cellular processes that include the regulation of the chromatin 3D structure and mRNA splicing (Shukla et al., 2011). Two technical replicates for each epigenomic mark were used in the analysis.

Using RNA-seq experimental data from the ENCODE Consortium, we assessed the practical significance of our results by associating the detection and classification of differential combinatorial patterns from called peaks of H3K36me3, H3K27me3, and EZH2 with the direction of gene expression (Section 3.5.3). The quantification of gene expression for the analyzed data proceeded as follows. First, we used Salmon (Patro et al., 2017) to quantify transcript expression from cell-specific RNA-seq experiments. We then calculated, using the R package $tximport$ (Soneson et al., 2015), estimated counts using abundance estimates (transcripts per million, TPM) scaled up to the average transcript length over samples and library size. This step ensures that counts computed from Salmon are not correlated with the average transcript length.

For the three cell line analysis presented in Baldoni et al. (2019a), we calculated the number of ChIP-seq reads from H3K4me3 and H3K27ac overlapping gene promoters. Promoter regions extend around the transcription start site 2000 base pairs upstream and 200 base pair downstream. Read counts from RNA-seq and ChIP-seq were normalized for sequencing depth using DESeq2. For the two cell line scenario (Section 3.5.1), differentially transcribed genes were defined through log2 fold changes of H3K36me3 ChIP-seq read counts. We observed several cases in which differentially expressed genes (defined by RNA-seq data) did not exhibit differential enrichment for H3K36me3. However, due to the activating roles of H3K36me3 on gene transcription, we follow the ideas presented by (Steinhauser et al., 2016) and (Ji et al., 2013) and defined differentially transcribed genes based on log2 fold changes of H3K36me3 ChIP-seq read counts.

Sensitivity and specificity metrics were calculated on the window-level as follows. For non-overlapping genomic windows $b_1, \ldots, b_M$, let $g_j = I(b_j \in$ differentially transcribed gene) and $d_j = I(b_j \in$ differential peak) denote the indicators of genomic windows being associated with either differential gene bodies or differential peaks, respectively, for $j = 1, \ldots, M$, for a given method and nominal FDR level. Then, the observed sensitivity (TPR), specificity (1-FPR), and FDR were calculated as follows:

$$\text{Sensitivity} = \frac{\sum_{j=1}^{M} g_j d_j}{\sum_{j=1}^{M} g_j}$$

$$\text{Specificity} = \frac{\sum_{j=1}^{M} (1-g_j)(1-d_j)}{\sum_{j=1}^{M} (1-g_j)}$$

$$\text{FDR} = \frac{\sum_{j=1}^{M} (1-g_j) d_j}{\sum_{j=1}^{M} d_j}$$

We compared the genome-wide performance of the presented model with the current DPCs ChIPComp, csaw, DiffBind, diffReps, RSEG, and THOR. For the methods that require a set of candidate regions to be specified *a priori*, ChIPComp and DiffBind, peaks were called in advance using MACS (Zhang et al., 2008) and used as input to the software for differential call. We benchmarked methods regarding the coverage of differentially transcribed gene bodies, the number and average size of differential peak calls, $\log_2$ fold change (LFC) of read counts, Spearman correlation of $\log_2$-transformed read counts between cell lines, and computation time. Metrics for sensitivity and specificity were defined on the window level and based on the coverage of differentially transcribed gene bodies by called peaks. For broad marks, read counts were computed using non-overlapping windows of 500bp. For the remaining short marks, we computed read counts using non-overlapping windows of 250bp. Results presented in this section pertain to the analysis of two cell lines, namely Helas3 and Hepg2. A discussion about the choice of the window size is presented in Section 3.3.2. Results from the analysis of more than two cell lines are presented in Baldoni et al. 2019a. Data accessing code, data pre-processing steps, method-specific parameters, and code to replicate the presented results are detailed in Appendix B.

3.5.1 Analysis of ChIP-seq Data From Broad Marks

Methods were benchmarked regarding the coverage of differentially transcribed gene bodies. The histone modification H3K36me3 is known to be associated with gene transcription and enriched regions of this mark are usually deposited on transcribed gene bodies. Hence, the location of differential peaks of H3K36me3 is expected to agree with the location of differentially expressed genes. Following the analysis presented by (Steinhauser et al., 2016), we defined a set of protein coding genes exhibiting $|\text{LFC}| > 2$ of ChIP-seq read counts between the two analyzed cell lines as true differentially transcribed genes. Results using different threshold levels are presented in Baldoni et al. 2019a and agree with those presented here. Protein-coding

genes with total read count across cell lines under the 25^{th} percentile were excluded from the analysis. Normalization by the median log-ratios of each replicate over the geometric mean was performed to avoid spurious differences due to sequencing depth.

In Figure 3.11A, we show receiver operating characteristic (ROC) curves for various methods and different nominal levels of FDR threshold for differential peaks of H3K36me3. Similar to the analysis of broad histone modification marks presented by Xing et al. (2012), we computed the observed true positive rates (false positive rates) on the window-level as the proportion of windows called as differential out of the total number of windows associated (not associated) with differentially transcribed genes. Our method had the best overall performance among all DPCs as its differential peaks were able to cover most of differentially transcribed gene bodies while still maintaining a low number of false positives. Methods that tended to call short peaks, such as ChIPComp and DiffBind, were the ones with the lowest sensitivity among all methods. ChIPComp and DiffBind have been previously shown to be dependent on the set of candidate peaks and to perform best in scenarios with short peaks (Figure 3.11B; Steinhauser et al. (2016); Lun and Smyth (2015)). In Figure 3.11C, we show the observed sensitivity (y-axis) and the average differential peak size (kbp; x-axis) for various methods under different nominal FDR levels (the observed FDR is annotated next to each data point). Our model and RSEG, two HMM-based methods, tended to call broader differential peaks and exhibited better sensitivity than other methods. Yet, differential peaks called by RSEG often did not correspond to differential regions of enrichment (Figure 3.11D), a behavior that has been noted by others (Starmer and Magnuson, 2016) and also seen in simulated data (Figure 3.10). Our HMM-based method with a non-linear normalization scheme via model offsets allowed us to maintain a low observed FDR and a higher sensitivity than other DPCs. In Figure 3.11F, we show examples of differential peak calls for the enhancer EZH2. Our method was among the fastest algorithms due to our computational scheme, taking approximately 1 hour to analyze genome-wide data (Figure 3.11E).

3.5.2 Analysis of ChIP-seq Data From Short Marks

We further evaluated the performance of the proposed method on data sets characterized by short peaks, namely the histone modifications H3K4me3 and H3K27ac and the transcription factor CTCF. The goal of our analysis was to assess whether our method was robust to different types of data and still able to call short differential regions of enrichment. In these scenarios, differential peaks are usually observed in isolated genomic regions and exhibit a high SNR. It has been shown that certain HMM-based approaches, including

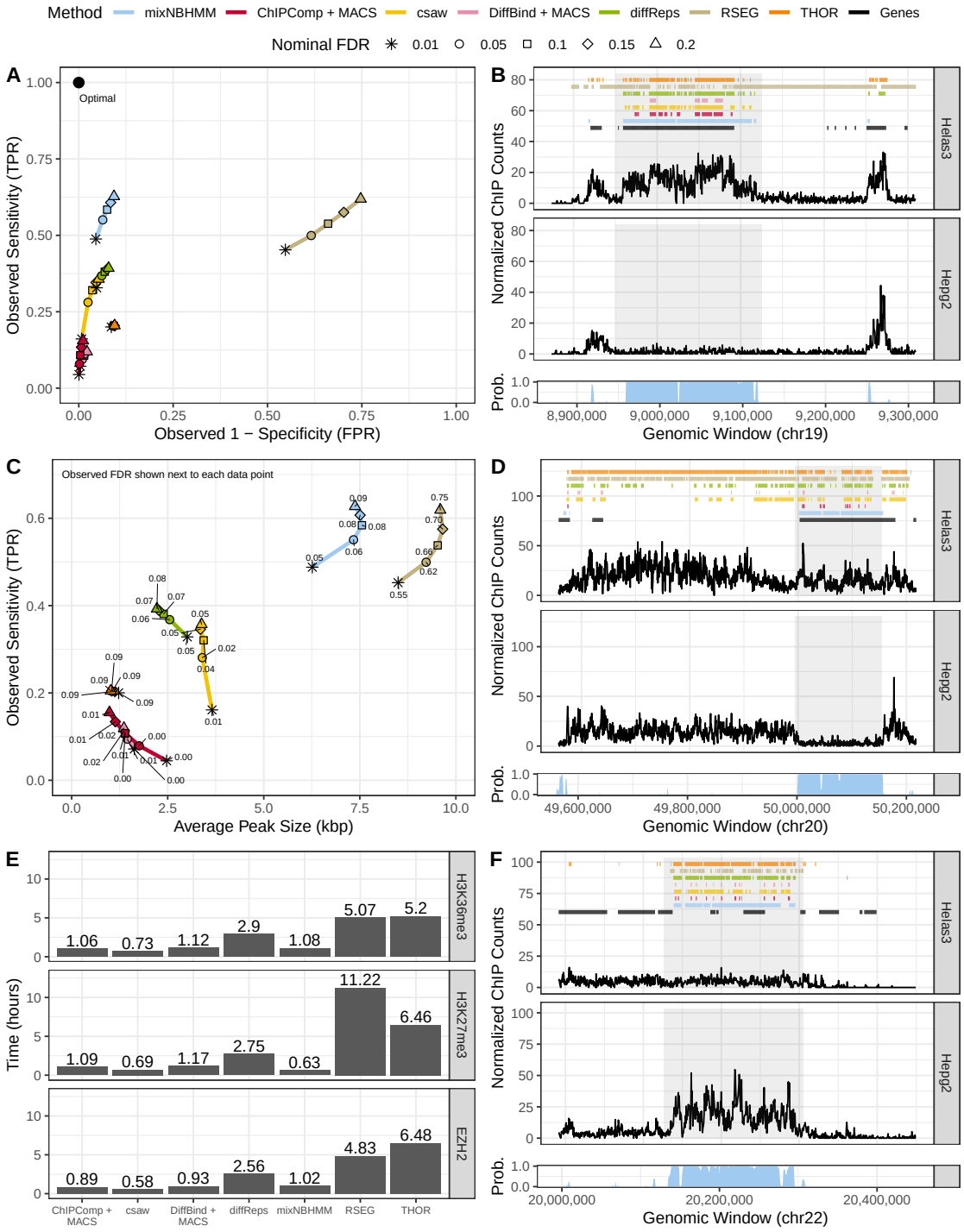

Figure 3.11: Analysis of broad ENCODE data. (A): ROC curves of H3K36me3 differential peak calls. (C): average number (y-axis) and size (x-axis) of H3K27me3 called peaks for various methods and different nominal FDR thresholds. (B), (D), and (F): example of peak calls from H3K36me3, H3K27me3, and EZH2, respectively, under a nominal FDR control of 0.05. Posterior probabilities of the HMM differential state are shown at the bottom of each panel. (E): computing time of genome-wide analysis from various methods.

RSEG, have low accuracy under short histone modification marks and TF data (Hocking et al., 2016). As we show, the model proposed in this article performs comparably to the evaluated DPCs known to perform best in short data (ChIPComp, DiffBind; (Steinhauser et al., 2016)) and appeared to be more efficient regarding the computation time in certain scenarios.

We calculated the LFC and the Spearman correlation between cell lines Helas3 and Hepg2 based on ChIP-seq read counts mapped onto differential peaks called by each method. Read counts were previously normalized by the median log-ratios of each replicate over the geometric mean to avoid spurious differences due to sequencing depth. As these marks are characterized by short peaks, ideal methods would show high absolute LFC and negative correlation between read counts mapped on differential peaks. Figure 3.12 shows the main results from our analysis using short data sets. In Figures 3.12 A and C we show the median LFC and the Spearman correlation of ChIP-seq counts for differential CTCF and H3K4me3 peak calls (sorted by the absolute LFC), respectively, under a nominal FDR control of 0.05. We present separate curves regarding the signal of observed enrichment to better characterize the direction of change. The results show that the HMM-based methods RSEG and THOR were among those with the lowest absolute LFC among all methods, which confirms their sub optimal performance in the scenario of short peaks (Hocking et al., 2016). In addition, we observed that ChIPComp had the best performance overall as it was able to call differential peaks with the highest absolute LFC and the lowest correlation between read counts of the two analyzed cell lines. Our model was able to properly call truly short differential peaks (Figures 3.12 B, D, and F) and was comparable to the non-HMM based methods regarding the computation time (Figure 3.12E), jointly calling differential peaks in less than 1.5 hour.

3.5.3 Genomic Segmentation and Classification of Chromatin States

Lastly, we analyzed data from the cell line Helas3 to segment its genome regarding the activity of marks H3K36me3, H3K27me3, and EZH2. We considered each mark as a separate experimental condition ($G = 3$) and sought to jointly classify local chromatin states in Helas3 based upon the presence or absence of enrichment from each mark. It is known that EZH2 catalyzes the methylation of H3K27me3, a repressive mark, and H3K36me3 is associated with transcribed genes (Section 3.2). Hence, we expected regions of enrichment in consensus for these marks to be rare and differential regions to be mostly represented by either transcribed chromatin states (enrichment for H3K36me3 alone) or repressed chromatin states (enrichment co-occurrence for H3K27me3 and EZH2). The analyses presented in this section highlight the applicability

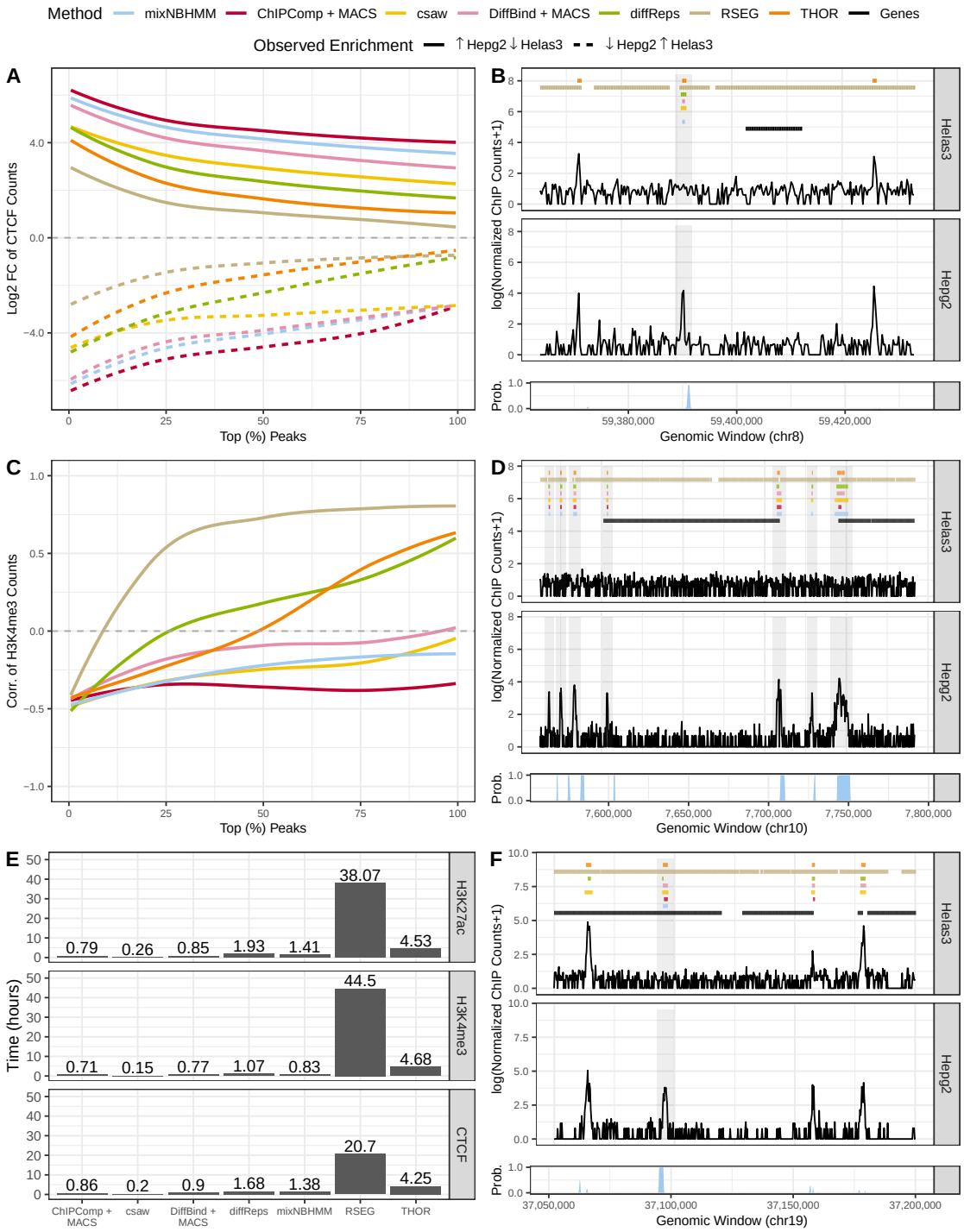

Figure 3.12: Analysis of short ENCODE data. (A) and (C): median LFC and correlation between cell lines of ChIP-seq counts from differential peaks for CTCF and H3K4me3, respectively. (B), (D), and (F): example of peak calls from CTCF, H3K4me3, and H3K27ac, respectively. Posterior probabilities of the HMM differential state are shown at the bottom of each panel. (E): computing time of genome-wide analysis from various methods. Results are shown under a nominal FDR control of 0.05.

of our method in the context of genomic segmentation (Ernst and Kellis, 2012), a distinct problem not tackled by current DPCs.

First, we segmented the genome using the Viterbi sequence of most likely HMM states to understand the distribution of genomic regions associated with consensus background, differential, and consensus enrichment states (Figure 3.13). While the majority of the genomic regions exhibited no enrichment for any of the analyzed marks, regions of consensus enrichment were rare and represented only 2% of the analyzed genome (Figure 3.13A), as expected. Consensus background and differential regions mostly covered intergenic (66%) and protein-coding genic regions (61%), respectively. Differential genomic windows were mostly representing either transcribed chromatin states or repressed chromatin states (Figure 3.13B). All differential combinatorial patterns expected to be rare had associated mixture proportion estimates less than 0.02 (see Appendix B for a discussion on pruning rare states).

These results suggest that protein-coding genes overlapping differential regions should be either silenced (e.g. genes associated with repressed chromatin states) or highly expressed (e.g. genes associated with transcribed chromatin states). To assign combinatorial patterns to differential peaks, we chose the combination pertaining to the most frequent mixture component across windows by using the maximum estimated mixture model posterior probability, $Pr(W_{jl} = 1|Z_j = 2, \mathbf{y}_{..j}, \mathbf{x}; \mathbf{\Psi}^{(t)})$, $j = 1, \ldots, M$. For genes overlapping differential regions associated with either transcribed or repressed chromatin states, we computed the distribution of transcripts per million (TPM) from matching RNA-seq experiments. Genes associated with transcribed chromatin states had a significantly higher distribution of TPM counts than genes associated with repressed chromatin states (Figure 3.13C). We detected broad differential regions of enrichment and the classification of differential combinatorial patterns agreed with their biological roles as well as the expression levels of associated gene bodies. Figure 3.13D shows an example of a genomic region with differential enrichment for the analyzed marks. We compare our results to ChromHMM with 3 states, a method developed for chromatin segmentation. Our method offers the benefit of simultaneously detecting differential peaks and classifying the combinatorial pattern of enrichment through mixture model posterior probabilities even in the context of genomic segmentation with highly diverse epigenomic marks (Figure 3.14). By using the BIC for model selection (Appendix B), one can choose the number of biologically relevant mixture components to be included in the model, a task that may not be as straightforward in methods such as ChromHMM (see Supplementary Figure 4 in Ernst and Kellis (2012)).

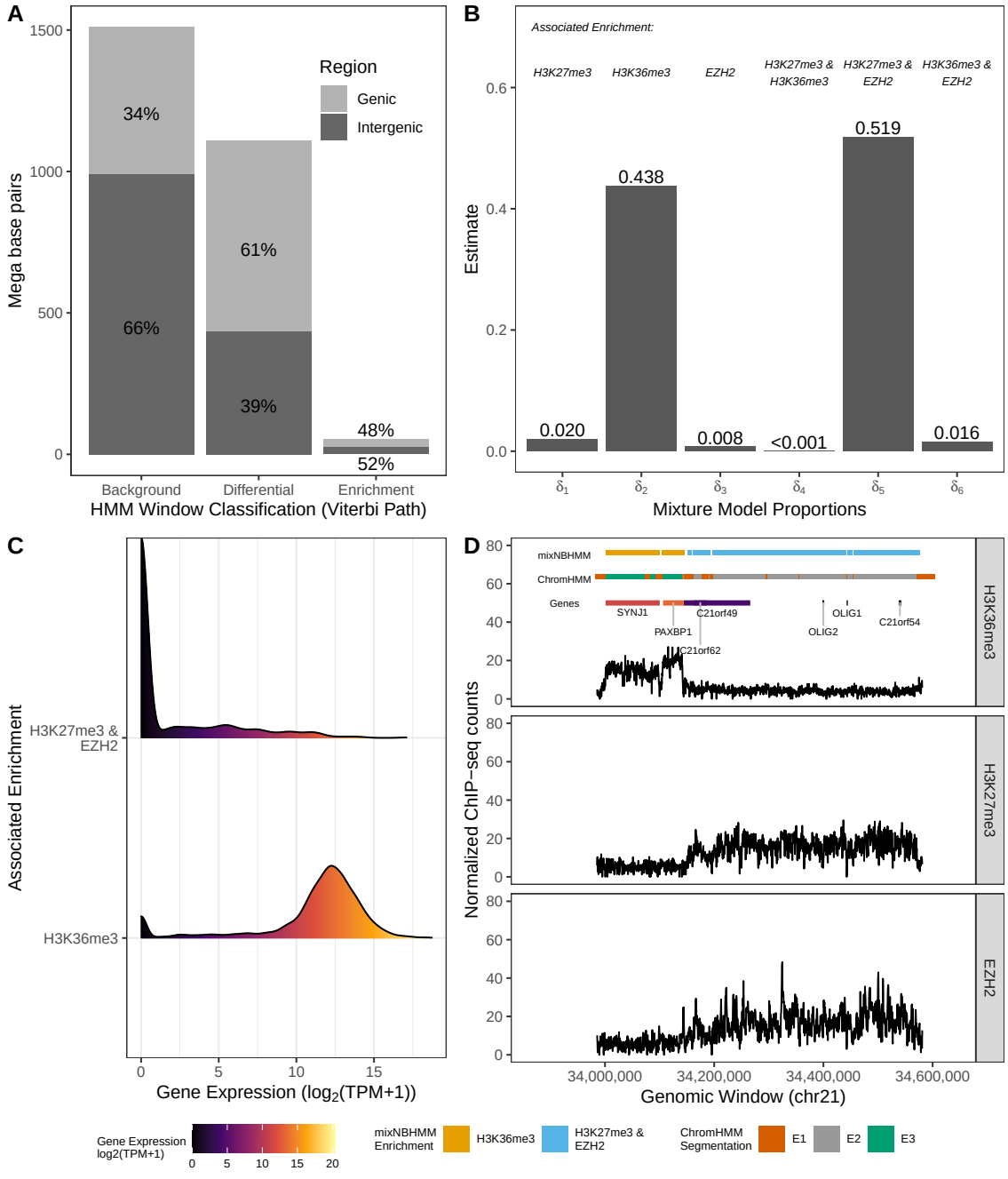

Figure 3.13: Genomic chromatin state segmentation and classification. (A): distribution of base pairs (y-axis) and the Viterbi sequence of states (x-axis). (B): estimated mixture probabilities and the associated differential combinatorial patterns. (C): density estimate from expression of genes intersecting differential peaks associated with the enrichment of H3K36me3 alone or the enrichment of H3K27me3 and EZH2 in consensus. (D): example of a genomic region with differential peaks and genes, colored according to their classification and expression levels, respectively.

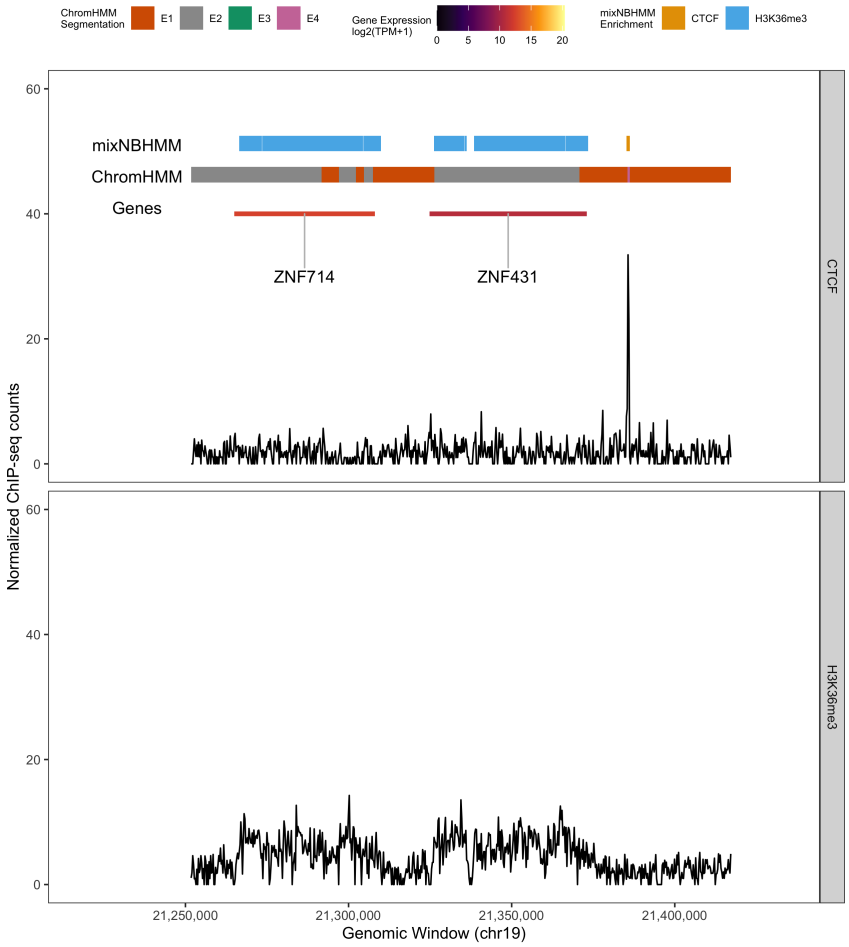

Figure 3.14: Genomic segmentation analysis of H3K36me3 and CTCF in Helas3 cell line. The chosen model parametrization and the normalization for non-linear biases via model offsets allow the segmentation of highly diverse epigenomic marks. The implemented hidden Markov model is able to properly account for the differences in length of enrichment regions between CTCF (short) and H3K36me3 (broad). Results from ChromHMM are shown for comparative purposes.

3.6 Discussion

We presented a flexible and efficient statistical model designed to call differential regions of enrichment from ChIP-seq experiments with multiple replicates and multiple conditions. Our model has three main advantages over current methods tailored for differential peak detection. First, it uses an HMM-based approach that accounts for the local dependency of ChIP-seq counts and is able to precisely detect broad and short differential regions of enrichment. Second, it utilizes a GLM-based framework with model offsets that account for potential non-linear biases in the data as well as a constrained parametrization across HMM states and mixture components. Our implementation of the RCEM algorithm led to genome-wide analyses of ChIP-seq data under a computational time comparable to some of the fastest current methods and was at least 5 times faster than current HMM-based algorithms. Lastly, our method allows the simultaneous detection and classification of differential combinatorial patterns of enrichment from its embedded mixture model and the associated posterior probabilities under any number of conditions. Our software has been implemented into an R package and is available for download (see Appenidix B).

CHAPTER 4: DEVELOPING STATISTICAL METHODOLOGY FOR THE ANALYSIS OF SINGLE-CELL CHIP-SEQ DATA: A COMPARATIVE STUDY OF CURRENT ALGORITHMS AND METHODOLOGICAL ADVANCES

4.1 Introduction

In the past decade, advances on single-cell epigenomic profiling of heterogeneous samples have allowed researchers to characterize previously unknown regulatory mechanisms within and between subpopulations of cells (Clark et al., 2016). Scientific areas that have recently benefited from such advances include developmental biology (Rotem et al., 2015; Buenrostro et al., 2018), cancer research (Grosselin et al., 2019; Granja et al., 2019), and immunology (Abdelsamed et al., 2020; ElTanbouly et al., 2020). Many single-cell transcriptomic studies have also benefited from the advances of single-cell epigenomic sequencing technologies, such as the single-cell Assay for Transposase Accessible Chromatin using sequencing (scATAC-seq; Cusanovich et al. (2018)), which overcame technological limitations and made the simultaneous analysis of thousands of cells possible. Utilizing either array-based, droplet-based, or combinatorial indexing coupled with split-pooling (Chen et al., 2019), scATAC-seq assays have become increasingly popular in recent years by allowing the discovery of open chromatin landscapes in individual cells, therefore providing an additional source of information to downstream transcriptomic applications. Yet, scATAC-seq allows a partial understanding of the mechanisms that form the epigenome, which also contains interactions between cellular elements and the genetic material that may affect processes within and between individual cells.

Single-cell chromatin immunoprecipitation followed by sequencing (scChIP-seq) is another single-cell epigenomic assay that allows the detection of biologically active regions in the chromatin with respect to a particular histone modification of interest (Rotem et al., 2015). The state of the art of scChIP-seq assays utilizes droplet microfluidics and unique molecular barcodes to simultaneous profile the activity of an epigenomic mark in thousands of cells. Recent studies utilizing scChIP-seq technologies have successfully assessed the roles of the histone modifications H3K27me3 and H3K4me3 on human breast cancer patient-derived xenografts (PDX) samples as well as H3K4me3 and H3K4me2 in mixed populations of embryonic cells and embryonic fibroblasts (Grosselin et al., 2019; Rotem et al., 2015). By utilizing the scChIP-seq

technology, these studies have explained parts of the heterogeneity seen in single-cell transcriptomic assays due to the activity of activating and repressing epigenomic marks in subpopulations of cells, a task that would not be possible to accomplish with conventional bulk assays.

In recent years, several methods designed for the analysis of scATAC-seq data have been proposed in the literature (González-Blas et al., 2019; Fang et al., 2019; Cusanovich et al., 2018; Baker et al., 2019). A recent study compared the performance of such methods in an extensive analysis utilizing simulated and real data (Chen et al., 2019). In summary, the ultimate goal of these methods is the clustering followed by the subsequent peak calling within subpopulations of cells exhibiting similar epigenomic profiles regarding the accessibility of their chromatin landscape. A few characteristics are shared across nearly all scATAC-seq methods benchmarked by Chen et al. (2019). First, current methods require as input a set of pre-specified genomic coordinates that are thought to characterize well the subpopulation of cells. Second, methods do not account or explicitly model the local dependency of single-cell counts in their analytical framework, a common characteristic of data resulting from single-cell and bulk ChIP-seq assays. Third, methods rely on a two-step procedure for clustering and characterization of subpopulations of cells regarding the epigenomic activity of interest. Although these features may be suitable for the analysis of scATAC-seq data, data resulting from scChIP-seq pose challenges to these methods. Since candidate peaks are often specified using bulk data, the choice of the peak calling algorithm and its parametrization can highly influence the final set of peaks, especially for broad marks, as we show in the Chapter 2 of this dissertation. Moreover, the low sequencing depth and relatively high noise of scChIP-seq experiments, in addition to the broadness of regions of activity from certain epigenomic marks, may cause these methods to have a limited performance in the analysis of scChIP-seq data (see Section 4.4). Due to these issues, current analyses of scChIP-seq data often use ad hoc approaches that are tailored for the particular problem at hand (Grosselin et al., 2019) without proper statistical justification on their use. Hence, there is a rising need for the development of statistical methods tailored for the analysis of scChIP-seq data.

Here, we present a benchmarking study of scATAC-seq methods on simulated scChIP-seq and scATAC-seq data (Section 4.4). In an extensive simulation study, we show that existing methods designed for scATAC-seq data are not optimized for some of the key characteristics of scChIP-seq experiments, namely the local dependency of counts, low signal to noise ratio, and low sequencing depth. Utilizing real data from scChIP-seq experiments (Grosselin et al. (2019); Section 4.2), we show that these methods exhibit varying

performance and that the results can be highly dependent on the choice of window size, a common problem in the analysis of bulk ChIP-seq experiments with broad regions of enrichment.

The methodological advances presented in this chapter are two-fold. First, we present a statistical model tailored to select candidate regions with differential activity of the epigenomic mark of interest from sparse scChIP-seq data. Utilizing simulated data, we show that selected differential regions better discriminate subpopulation of cells in downstream analysis by methods commonly used for single-cell clustering. This procedure is in contrast to detecting candidate enriched regions from pooled/aggregated data, a standard initial step in scATAC-seq data analysis. As a result, methods originally developed for scATAC-seq data exhibit a significant improvement in performance when clustering subpopulations of cells from scChIP-seq data. Finally, we propose the use of a model-based framework for joint clustering and characterization of structurally distinct cells (e.g., cells exhibiting varying patterns of enrichment of counts). The presented model is based on a class of mixture of hidden Markov models tailored for single-cell epigenomics (Section 4.3). For such a class of models, which accounts for the local dependency of sparse counts under a high resolution, we adapt the initialization algorithm presented by Smyth (1997) for fast initialization of the algorithm based on single-cell epigenomics data. The presented algorithm helps in the identification of the number of homogeneous population of cells regarding the activity of the epigenomic mark of interest, a necessary task in all scATAC-seq methods.

4.2 Analysis of scChIP-seq Data From Human Breast Cancer Patient-Derived Xenografts

Here, we utilize data from a scChIP-seq experiment from the histone modification mark H3K27me3 of human breast cancer patient-derived xenograft (PDX) samples (Grosselin et al., 2019). Specifically, scChIP-seq data from a pair of luminal estrogen receptor-positive breast Tamoxifen-resistant ($n = 200$ tumor cells) and Tamoxifen-sensitive ($n = 622$ tumor cells) PDXs were produced and the enrichment for H3K27me3 on 50 kb non-overlapping genomic windows was measured for all encapsulated single cells. Upon normalization of the raw counts and exclusion of outlier cells with total read counts less than 1600 or above the upper first percentile, principal component analysis (PCA) was applied and a consensus clustering algorithm was used to cluster the remaining cells ($n = 373$) into similar groups regarding the activity of H3K27me3. Grosselin et al. (2019) showed the existence of a subpopulation of cells from the drug-sensitive tumor that shared common characteristics with cells from resistant tumors. Specifically, these cells displayed similar loss of H3K27me3 activity, a mark associated with stable transcriptional repression, associated with

marker genes known to promote resistance to Tamoxifen treatment. Their analysis revealed differences in cells subpopulations that would be undetectable utilizing bulk ChIP-seq assays.

Due to the lack of statistical methods tailored for scChIP-seq data in the literature, we applied some of the current scATAC-seq methods on the scChIP-seq data from Grosselin et al. (2019) (Figure 4.15A-B). We observed that these methods exhibit varying performance on the scChIP-seq data regarding the clustering assignments of single cells (Figure 4.15C). While certain model-based algorithms tailored for scATAC-seq data exhibited a somewhat similar clustering assignment to the one from the annotated data set from Grosselin et al. (2019) (via adjusted Rand index, ARI; Rand (1971)), others failed to discriminate clusters of cells by (Figure 4.15D).

In Figure 4.16, we show results from current scATAC-seq methods with read counts calculated under different window sizes. It is worth noting that genomic windows of 50,000bp are not ideal for the analysis of H3K27me3, for which the enrichment can be present in much narrower windows. However, we include these results here for comparison purposes with the original analysis presented by (Grosselin et al., 2019). As methods rely on a set of candidate peaks (called on bulk data) to scrutinize single-cell subpopulations, results can be highly sensitive to the choice of window size, as reflected by the low ARI values for certain methods (e.g. SnapATAC). On the other hand, methods such as cisTopic, Cusanovich2018, and Scasat, exhibited very minor changes in clustering assignments under different data resolutions, as reflected by their high ARI values.

The lack of proper methods tailored for this type of data and the high variability of results from scATAC-seq methods make the choice of the best approach for the problem of clustering cells with scChIP-seq data challenging. We show in Section 4.4 of this chapter that the clustering performance of these methods can be highly sensitive to the sparsity, the low signal to noise ratio, and the local dependency present in scChIP-seq counts. By relying on a two-step approach for single-cell clustering (peak calling on the bulk data followed by clustering of single-cells based on candidate peaks), these methods exhibit poor performance in realistic scenarios. To the best of our knowledge, there is a lack of methods in the literature to detect differential regions of enrichment from single-cell data, which compromises the performance of current scATAC-seq methods in scChIP-seq data. As we show in Section 4.4.2, differential regions detected with single-cell data better discriminate sub populations of single-cells and their use as candidate regions improves the clustering assignments of existing approaches.

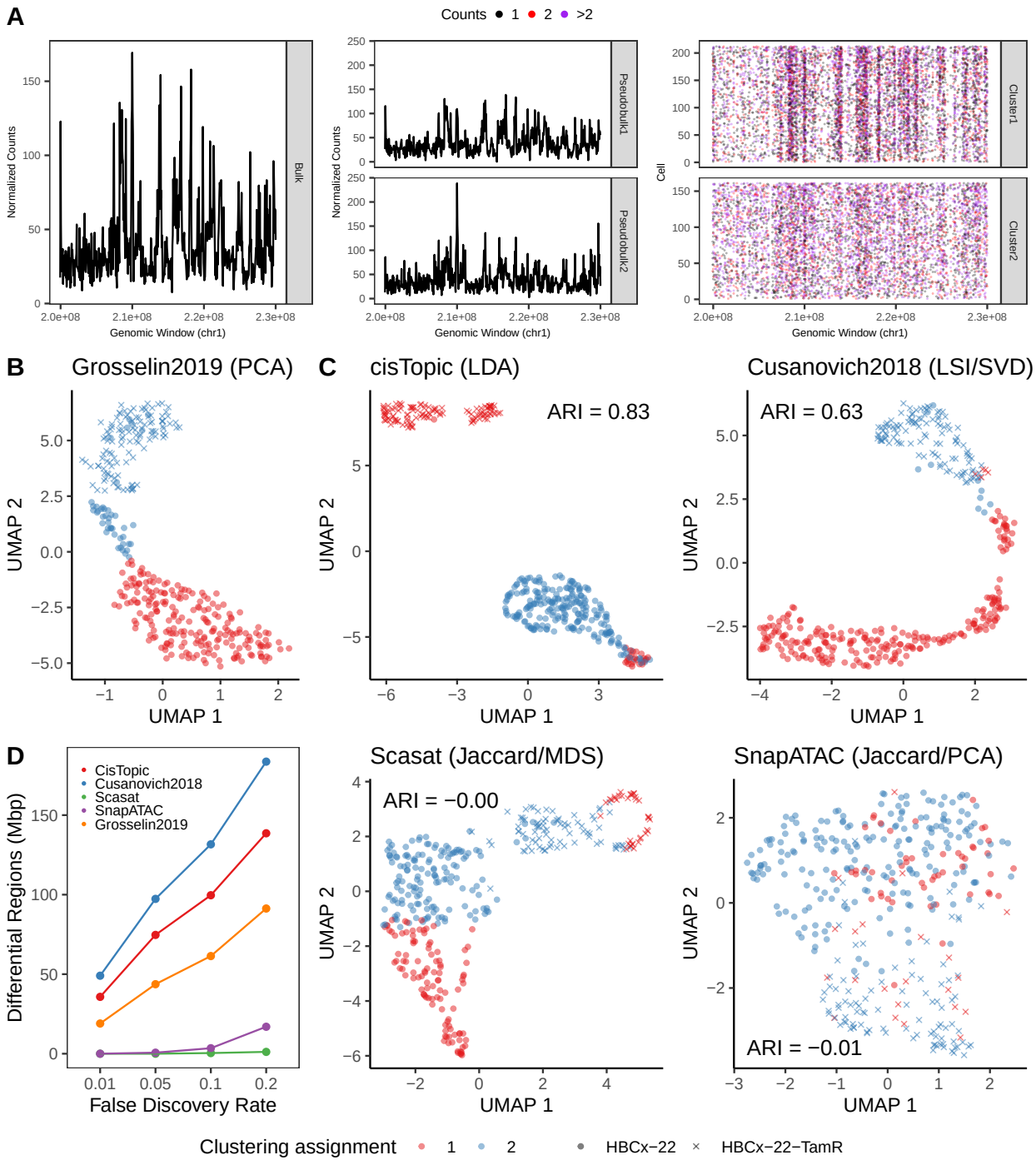

Figure 4.15: Analysis of H3K27me3 scChIP-seq data from drug sensitive (HBCx-22) and drug resistant (HBC-22-TamR) human breast cancer PDXs samples (Grosselin et al., 2019) with scATAC-seq methods. (A): original data from bulk and annotated pseudo bulks (clusters) using 50kb non-overlapping windows. (B): UMAP representation of original results from Grosselin et al. (2019). (C): UMAP projections of results using scATAC-seq methods, (D): total size of differential regions of enrichment between pseudo bulk of clustered cells for each method and different FDR values. Differential peaks were called using the methods presented in Chapter 3.

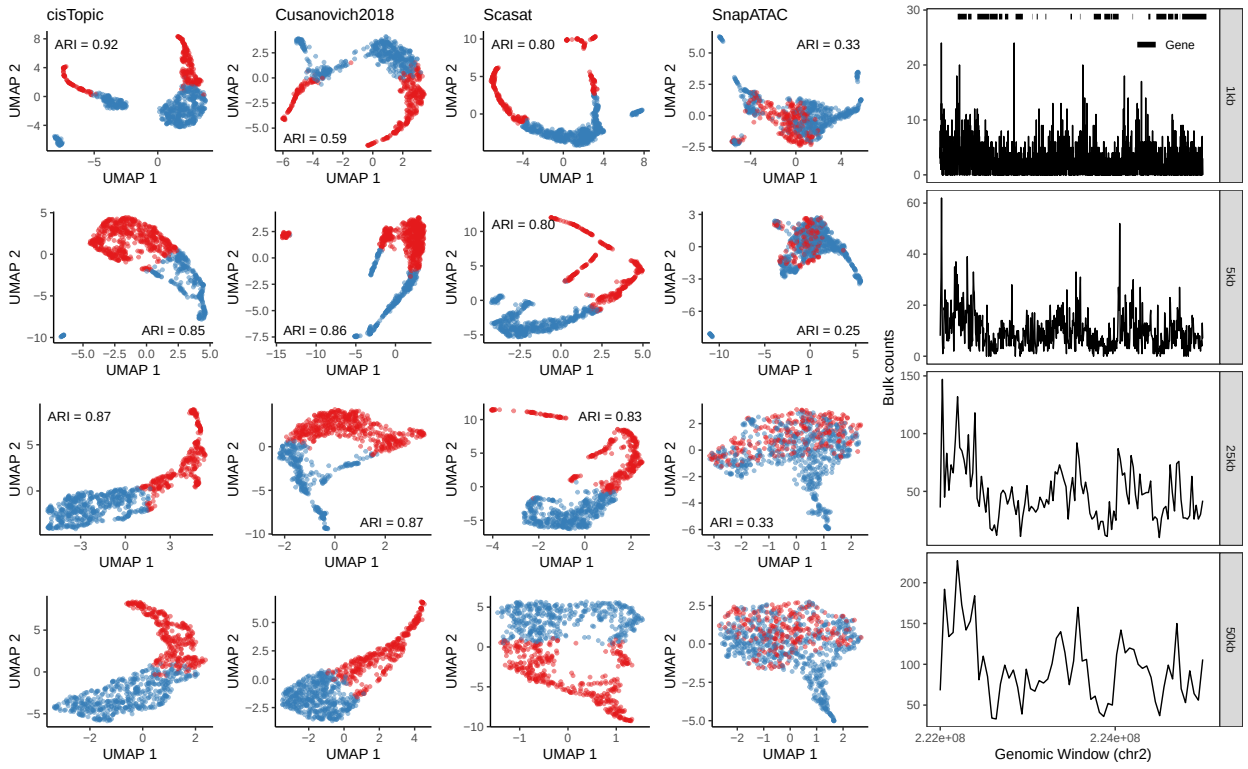

Figure 4.16: Application of current scATAC-seq methods on scChIP-seq data (Grosselin et al., 2019) under different resolutions. For each method, ARI compares clustering assignments between consecutive genomic resolutions.

To address some of the limitations of current scATAC-seq methods, this chapter present two strategies for single-cell clustering from scChIP-seq data. First, we introduce a hidden Markov model tailored to detect genomic regions with differential protein-DNA binding activity from epigenomic marks. Regions with differential activity better discriminate single-cell subpopulations and lead to a substantial improvement in current scATAC-seq methods' performance. Second, we propose the use of a class of mixture of hidden Markov models (Smyth, 1997) for simultaneous clustering and characterization of single-cells. Such a model is able to detect structural differences among single-cells, while accounting for the local dependency of sparse counts in high resolutions. Cells exhibiting similar structural patterns can then to be clustered together by making use of cluster-cell membership posterior probabilities. In addition, we propose a modified version of the initialization scheme presented by Smyth (1997) for the implemented EM algorithm that is based on Hellinger distances between cells. The presented scheme helps with the convergence of the clustering framework and helps with choosing the appropriate number of single-cell clusters, a necessary task for all benchmarked clustering methods for scATAC-seq data.

4.3 Methods

4.3.1 A Hidden Markov Model for Selecting Differentially Enriched Genomic Regions From Single-cell Data

4.3.1.1 Model Setup

Let Y_{ij} denote a binary indicator for the presence of sequencing reads on window i of cell j, for all $i = 1, \ldots, M$ and $j = 1, \ldots, N$. The binarization of read counts counts is a technique used by several existing algorithms (Cusanovich et al., 2018; González-Blas et al., 2019; Baker et al., 2019; Fang et al., 2019) and helps with challenges arising due to differences in sequencing depth among cells as well as the effects of PCR amplification artifacts (Chen et al., 2019). In addition, we assume that data is collected from diploid organisms in which no more than two copies of the DNA pertaining to a given genomic coordinate can be selected from an encapsulated single-cell. The data from N single-cells across M genomic windows can be organized in a $M \times N$ matrix $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$, such that $\mathbf{Y}_j = (Y_{1j}, \ldots, Y_{Mj})'$ for all $j = 1, \ldots, N$. At the i^{th} window, let $\mathbf{y}_{i\cdot} = (y_{i1}, \ldots, y_{iN})'$ denote the $N \times 1$ vector of observed scChIP-seq window read counts across all single cells, and let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ denote the corresponding observed matrix of single-cell read counts. We assume that each genomic window belongs to one of three possible hidden states: consensus background (state 1), differential (state 2), and consensus enrichment (state 3). Windows exhibiting low

(high) enrichment across all cells will be modeled by an emission distribution pertaining to the consensus background (enrichment) state. Windows exhibiting differential enrichment will be modeled by an emission distribution pertaining to the differential state.

To model transitions between states, we assume a single latent discrete time stationary Markov chain $\mathbf{Z} = \{Z_i\}_{i=1}^{M}$, $Z_i \in \{1, 2, 3\}$, with state-to-state transition probabilities $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{33})'$ and initial probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)'$, such that $\sum_{s=1}^{3} \gamma_{rs} = 1$ and $\sum_{s=1}^{3} \pi_s = 1$ for $r \in \{1, 2, 3\}$. To facilitate the notation, let $f_r(\mathbf{y}_{i.}|\boldsymbol{\psi}_r)$ denote the emission distribution corresponding to the r^{th} hidden state, where $\boldsymbol{\Psi} = (\boldsymbol{\pi}', \boldsymbol{\gamma}', \boldsymbol{\psi}')'$ denotes the vector of all model parameters, $\boldsymbol{\psi} = (\beta_1, \beta_2, \beta_3)'$ denotes each state's set of emission distribution-specific parameters, and $\mathcal{Z}$ denotes the set of 3^M possible state paths of length M. Then, the likelihood function pertaining to the proposed HMM may be written as

$$f(\mathbf{y}|\boldsymbol{\Psi}) = \sum_{\mathbf{Z} \in \mathcal{Z}} \left\{ \prod_{r=1}^{3} \pi_r^{I(Z_1 = r)} \times \left(\prod_{i=2}^{M} \prod_{r=1}^{3} \prod_{s=1}^{3} \gamma_{rs}^{I(Z_{i-1} = r, Z_i = s)} \right) \times \right. \tag{4.11}$$
$$\left. \times \left(\prod_{i=1}^{M} f_1\left(\mathbf{y}_{i.}|\beta_1\right)^{I(Z_i = 1)} f_2\left(\mathbf{y}_{i.}|\beta_2\right)^{I(Z_i = 2)} f_3\left(\mathbf{y}_{i.}|\beta_3\right)^{I(Z_i = 3)} \right) \right\}.$$

We assume that read counts pertaining to genomic windows from the consensus background ($r = 1$), differential ($r = 2$), and consensus enrichment ($r = 3$) states follow a Bernoulli distribution with state-specific parameter β_r. We ignore possible differential combinatorial patterns existing in the data due to single-cell unknown subpopulations and assume independence of read counts across cells, conditional upon the HMM state. The emission distribution of the consensus background, differential, and consensus enrichment states, respectively, can be written as

$$f_r(\mathbf{y}_{i.}|\boldsymbol{\psi}_r) = \prod_{j=1}^{N} p_{rij}^{y_{ij}} (1 - p_{rij})^{1 - y_{ij}}, \quad r \in \{1, 2, 3\}, \tag{4.12}$$

with $y_{ij} \in \{0, 1\}$, such that $\log(p_{rij}/(1 - p_{rij})) = \beta_r + u_{ij}$. The inclusion of offsets u_{ij} allows the adjustments for technical artifacts such as differences in sequencing depth between cells. When $u_{ij} = 0$, β_r represents the log-odds of observing at least a read count on genomics windows associated with state r.

The presented model allows the detection of genomic windows exhibiting differential enrichment for the epigenomic mark of interest from scChIP-seq studies. As we show in Section 4.4, by utilizing these regions

as candidate peaks in scATAC-seq methods improves substantially their performance in scChIP-seq data. State of the art analysis pipelines in single-cell epigenomics rely on sets of candidate regions that are found to be enriched in bulk (pooled) data or aggregated single-cell counts (Chen et al., 2019). Such an approach may compromise the downstream analysis in scChIP-seq data as epigenomic marks are often found to be active in broad genomic domains. As a result, pooled or aggregated single-cell counts may mask, pool, or even ignore differentially enriched windows that are important to discriminate single-cell subpopulations.

4.3.1.2 Estimation

We utilize the EM algorithm to estimate the parameters of the presented model. In the t^{th} step of the EM algorithm, the Q function of the complete data log-likelihood can be written as

$$\begin{aligned} Q\left(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(t)}\right) = & \sum_{r=1}^{3}\left\{Pr\left(Z_1 = r|\mathbf{y};\boldsymbol{\Psi}^{(t)}\right)\log(\pi_r)\right\} + \\ & + \sum_{i=2}^{M}\sum_{r=1}^{3}\sum_{s=1}^{3}\left\{Pr\left(Z_{i-1} = r, Z_i = s|\mathbf{y};\boldsymbol{\Psi}^{(t)}\right)\log(\gamma_{rs})\right\} + \\ & + \sum_{i=1}^{M} Pr\left(Z_i = 1|\mathbf{y};\boldsymbol{\Psi}^{(t)}\right)\log f_1\left(\mathbf{y}_{i.}|\beta_1\right) + \sum_{i=1}^{M} Pr\left(Z_i = 2|\mathbf{y};\boldsymbol{\Psi}^{(t)}\right)\log f_2\left(\mathbf{y}_{i.}|\beta_2\right) + \\ & + \sum_{i=1}^{M} Pr\left(Z_i = 3|\mathbf{y};\boldsymbol{\Psi}^{(t)}\right)\log f_3\left(\mathbf{y}_{i.}|\beta_3\right). \end{aligned} \tag{4.13}$$

In the E-step of the EM algorithm, we compute the posterior probabilities from Equation 4.13. The quantities $Pr\left(Z_i = r|\mathbf{y};\boldsymbol{\Psi}^{(t)}\right)$ and $Pr\left(Z_{i-1} = r, Z_i = s|\mathbf{y};\boldsymbol{\Psi}^{(t)}\right)$ can be calculated through the Forward-Backward algorithm in a similar fashion as presented in Appendix B. The Q function is maximized with respect to the parameters $\boldsymbol{\Psi} = (\boldsymbol{\pi}', \boldsymbol{\gamma}', \beta_1, \beta_2, \beta_3)'$ during the M-step of the algorithm. Estimates for the initial and transition probabilities can be directly calculated as $\hat{\pi}_r^{(t+1)} = Pr\left(Z_1 = r|\mathbf{y};\boldsymbol{\Psi}^{(t)}\right)$ and $\hat{\gamma}_{rs}^{(t+1)} = \sum_{i=2}^{M} Pr(Z_{i-1} = r, Z_i = s|\mathbf{y};\boldsymbol{\Psi}^{(t)}) / \sum_{i=2}^{M} Pr(Z_{i-1} = r|\mathbf{y};\boldsymbol{\Psi}^{(t)})$, respectively, restricted to $\sum_{r=1}^{3}\hat{\pi}_r^{(t+1)} = 1$ and $\sum_{s=1}^{3}\hat{\gamma}_{rs}^{(t+1)} = 1$, for $r \in \{1, 2, 3\}$. Estimating $(\beta_1, \beta_2, \beta_3)'$ from Equation 4.13 can be seen as obtaining parameter estimates from a series of weighted logistic regression models. We independently estimate these quantities via the algorithm BFGS (Fletcher, 2013).

Upon convergence of the algorithm, the final set of HMM posterior probabilities can be used to segment the genome into consensus background, differential, or consensus enrichment windows. Approaches that control the total false discovery rate (FDR) via posterior probabilities (Efron et al., 2001) or that estimate the most likely sequence of hidden states (Viterbi, 1967) can be used for such purposes. Let

$\hat{\rho}_{i2} = Pr\left(Z_i = 2|\mathbf{y}; \hat{\mathbf{\Psi}}\right)$ denote the estimated posterior probability that the i^{th} genomic window belongs to the differential HMM state, $i = 1, \ldots, M$. For a cutoff of posterior probability α, the total FDR is $\sum_{i=1}^{M}(1 - \hat{\rho}_{i2})I(\hat{\rho}_{i2} \geq 1 - \alpha) / \sum_{i=1}^{M} I(\hat{\rho}_{i2} \geq 1 - \alpha)$, where $I(\cdot)$ is an indicator function. The posterior probability cutoff is then chosen by controlling the total FDR. Differential regions of enrichment are formed by merging adjacent windows that meet a given FDR threshold level for the differential HMM state. These regions can then be used as candidate peaks in scATAC-seq methods for clustering of single-cells from scChIP-seq data (Section 4.4.2).

4.3.2 A Mixture of Hidden Markov Models for Simultaneous Clustering and Characterization of Single-cells

Here, we present a statistical model defined as a mixture of hidden Markov models (MHMM) for single-cell clustering. The MHMM has been extensively studied in the literature (Smyth, 1997; Jebara et al., 2007; Vermunt et al., 2008) and here we present a modification of this model to account the data characteristics of scChIP-seq counts on a high resolution of read counts. In this context, the purpose of the MHMM is to cluster similar cells that share common structural characteristics with respect to the local dependency of observed counts. Specifically, the model assumes that sequences of counts are generated from a latent set of L hidden Markov models (L known *a priori*) and it allows the simultaneous estimation of model parameters and clustering of similar single-cells.

As such, the MHMM does not account for longitudinal differences in the read count distribution across the genome between cells. For instance, multiple realizations of the same process (or hidden Markov chain) are deemed similar and clustered together by the algorithm, despite differences with respect to the location of enrichment regions in the data. Yet, it allows the clustering of cells that share common characteristics of count enrichment across the genome. The presented model could be used for the purpose of distinguishing outliers cells, a common task in single-cell epigenomic studies (Jia et al., 2018) that can impact the clustering performance of current methods (Chen et al., 2019). Hence, the presented statistical model needs further consideration for the problem of distinguishing multiple realization of the same hidden process and such a task will be left as a future research project.

4.3.2.1 Model Setup

Let Y_{ij} denote a binary indicator for the presence of a sequencing read on window i of cell j, for all $i = 1, \ldots, M$ and $j = 1, \ldots, N$. The data from N single-cells across M genomic windows can be organized in a $M \times N$ matrix $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$, such that $\mathbf{Y}_j = (Y_{1j}, \ldots, Y_{Mj})'$ for all $j = 1, \ldots, N$. Next, we

assume that for each cell j there is an associated latent L-dimensional variable $\mathbf{W}_j = (W_{1j}, \ldots, W_{Lj})'$ that indicates the cell's membership to one of L (known) possible subpopulations of cells. We assume that $\mathbf{W}_j \sim$ Multinomial$(1, \boldsymbol{\delta})$ with $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_L)'$ and $\sum_{l=1}^{L} \delta_l = 1$. Let $\mathbf{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_N)$ denote the $L \times N$ latent binary matrix of subpopulation-cell membership. The ultimate goal of the presented model is to estimate the posterior probabilities of subpopulation-cell membership $\mathbf{W}$ conditional on the observed data $\mathbf{y}$.

We assume that there is a latent first-order Markov chain $\mathbf{Z} = (Z_1, \ldots, Z_M)'$, such that $Z_i \in \{1, \ldots, K\}$ with $K = 2L$, that guides the presence and absence of read enrichment for each of the L subpopulations of cells. Conditionally on cell's memberships, one can define such a Markov chain in terms of initial probabilities $\boldsymbol{\pi} = (\boldsymbol{\pi}_1 \ldots, \boldsymbol{\pi}_L)'$, with $\boldsymbol{\pi}_l = (\pi_{l1}, \pi_{l2})'$ for all $l = 1, \ldots, L$ and $\sum_{l=1}^{L} \sum_{u=1}^{2} \pi_{lu} = 1$, and transition probabilities $\boldsymbol{\gamma}$ such that

$$\boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\gamma}_2 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\gamma}_3 & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\gamma}_L \end{pmatrix}, \quad \text{where} \quad \boldsymbol{\gamma}_l = \begin{pmatrix} \gamma_{l,11} & \gamma_{l,12} \\ \gamma_{l,21} & \gamma_{l,22} \end{pmatrix} \quad \forall l = 1, \ldots, L.$$

The data generating process according to this model can interpreted as follows. Conditionally on the j^{th} cell's cluster membership $\mathbf{W}_j$, a sequence of binary counts indicating the presence of sequencing reads on genomic windows is generated from one of the l absorbing subchains with transition probabilities $\boldsymbol{\gamma}_l$. Such a model does not allow transitions between subchains, and each subchain characterizes the data from a particular subpopulation of cells. This model is a finite mixture of hidden Markov models (MHMM), which allows the marginal likelihood for the observed data to be written as a mixture of L probabilistic distributions of independent HMMs (Equation 1, in Smyth (1997)).

Conditionally on $\mathbf{Z}$, we assume independence among cell's observed and latent data $(\mathbf{Y}_j, \mathbf{W}_j)$ for all $j = 1, \ldots, N$. Hence, the complete data likelihood for the the presented MHMM can be written as

$$Pr(\mathbf{Y}, \mathbf{W}, \mathbf{Z}|\boldsymbol{\Psi}) = \prod_{j=1}^{N} Pr(\mathbf{Y}_j|\mathbf{W}_j, \mathbf{Z}; \boldsymbol{\Psi}) Pr(\mathbf{Z}|\mathbf{W}_j; \boldsymbol{\Psi}) Pr(\mathbf{W}_j|\boldsymbol{\Psi}). \tag{4.14}$$

To define each term of the right hand side of Equation 4.15, we first define the set $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_L)$ such that $\mathbf{S}_l = \{2l-1, 2l\}, \forall l = 1, \ldots, L$. Then, we have

$$
\begin{aligned}
Pr(\mathbf{W}_j|\mathbf{\Psi}) &= \prod_{l=1}^{L} \delta_l^{W_{lj}}, \qquad (4.15)\\
Pr(\mathbf{Z}|\mathbf{W}_j;\mathbf{\Psi}) &= \prod_{l=1}^{L} \left\{ \prod_{u=1}^{2} \pi_{lu}^{I(Z_l=S_{lu},W_{lj}=1)} \prod_{i=2}^{M} \prod_{u=1}^{2} \prod_{v=1}^{2} \gamma_{l,uv}^{I(Z_{l-1}=S_{lu},Z_l=S_{lv},W_{lj}=1)} \right\}, \quad \text{and}\\
Pr(\mathbf{Y}_j|\mathbf{W}_j,\mathbf{Z};\mathbf{\Psi}) &= \prod_{l=1}^{L} \left\{ \prod_{i=1}^{M} \prod_{u=1}^{2} f_{lu}(y_{ij}|\beta_{lu})^{I(Z_l=S_{lu},W_{lj}=1)} \right\}, \quad \forall \quad j = 1, \ldots, N.
\end{aligned}
$$

We assume that the emission distribution under state S_{lu} (the u^{th} component of the l^{th} subchain), $f_{lu}(y_{ij}|\boldsymbol{\beta}_{lu})$, is a Bernoulli distribution with parameter $p_{ij}^{(lu)} = \exp(\beta_{lu} + o_{ij})/(1 + \exp(\beta_{lu} + o_{ij}))$. Such a parametrization allows the inclusion of continuous covariates and offsets o_{ij} that are thought to help with the normalization of counts across cells and genomic windows.

4.3.2.2 Estimation

We utilize the EM algorithm to estimate the model parameters and the posterior probabilities of cluster membership for each cell, conditional on the observed data. Specifically, the Q-function of the EM algorithm for the presented model is

$$
\begin{aligned}
Q(\mathbf{\Psi}|\mathbf{\Psi}^{(t)}) &= \sum_{j=1}^{N} E_{\mathbf{Z},\mathbf{W}_j} \left[\log \left[Pr(\mathbf{y}_j|\mathbf{W}_j,\mathbf{Z};\mathbf{\Psi}) Pr(\mathbf{Z}|\mathbf{W}_j;\mathbf{\Psi}) Pr(\mathbf{W}_j|\mathbf{\Psi}) \right] |\mathbf{y}_j;\mathbf{\Psi}^{(t)} \right],\\
&= \sum_{j=1}^{N} \left\{ \sum_{l=1}^{L} P\left(W_{lj=1}|\mathbf{y}_j;\mathbf{\Psi}^{(t)}\right) \log(\delta_{ls}) + \right.\\
&+ \sum_{l=1}^{L} \sum_{u=1}^{2} P\left(Z_1 = S_{lu}, W_{jl} = 1|\mathbf{y}_j;\mathbf{\Psi}^{(t)}\right) \log(\pi_{lu}) +\\
&+ \sum_{l=1}^{L} \sum_{i=2}^{M} \sum_{u=1}^{2} \sum_{v=1}^{2} P\left(Z_{i-1} = S_{lu}, Z_i = S_{lv}, W_{jl} = 1|\mathbf{y}_j;\mathbf{\Psi}^{(t)}\right) \log(\gamma_{l,uv}) +\\
&+ \left. \sum_{l=1}^{L} \sum_{i=1}^{M} \sum_{u=1}^{2} P\left(Z_i = S_{lu}, W_{jl} = 1|\mathbf{y}_j;\mathbf{\Psi}^{(t)}\right) \log f_{lu}(y_{ij}|\beta_{lu}) \right\}, \qquad (4.16)
\end{aligned}
$$

which follows from the linearity property of expectation and the independence between $\mathbf{W}_j$ and $\mathbf{Y}_{j'}^{'}$, for all $j \neq j^{'}$.

The posterior probabilities present in Equation 4.16 can be calculated via Forward-Backward algorithm for each cell and for each subchain. Let $F^j_{i,S_{lu}}$ and $B^j_{i,S_{lu}}$ denote the forward and backward probabilities, respectively, for the j^{th} cell at the i^{th} genomic position associated with the u^{th} state of the l^{th} subchain. Then, one can show that

$$
\begin{aligned}
P\left(Z_i = S_{lu} W_{lj} = 1 | \mathbf{y}_j; \mathbf{\Psi}^{(t)}\right) &= \frac{F^j_{i,S_{lu}} B^j_{i,S_{lu}} \delta^{(t)}_l}{\sum_{k=1}^{L}\sum_{v=1}^{2} F^j_{M,S_{kv}} \delta^{(t)}_k}, \\
P\left(Z_{i-1} = S_{lu}, Z_i = S_{lv}, W_{lj} = 1 | \mathbf{y}_j; \mathbf{\Psi}^{(t)}\right) &= \frac{\delta^{(t)}_l F^j_{(i-1),S_{lu}} \gamma^{(t)}_{l,uv} f_{lv}\left(y_{ij} | \beta^{(t)}_{lv}\right) B^j_{i,S_{lv}}}{\sum_{k=1}^{L} \delta^{(t)}_k \sum_{v=1}^{2} F^j_{M,S_{kv}}}, \quad \text{and} \\
P\left(W_{lj} = 1 | \mathbf{y}_j; \mathbf{\Psi}^{(t)}\right) &= \frac{\delta^{(t)}_l \sum_{v=1}^{2} F^j_{M,S_{lv}}}{\sum_{k=1}^{L} \delta^{(t)}_k \sum_{v=1}^{2} F^j_{M,S_{kv}}}, \qquad (4.17)
\end{aligned}
$$

for all $j = 1, \ldots, N$ and $l = 1, \ldots, L$.

The E-step of the EM algorithm proceeds as follows. First, one estimates (in parallel for all $l = 1, \ldots, L$ subchains) the quantities $F^j_{i,S_{lu}}$ and $B^j_{i,S_{lu}}$, for all $j = 1, \ldots, N$, $i = 1, \ldots, M$, and $u = 1, 2$. Second, estimates for $P\left(W_{lj} = 1 | \mathbf{y}_j; \mathbf{\Psi}^{(t)}\right)$, for all $l = 1, \ldots, L$ and $j = 1, \ldots, N$, are calculated using Equation 4.17. Third, one estimates (in parallel for all $l = 1, \ldots, L$ subchains) all remaining posterior probabilities $P\left(Z_i = S_{lu} W_{lj} = 1 | \mathbf{y}_j; \mathbf{\Psi}^{(t)}\right)$ and $P\left(Z_{i-1} = S_{lu}, Z_i = S_{lv}, W_{lj} = 1 | \mathbf{y}_j; \mathbf{\Psi}^{(t)}\right)$. In the M-step, estimates for $\boldsymbol{\delta}$, $\boldsymbol{\pi}$, and $\boldsymbol{\gamma}$ can be obtained from closed form solutions. Finally, the estimation of the parameters β_{lu} can be done via weighted logistic regression in parallel for all L subchains. At convergence, the posterior probabilities $P\left(W_{lj} = 1 | \mathbf{y}_j; \mathbf{\Psi}^{(t)}\right)$ allow the clustering of cells sharing similar data generating processes of sequencing reads.

The EM algorithm was efficiently implemented to facilitate the application to a large number of cells and clusters. By assuming the Bernoulli distribution for the binary counts, one can efficiently store data in memory from thousands of single-cells in high resolution in sparse matrix format. In addition, the computation of posterior probabilities during the EM algorithm can be efficiently done in parallel and all output can be stored (and loaded) during the E-step of the algorithm either in binary or HDF5 format using the C++ library Armadillo (Sanderson and Curtin, 2016). During the M-step, the aggregation of counts for each subchain simplifies the problem of estimation for thousands of cells in high resolutions and leads to a fast optimization via weighted logistic regression.

4.3.2.3 Initializing the EM Algorithm and Learning the Number of Clusters

We introduce an initialization scheme to obtain the initial parameter estimates from the MHMM for the EM algorithm, namely $\boldsymbol{\Psi}^{(0)} = (\boldsymbol{\pi}^{(0)}, \boldsymbol{\gamma}^{(0)}, \boldsymbol{\delta}^{(0)}, \boldsymbol{\beta}^{(0)})'$. The presented scheme, which aids the determination of the number of single-cell clusters, is a modified version of the scheme from Smyth (1997), which can be time-consuming in the context of single cell epigenomics. Specifically, Smyth (1997) uses a HMM likelihood-based distance matrix for the initialization of the EM algorithm for the MHMM. For a total of N cells (where N can be the tens of thousands), their approach would require the evaluation of N^2 calculations of the M-dimensional sets of forward probabilities (where M can be greater than 10^6). Instead, we propose the use of a fast cell-to-cell Hellinger distance calculation based on posterior probabilities from enrichment states from individual HMMs to determine the initial number of clusters and initial parameters of the MHMM:

1. In parallel, fit N 2-state HMMs, one to each single-cell $\mathbf{Y}_j$, $j = 1, \ldots, N$, with initial probabilities $\boldsymbol{\pi}^{(j)} = (\pi_1^{(j)}, \pi_2^{(j)})'$, transition probabilities $\boldsymbol{\gamma}^{(j)} = (\gamma_{11}^{(j)}, \ldots, \gamma_{22}^{(j)})'$, and Bernoulli emission distributions with parameters 2 and $p_i^{(ju)} = \exp(\beta_u^{(j)} + o_{ij})/(1 + \exp(\beta_u^{(j)} + o_{ij}))$, for $u = 1, 2$ and $i = 1, \ldots, M$. Aggregation of counts and a rejection-controlled EM algorithm can be used for computational efficiency.

2. From the estimated window-based posterior probabilities of enrichment of each HMM, w_{ij}, $i = 1, \ldots, M$ and $j = 1, \ldots, N$, calculate the cell-to cell $N \times N$ Hellinger distance matrix D with entries $d_{jj'} = \sqrt{2 \sum_{i=1}^{M} \left(\sqrt{w_{ij}} - \sqrt{w_{ij'}}\right)^2}$, for all $j = 1, \ldots, N$ and $j' \neq j$.

3. Use the distance matrix D to learn the number of clusters L and cluster cells into L groups.

4. Initialize the MHMM with L clusters and parameters $\boldsymbol{\Psi}^{(0)}$ created by taking cluster-specific averages across cells of $\hat{\boldsymbol{\pi}}^{(j)}$, $\hat{\boldsymbol{\gamma}}^{(j)}$, and $\hat{\boldsymbol{\beta}}^{(j)}$. Initialize $\delta_l^{(0)}$ as the proportion of cells assigned to cluster l, $l = 1, \ldots, L$.

The purpose of the above initialization scheme is to improve the convergence time of the EM algorithm, in which the information from all cells and genomic windows are then utilized for single-cell clustering. In simulation studies, (Smyth, 1997) compared different initialization schemes that differed regarding the choice of the initial transition probability matrix. The presented scheme led to the highest log-likelihood value for the observed data in comparison to the random initialization with and without block diagonal transition matrices. Here, we evaluated the performance of the current method with different initialization values and observed

that the presented scheme leads to the fastest convergence of the EM algorithm with final parameter estimates close to the true values. However, further consideration is needed to understand the rate of convergence of the implemented method as a function of the number of cells, genomic windows, and clusters in the data.

Clustering algorithms, such as hierarchical clustering, can be used for the purpose of learning the number of clusters L in the data, which is a necessary task in current scATAC-seq methods. While the computing time of steps (1) and (2) can be significantly reduced by utilizing parallel computing, suitable modifications could be implemented to further reduce the computing burden. These include downsampling single-cells or genomic windows, using C++ libraries developed for large-scale distance computations of sparse matrices, which can be created by FDR-thresholding posterior probabilities of enrichment for each cell, or even utilizing a cell-specific moving-average smoothing in step (1) to reduce the noise and avoid the computations of individual HMMs.

Figure 4.17 shows the application of MHMM on a simulated dataset. We simulated counts for 150 cells and 3 clusters (50 cells/cluster; Figure 4.17A) and applied the initialization scheme to compute the initial values of the EM algorithm (Figure 4.17B). The distance matrix D allows the visualization of the number of similar subpopulations in the data. Upon convergence of the EM algorithm, one can visualize the cell-cluster posterior probability membership and group cells into similar subpopulations.

4.4 Simulation Studies

4.4.1 Benchmarking Study of Current scATAC-seq Methods on scChIP-seq Data

We evaluated the performance of current algorithms designed for scATAC-seq data on simulated scATAC-seq and scChIP-seq assays. Methods were compared regarding their accuracy of clustering assignment by means of the ARI and the adjusted mutual information (AMI, Vinh et al. (2010)) with the gold-standard labels of simulated data clusters. In addition, methods we compared regarding the computing time in analyzing the data. Under similar scenarios regarding sequencing depth and number of cells/clusters, current approaches failed to characterize subpopulation of cells when the enrichment of reads was not concentrated in short genomic regions, but rather spread across broad regions, a characteristic of scChIP-seq data.

Read counts were simulated based on real scATAC-seq and scChIP-seq experiments as follows. First, the genome was tiled into M non-overlapping windows of 250bp and sequencing reads from annotated single-cell subpopulations released by Buenrostro et al. (2018) (scATAC-seq) and Grosselin et al. (2019) (scChIP-seq) were mapped onto the resulting non-overlapping windows and tabulated. For a given sequencing depth level

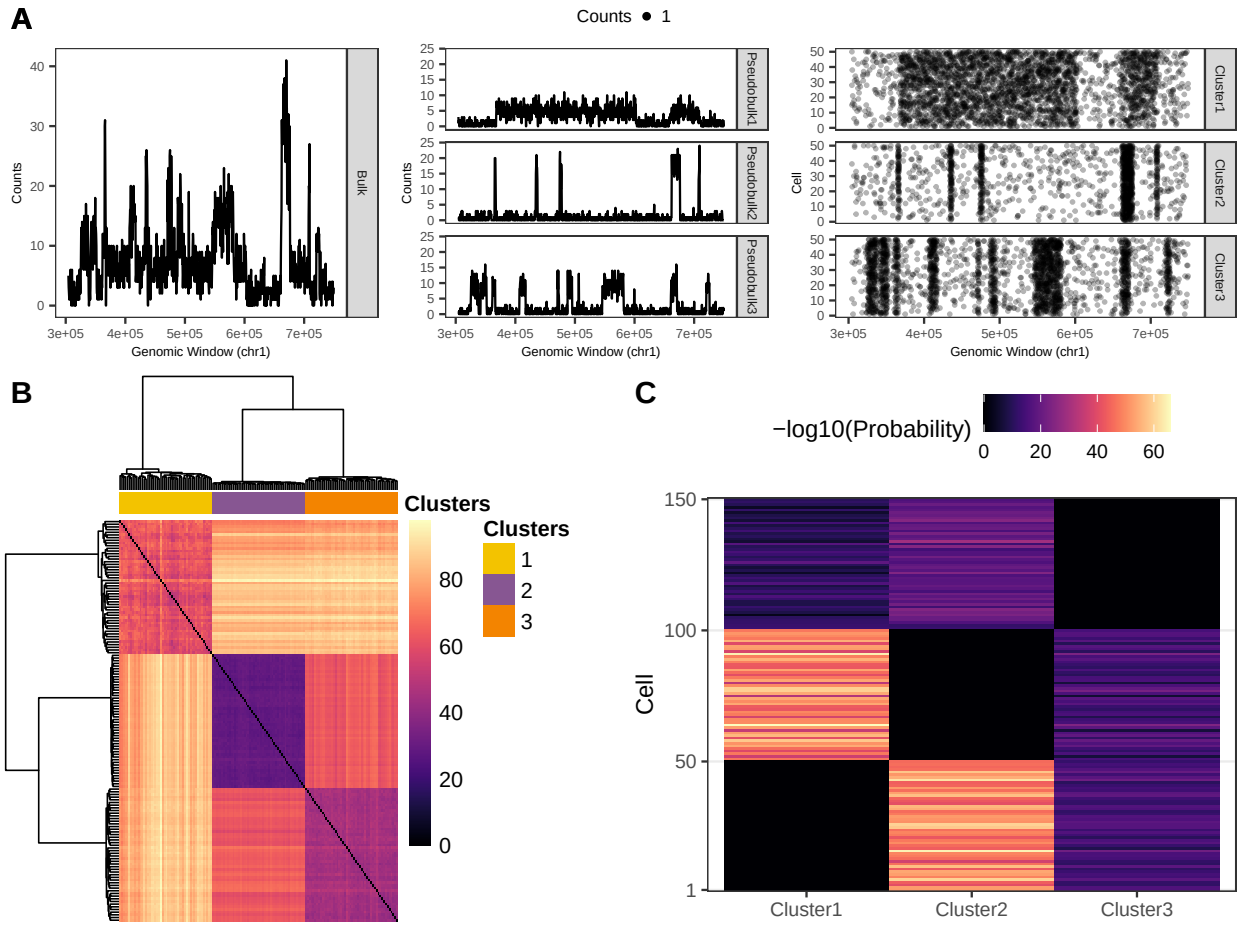

Figure 4.17: Application of MHMM on a simulated data. (A): counts for the bulk, pseudobulk, and clusters of cells. (B): heatmap of the cell-to-cell Hellinger distance matrix D from the initialization scheme. (C): estimated posterior probabilities of cluster membership for simulated cells upon convergence of the algorithm. Colors lighter than purple in the continuous scale indicate estimated posterior probabilities lower than 10^{-20}.

d and noise level z, read counts from window i, $i = 1, \ldots, M$, of each cell were simulated according to a binomial distribution with parameters 2 and $p_i^{(d,n)} = d(1-z)w_i/2 + dz/(2M)$, where w_i is the observed proportion of reads from real data associated with genomic window i. Under this distributional assumption, the expected sequencing depth of a cell is d. For lower noise levels, the distribution of counts across the genome resembles the one from real data on the single-cell level. For higher noise levels, the distribution of counts resembles the one from independent binomial trials. We assume in this simulation study that data is collected from diploid organisms in which no more than two copies of the DNA pertaining to a given genomic coordinate can be selected from an encapsulated single-cell. A similar assumption was made in the simulation study presented in Chen et al. (2019).

We evaluated depth levels of 5,000, 10,000, and 25,000 reads per cell (25,000 read pairs is the recommended sequencing depth by 10X Genomics for scATAC-seq libraries). In Grosselin et al. (2019), the median (mean) genome-wide sequencing depth was 3,651 (10,228) reads per cell. Due to the sparsity of scChIP-seq reads, we only assessed different noise levels for scATAC-seq data (0 and 0.25). A noise level of 0.25 indicates that, on average, read counts from 25% of the genomic windows were simulated under the assumption that genomic windows had equal probability of having an assigned sequencing read for a given sequencing depth (see formula in the previous paragraph). In all the scenarios, the sequencing depth and noise levels were assumed to be constant for all cells, in agreement to the benchmarkinng study presented by (Chen et al., 2019). By maintining a constant sequencing depth and noise level among single-cells, one can better measure the performance of current algorithm under contolled scenarios without any nuisance source of variation. We studied the performance of current methods in scenarios assuming 3 and 6 sub population of cells. Because the annotated data from Grosselin et al. (2019) only had 2 subpopulations of similar cells, we used an ad hoc approach to simulate clusters for scChIP-seq data. Specifically, reads from different clusters were simulated after repeatedly shuffling the observed proportions w_i in non-overlapping blocks of size 5000bp across the genome such that nearly either 1% or 5% of the final genome was formed by shuffled windows (estimated from real data; Figure 4.15). The simulated numbers of cells per cluster were 500, 1000, and 2500. In Grosselin et al. (2019), the numbers of cells per cluster were 212 and 161. Finally, we also assessed the capability of current method in detecting rare cell subpopulations in which the total number of cells for a given cluster was reduced to 10% of its original size. A hundred simulated datasets were generated for all evaluated scenarios.

For both scATAC-seq and scChIP-seq simulated data, peaks were called on the bulk data and then regions of enrichment (and their associated counts) were passed as input to the evaluated methods, which require a pre-specified set of candidate regions for single-cell clustering. We called peaks using the algorithm presented in Baldoni et al. (2019a), which is flexible for narrow and broad marks, using counts computed on non-overlapping windows of 500bp. For each method, we followed the analysis presented in Chen et al. (2019) and compared the methods' performance according to three commonly used clustering algorithms on their final feature matrix, namely K-means, hierarchical clustering, and Louvain (Kiselev et al., 2019). We used the adjusted Rand index and compared the final set of clustering assignments with the true cluster labels from simulated data.

4.4.1.1 Simulation Results

Results from this simulation study show that current methods developed for scATAC-seq experiments presented satisfactory results on scATAC-seq-like simulated data, as expected (Figure 4.18A; scenario with 3 clusters, 500 cells/cluster, and 10,000 reads per cell). The choice of the clustering algorithm did not appear to have an influence in the overall clustering performance and SnapATAC was the only method that exhibited difficulty in assigning single-cells to their respective clusters (Figure 4.18B). Regarding computing time, all methods were able to cluster cells in less than a minute, on average, although cisTopic was consistently more time consuming than other algorithms (Figure 4.18C). Figure 4.18D shows the UMAP projections of the feature matrix from each method for a given simulated data set. In general, we observed that most methods had satisfactory performance in terms of ARI (comparing with true simulated labels) in all simulated scenarios of scATAC-seq data except when noise was introduced to the data. When noise was present, a higher sequencing depth was necessary for all methods but SnapATAC to achieve an average ARI (AMI) greater than 0.87 (Table 4.6).

Next, we applied current methods on scChIP-seq simulated data (Figure 4.19A; scenario with 3 clusters, 500 cells/cluster, and 10,000 reads per cell). We observed that the clustering assignments from current methods had a moderate to low agreement with the true cluster memberships (Figure 4.19B), despite of the choice of the clustering algorithm, even in a scenario with excessively high sequencing depth (10,000 versus a median depth of 3,651 in Grosselin et al. (2019)). Similar to simulated scATAC-seq data, all methods performed the analysis under a reasonable computing time, although cisTopic was consistently more time consuming. Figure 4.19D shows the UMAP representation of the feature matrix from each method for a given simulated data set. Overall, current methods had limited performance in all simulated scenarios for

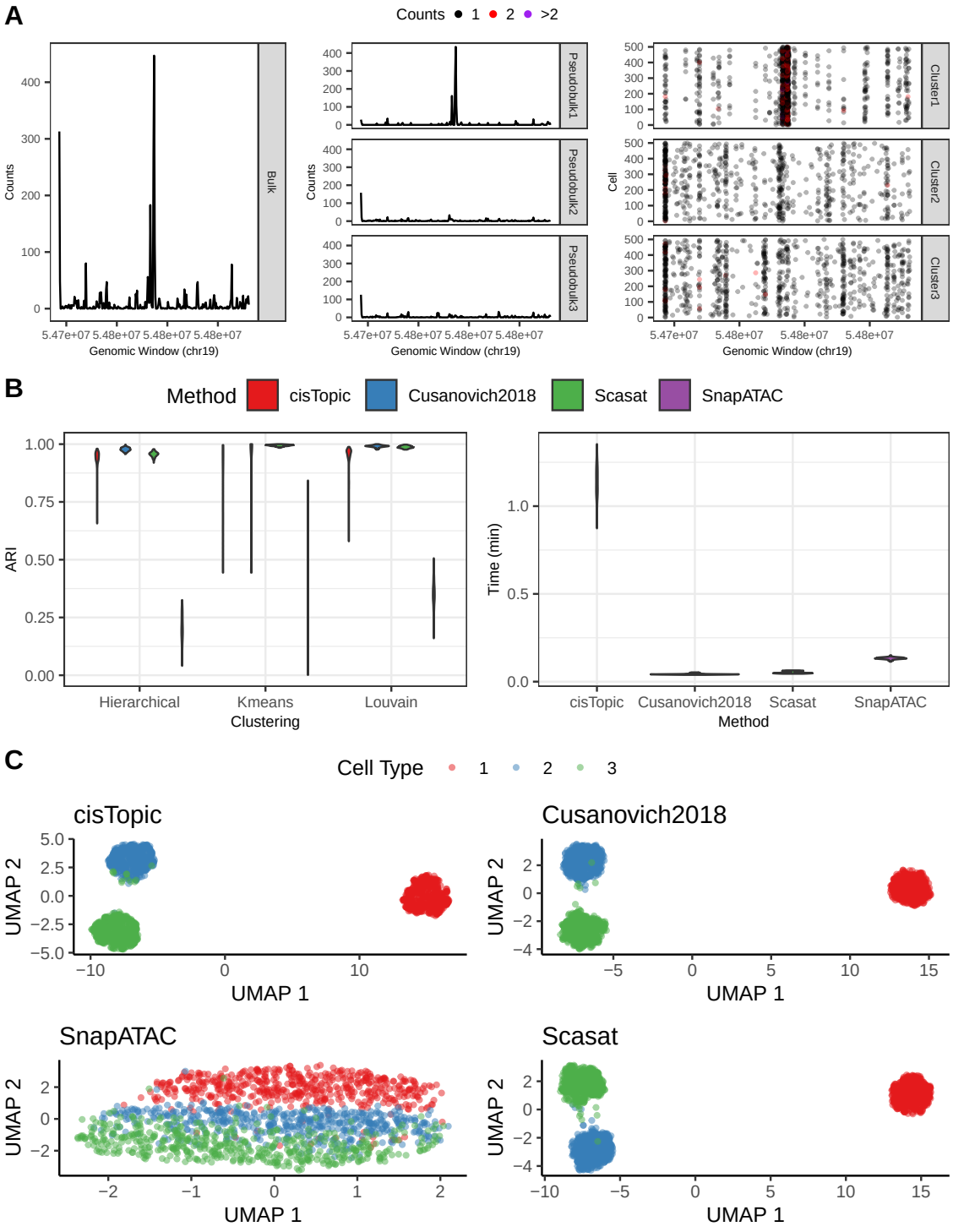

Figure 4.18: Results from simulated scATAC-seq data for the scenario with 3 clusters, 500 cells/cluster, 10,000 reads per cell, and no noise, on chromosome 19. (A): simulated counts from the bulk and pseudo bulk samples (and clusters). (B): distribution of ARI values and computing time across 100 simulated data sets for different methods and clustering algorithms. (C): UMAP projections of a simulated data for different methods. Colors indicate true single-cell cluster memberships.

Table 4.6: Performance of scATAC-seq methods on simulated scATAC-seq data under different sequencing depths (5,000 and 25,000) and different noise levels (0% and 25%) for 3 clusters, 500 cells/cluster, and no rare cell sub populations. Average (standard deviation) ARI, AMI, and computing time are shown.

Depth	Noise	Method	ARI	AMI	Time
5,000	0%	Cusanovich2018	0.88 (0.18)	0.88 (0.12)	0.04 (0.00)
		Scasat	0.86 (0.02)	0.84 (0.02)	0.05 (0.01)
		SnapATAC	0.14 (0.23)	0.15 (0.24)	0.13 (0.01)
		cisTopic	0.64 (0.14)	0.68 (0.08)	0.59 (0.05)
	25%	Cusanovich2018	0.77 (0.17)	0.78 (0.11)	0.04 (0.00)
		Scasat	0.68 (0.04)	0.69 (0.03)	0.05 (0.01)
		SnapATAC	0.11 (0.18)	0.12 (0.20)	0.13 (0.01)
		cisTopic	0.53 (0.06)	0.60 (0.03)	0.47 (0.04)
25,000	0%	Cusanovich2018	0.89 (0.22)	0.92 (0.17)	0.06 (0.00)
		Scasat	1.00 (0.00)	1.00 (0.00)	0.08 (0.01)
		SnapATAC	0.57 (0.40)	0.55 (0.38)	0.17 (0.01)
		cisTopic	0.91 (0.21)	0.93 (0.16)	2.64 (0.20)
	25%	Cusanovich2018	0.87 (0.24)	0.90 (0.18)	0.08 (0.03)
		Scasat	1.00 (0.00)	1.00 (0.00)	0.10 (0.05)
		SnapATAC	0.46 (0.42)	0.44 (0.41)	0.23 (0.10)
		cisTopic	0.89 (0.22)	0.91 (0.17)	3.27 (1.51)

scChIP-seq data regardless of the number of clusters and number of cells/cluster. The sequencing depth and cluster-to-cluster difference levels played a major role in the performance of the methods. Specifically, methods had a satisfactory performance in simulated scChIP-seq only under high sequencing depth levels (25,000 reads/cell) and high cluster-to-cluster difference levels (5% of differential regions; Table 4.7). However, these scenarios are somewhat unrealistic for scChIP-seq real datasets, in which the median depth was 3,651 and the average percentage of differential genomic regions appeared to be around 1% (Grosselin et al., 2019).

By reducing the signal-to-noise ratio of simulated scATAC-seq data, methods had a slight decrease in performance, an issue that was ameliorated in scenarios with higher sequencing depths per cell. Depth levels of 25,000 reads per cell is not unrealistic for scATAC-seq experiments (Chen et al., 2018). However, the scChIP-seq technology is highly influenced by background noise due to non-specific antibody pull-down (Clark et al., 2016) and current studies present assays with a moderate to low sequencing depth per cell (Rotem et al., 2015; Grosselin et al., 2019). Therefore, these findings support the development of robust algorithms tailored for the sparsity of the data, as well as the local dependency, often observed in scChIP-seq data.

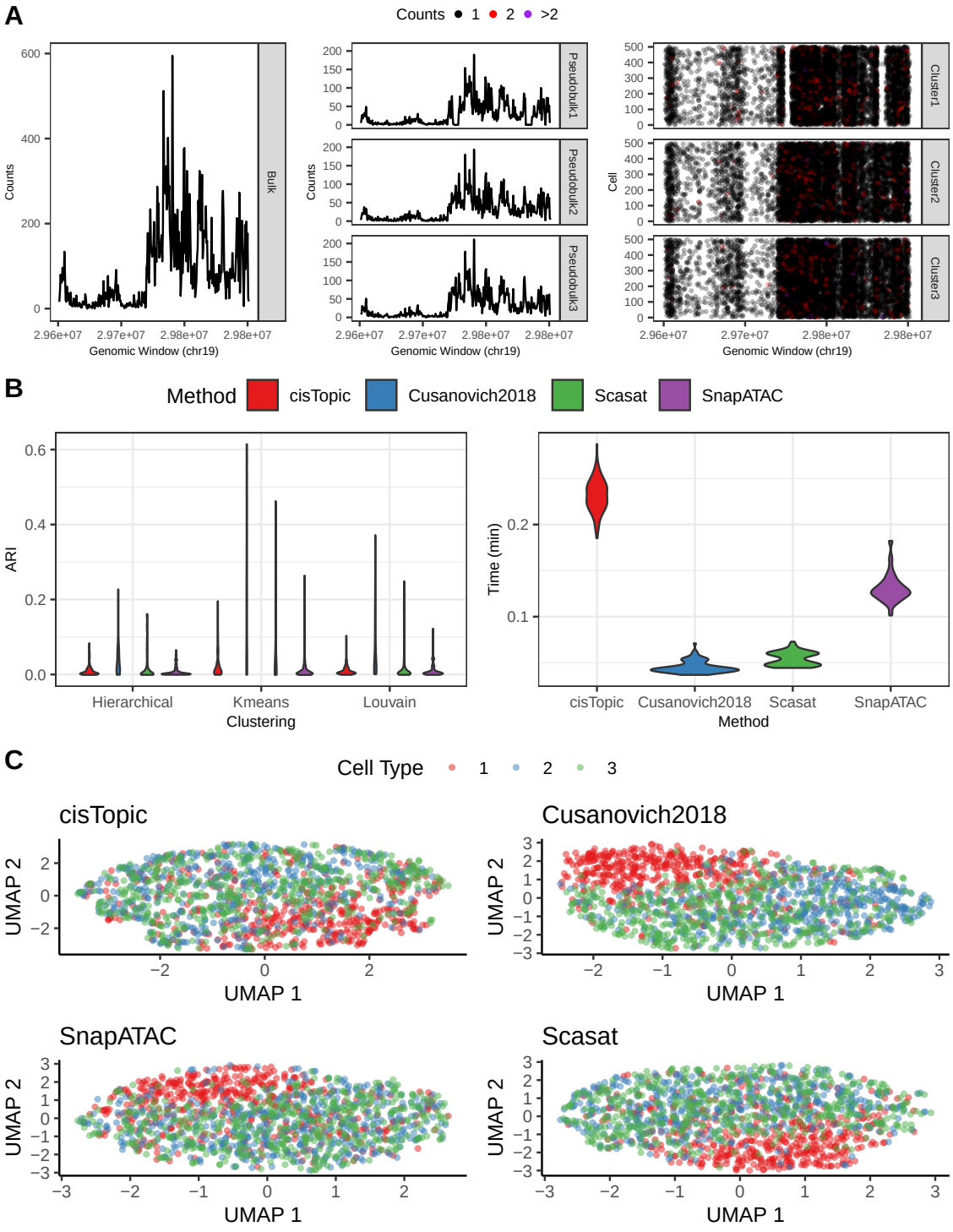

Figure 4.19: Results from simulated scChIP-seq data for the scenario with 3 clusters, 500 cells/cluster, 10,000 reads per cell, and no noise, on chromosome 19. (A): simulated counts from the bulk and pseudo bulk samples (and clusters). (B): distribution of ARI values and computing time across 100 simulated data sets for different methods and clustering algorithms. (C): UMAP projections of a simulated data for different methods. Colors indicate true single-cell cluster memberships.

Table 4.7: Performance of scATAC-seq methods on simulated scChIP-seq data under different sequencing depths (5,000 and 25,000) and cluster-to-cluster difference levels (1% and 5%). The scenario with 5,000 reads/cell and 1% difference level better approximates real data (Grosselin et al., 2019). Average (standard deviation) ARI, AMI, and computing time are shown.

Depth	Dissimilarity	Method	ARI	AMI	Time
5,000	1%	Cusanovich2018	0.04 (0.06)	0.04 (0.07)	0.05 (0.02)
		Scasat	0.00 (0.01)	0.00 (0.01)	0.06 (0.02)
		SnapATAC	0.00 (0.01)	0.00 (0.01)	0.15 (0.05)
		cisTopic	0.01 (0.01)	0.01 (0.01)	0.16 (0.06)
	5%	Cusanovich2018	0.69 (0.11)	0.64 (0.10)	0.04 (0.00)
		Scasat	0.38 (0.23)	0.36 (0.21)	0.05 (0.01)
		SnapATAC	0.10 (0.20)	0.10 (0.18)	0.13 (0.01)
		cisTopic	0.31 (0.12)	0.29 (0.11)	0.14 (0.01)
25,000	1%	Cusanovich2018	0.54 (0.21)	0.51 (0.19)	0.04 (0.00)
		Scasat	0.39 (0.25)	0.37 (0.24)	0.05 (0.01)
		SnapATAC	0.14 (0.23)	0.14 (0.22)	0.16 (0.05)
		cisTopic	0.13 (0.10)	0.12 (0.09)	0.47 (0.05)
	5%	Cusanovich2018	0.97 (0.13)	0.97 (0.10)	0.04 (0.00)
		Scasat	1.00 (0.00)	1.00 (0.01)	0.05 (0.01)
		SnapATAC	0.80 (0.38)	0.80 (0.38)	0.13 (0.01)
		cisTopic	0.94 (0.13)	0.93 (0.10)	0.47 (0.04)

4.4.2 Improving Single-cell Clustering With Differentially Enriched Candidate Regions

Next, we evaluated the improvement in performance of current scATAC-seq methods by utilizing the model presented in Section 4.3.1 to define differentially enriched regions as candidate peaks for single-cell clustering. To this end, we simulated data as follows. Data from three hundred cells from three distinct subpopulations were simulated according to three independent two-state Markov chains with transition probabilities $\gamma_{11} = 0.995$ (background-to-background transition probability) and $\gamma_{22} = 0.99$ (enrichment-to-enrichment transition probability). Sparse binary counts were simulated for each cell such that the mean (standard deviation) number of windows with at least one sequencing read was 250.16 (16.35), out of a total of 10,000 genomic windows. Following the simulation setup presented in Section 4.4.1, counts were simulated according to a Bernoulli distribution and reads were allocated to enrichment and background regions with a 70:30 proportion.

A total of 100 simulated datasets were generated. For each dataset, scATAC-seq methods were applied to cluster cells into subpopulations. Candidate peaks were defined in two distinct manners. First, peaks were called on the aggregated single-cell data (as in Section 4.4.1). Second, differentially enriched regions were

detected by making use of the statistical model presented in Section 4.3.1. We compared the performance of current methods for the two peak calling strategies regarding the cluster assignments with the true simulated cluster labels via the ARI and AMI metrics. We show in Section 4.4.2.1 that the presented model offers substantial benefits to the selection of candidate peaks from scChIP-seq data and improves the performance of current methods.

4.4.2.1 Simulation Results

Table 4.8 shows the results from the simulation study with candidate peaks called on pooled data and those called as differentially enriched regions. We show results for scATAC-seq methods Cusanovich2018, Scasat, SnapATAC, and cisTopic. Single-cells were clustered with three distinct clustering algorithm, namely hierarchical clustering, Kmeans, and Louvain. Overall we observed a significant improvement in performance in scATAC-seq methods by utilizing differentially enriched regions as candidate peaks. All methods exhibited an improvement in performance when differentially enriched regions were utilized as candidate peaks. From all methods, cisTopic exhibited the largest gains in performance by comparing the ARI and AMI metrics from pooled and differential peaks. The variability of the ARI and AMI metrics across a hundred simulated datasets also exhibited a substantial decrease for all methods but SnapATAC by utilizing the presented strategy for candidate peak selection. This fact indicates that most methods tend to be more precise in classifying single-cell subpopulations by utilizing differentially enriched regions as candidate peaks. In agreement to the simulation study presented in Section 4.4.1, SnapATAC exhibited the lowest performance among all methods

Figure 4.20 illustrate some of the results from the simulation study. As shown in panel D, the UMAP representation of the feature matrix output by each scATAC-seq method shows a better separation of single-cell subpopulations for nearly all assessed methods. Such a fact further illustrates the benefit of the presented model for single-cell clustering from scChIP-seq data. Yet, SnapATAC and its implemented Jaccard similarity feature matrix did not lead to UMAP representations with sufficiently distinguishable patterns of single-cell clusters. For this particular method, the selection of differentially enriched regions did not appear to help much with the clustering of similar cells.

4.5 Discussion

Here, we presented a comparative study of some of the current algorithms for the analysis of epigenomic data on both simulated and real data from single-cell ChIP-seq experiments. We proposed the use of a hidden Markov model for the selection of candidate regions exhibiting differential enrichment from sparse

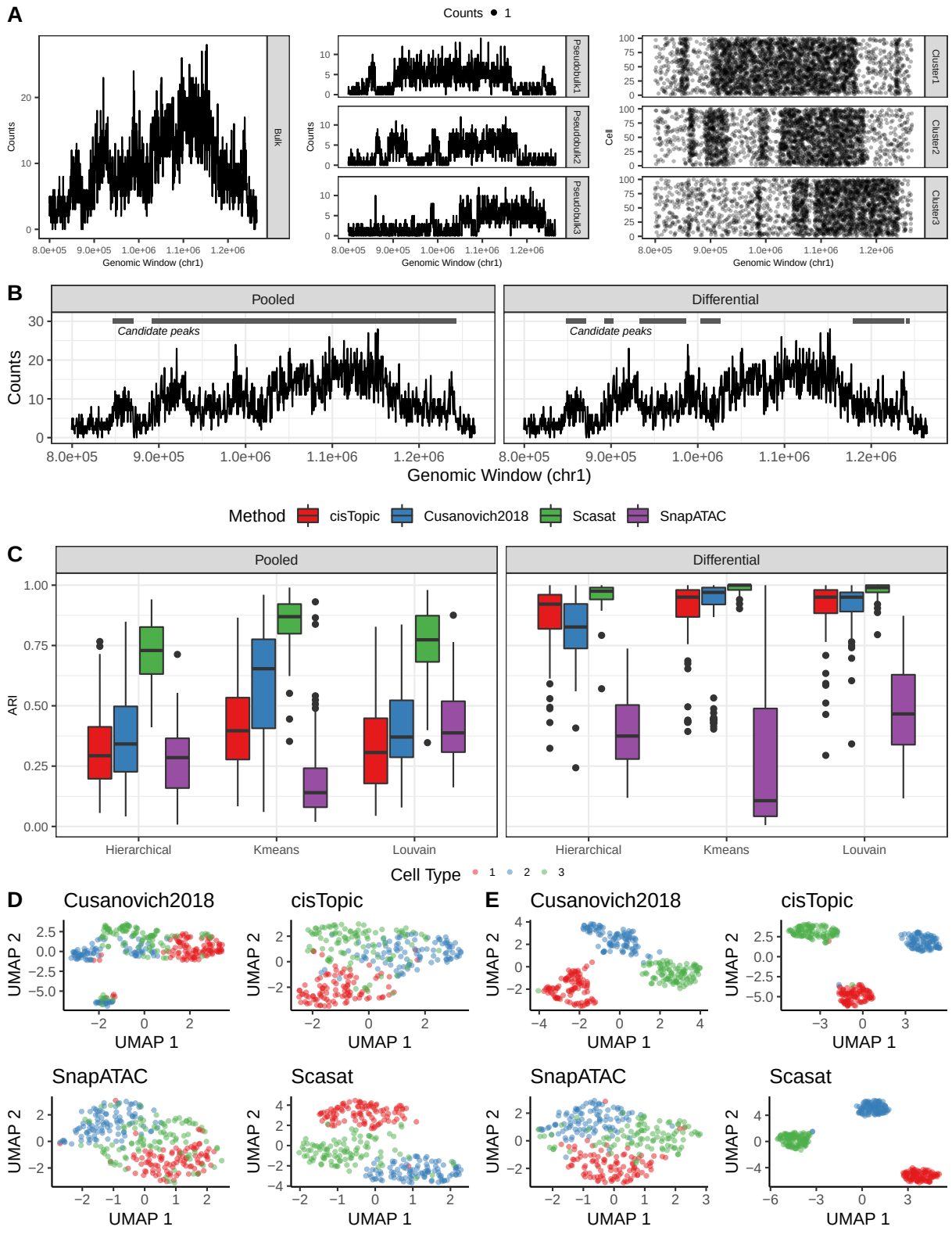

Figure 4.20: Performance of scATAC-seq methods on simulated scChIP-seq data with candidate peaks called either on bulk data (Pooled) or on single-cell data with 3-state HMM (Differential).

Table 4.8: Performance of scATAC-seq methods on simulated scChIP-seq data with candidate peaks called either on bulk data (Pooled) or on single-cell data with 3-state HMM (Differential).

Clustering Algorithm	Method	Pooled		Differential	
		ARI	AMI	ARI	AMI
Hierarchical	Cusanovich2018	0.36 (0.20)	0.36 (0.18)	0.81 (0.13)	0.78 (0.13)
	Scasat	0.72 (0.13)	0.68 (0.12)	0.96 (0.05)	0.94 (0.06)
	SnapATAC	0.27 (0.15)	0.28 (0.12)	0.39 (0.15)	0.38 (0.13)
	cisTopic	0.32 (0.16)	0.31 (0.15)	0.87 (0.13)	0.85 (0.13)
Kmeans	Cusanovich2018	0.60 (0.21)	0.57 (0.19)	0.88 (0.20)	0.88 (0.17)
	Scasat	0.85 (0.11)	0.81 (0.11)	0.99 (0.02)	0.98 (0.03)
	SnapATAC	0.20 (0.18)	0.21 (0.18)	0.29 (0.34)	0.30 (0.34)
	cisTopic	0.42 (0.19)	0.40 (0.17)	0.90 (0.14)	0.88 (0.12)
Louvain	Cusanovich2018	0.41 (0.16)	0.38 (0.14)	0.91 (0.09)	0.89 (0.11)
	Scasat	0.75 (0.15)	0.70 (0.15)	0.98 (0.03)	0.97 (0.04)
	SnapATAC	0.41 (0.14)	0.38 (0.12)	0.48 (0.18)	0.43 (0.15)
	cisTopic	0.34 (0.19)	0.31 (0.17)	0.90 (0.12)	0.88 (0.13)

single-cell data. Current methods developed for the analysis of scATAC-seq data often rely on the set of candidate peaks called on the bulk data for single-cell clustering. This approach may be subject to choice of the peak calling algorithm and its parametrization, and differences in sequencing depth across cells may mask the experimental signal from under-sequenced cells when calling peaks on the bulk data. Moreover, this approach may lead to suboptimal performance as enrichment regions from scChIP-seq data are often found to be extremely sparse and to expand through large genomic domains. In a simulation study, the presented model led to significant benefits in current scATAC-seq methods.

In addition, we proposed the use of a statistical method to cluster single-cells from heterogeneous samples into groups of cells sharing similar structural patterns of read count distribution across the genome. The proposed method allows the analysis of single-cell data on high genomic resolutions, without the need of relying on a set of candidate peaks for the characterization of single-cell subpopulations. The proposed model does not rely on a set of candidate peaks and utilizes the available data from all genomic positions, while being able to account for different sequencing depths in the datasets through model offsets. Lastly, we presented in this project an initialization scheme for the presented EM algorithm that aids with the determination of the number of single-cell subpopulations in the data, a task that is often necessary in the analysis of single-cell epigenomic data with current methods.

One of the limitations of the presented MHMM is that it lacks the ability to detect longitudinal differences in the read count distribution across the genome between cells. For instance, multiple realizations of the same process (or hidden Markov chain) are deemed similar and clustered together by the algorithm, despite differences with respect to the location of enrichment regions in the data. As a future research project, we will further consider alternative models and parametrization, such as the hierarchical Dirichlet process hidden Markov model (HDP-HMM), to properly account for such realizations of the same process.

CHAPTER 5: CONCLUSION AND FUTURE RESEARCH

In this dissertation, we introduce three statistical methods for the analysis of epigenomic data that are tailored to address some of the challenges faced by current approaches in contexts where the data is sparse and counts exhibit a local dependency across the genome.

In Chapter 2, we presented a multi-sample zero-inflated mixed-effects hidden Markov models (HMMs) to account for the excess of observed zeros in regions without epigenomic activity, the latent sample-specific differences, and the local dependency of sequencing read counts. We applied the presented methods in an extensive simulation study and in data sets from the ENCODE and Roadmap Epigenomics Projects and showed superior performance than current methods in data sets from histone modifications characterized by broad regions of enrichment, i.e., regions with more sequencing reads than one would expect in background regions.

In Chapter 3, we presented a statistical model to detect and classify differential epigenomic activity across conditions in multi-sample multi-condition designs. Our model is flexible for the analysis of broad (e.g., histone modifications H3K27me3 and H3K36me3) and short data sets (e.g., transcription factor CTCF and ATAC-seq data). We utilized an efficient implementation of the EM algorithm that allows the genome-wide analysis of multiple ChIP-seq data sets in a computing time that is comparable to some of the fastest algorithms available. Although the presented model performed well in both analyzed simulated and real data sets, more simulation studies are needed to better understand the performance of the model selection approach using the BIC for HMMs. In the genomic segmentation analysis presented in Section 3.5.3, there was a prior biological knowledge regarding the roles of the analyzed epigenomic marks (H3K36me3, H3K27me3, and EZH2) and we knew in advance that the optimal number of mixture components in the HMM differential state would be 2 (enrichment for H3K36me3 alone, or co-occurrence of H3K27me3 and EZH2). The number of mixture components from the best model chosen via BIC agreed with the expected optimal number of components. However, for analyses where there is no prior information regarding the activity of the analyzed marks, it will be useful to understand whether the use of th BIC for model selection is in fact appropriate.

In addition, the current implementation of our differential peak caller can be optimized for scenarios where one of the analyzed epigenomic marks (in a genomic segmentation analysis) is highly different than the others. In Figure 3.14, for instance, our model was robust to detect and classify differential patterns between CTCF and H3K36me3 in the Helas3 cell line with the non-linear normalization for sequencing depth via model offsets. However, as Figure 3.14 shows, it is clear that the mean of the read count distribution from enrichment regions of CTCF after normalization is not the same as the mean of the respective regions from H3K36me3. Although our model was robust to these differences, it is important for our differential peak caller to incorporate condition-specific parameters in the GLM-based framework from a methodological perspective. Such an implementation is left as a future research project.

In Chapter 4, we presented a comparative analysis of scChIP-seq data using current scATAC-seq algorithms. We showed that current approaches can have difficulties to deal with the sparsity of the data, which exhibits a local dependency of counts that is not commonly found in scATAC-seq experiments. One of the possible explanations for the suboptimal performance of such methods is that they rely on a 2-step approach by first calling peaks on the bulk data and then performing any sort of dimension reduction technique (e.g. PCA, SVD, and LDA) followed by the application of a clustering algorithm. We argue that, because the data is highly sparse, considering candidate peaks from the bulk data might not be ideal for scChIP-seq data for two reasons. First, it is a difficult task to distinguish differential regions of enrichment from regions where there is background signal due to sparsity and noisy aspects of the data. Because candidate regions of enrichment can be excessively broad (as they are obtained from the bulk data), methods might have not enough power to distinguish subpopulations of cells (as they computations are done on the single-cell level with sparse counts). Second, consensus regions of enrichment are inevitably considered as candidate regions by using the 2-step approach. Such regions are not informative to distinguish subpopulations of cells and, therefore, should be ideally removed from the analysis by these algorithms. The consideration of such regions in addition to regions that are truly differential may actually worsen the performance of current methods since it is already a difficult task to distinguish what is a true signal and what is noise in scChIP-seq data. To address these issues, we introduced a statistical model to clustering single-cells from heterogeneous scChIP-seq data sets in high resolutions with a initialization scheme that allows the estimation of the number of sub populations present in the data. As a future research project, we plan on integrating in this model the ability to distinguish subpopulations of cells by consider not their structural differences regarding their likelihood for a finite set of HMMs but, in fact, their longitudinal differences regarding the presence/absence

of enrichment of reads along the genome. We plan also to explore alternative methodologies for unstructured data that are based on neural networks and deep learning algorithms, a set of tools that has been proven extremely powerful in other areas such as transcriptomics and proteomics.

In summary, the methods developed in this dissertation aim to address current challenges in the analysis of bulk and single-cell ChIP-seq data and are relevant for biomedical researchers interested in the field of epigenomics.

APPENDIX A: TECHNICAL DETAILS FOR CHAPTER 2

Data

The data utilized in Chapter 1 pertaining to the ENCODE Consortium and Roadmap Project are listed in Table A.9.

Table A.9: GEO sample accession codes of the analyzed data from the ENCODE Consortium and Roadmap Project in Chapter 1.

Cell Line	H3K27me3	H3K36me3	RNA-seq
H1hesc	GSM733748	GSM733725	GSM758566
HelaS3	GSM733696	GSM733711	GSM765402
Hepg2	GSM733754	GSM733685	GSM758575
Huvec	GSM733688	GSM733757	GSM758563
Nhek	GSM733701	GSM733726	GSM765401
CD4 Memory Primary	GSM772998	GSM772964	GSM669618
CD4 Naïve Primary	GSM772947	GSM772932	GSM669617
CD8 Naïve Primary	GSM772871	GSM772872	GSM669619
CD34 Mobilized Primary	GSM669945	GSM621459	GSM909310

The following steps were conducted to process the data. First, we removed PCR duplicates from the BAM files using SAMTools (Li et al., 2009) and converted the resulting indexed and sorted files to BED format using BEDTools (Quinlan and Hall, 2010), as JAMM only accepts such a format as input. Then, the fragment length of each ChIP-seq experiment was estimated using MACS2 and its sub-command *predictd*. Finally, using the estimated fragment length, read counts from all cell lines were tabulated for their ChIP replicates and input control experiments using fixed-step and non overlapping windows of size 250bp, 500bp, 750bp, and 1000bp through the R package *bamsignals* (Mammana and Helmuth, 2016). For all non-overlapping window-based methods (JAMM, MOSAiCS, RSEG, Zerone, and ZIMHMM), we assessed their performance with different window sizes. See Section D.3 in Baldoni et al. 2019b for a discussion about results with different window sizes.

All the methods considered in the data applications outputted a set of genomic regions of enrichment that were used for benchmark purposes. MACS2 and JAMM output a list of peak regions in BED6+3 (*broadPeak*) and BED6+4 formats (*narrowPeak*), respectively. Peak calls from BCP, CCAT, MOSAiCS, and RSEG were obtained from the output BED files. As recommended by Ibrahim et al. (2014), we used the set of filtered peaks from JAMM to avoid the inclusion of artifact peaks (single basepair peaks or peaks with very few reads). For Zerone, we used the outputted Viterbi sequence of predicted states to merge adjacent windows

and form regions of enrichment. For comparative purposes, enrichment regions detected by our method were defined in a similar fashion according to the Viterbi sequence of predicted states. We observed a similar performance by thresholding posterior probabilities using an FDR level of 0.05. For a comparison between the Viterbi and the FDR thresholding approach, see Table 2.3.

The following parametrization was used when calling peaks from the benchmarked methods. For BPC, *peakranger bcp –format bam –report –verbose –geneannotfile 'gene' –data 'sample' –control 'control' –output 'output'*. For CCAT, *peakranger ccat –format bam –report –verbose –geneannotfile 'gene' –data 'sample' —control 'control' –output 'output'*. For JAMM, *JAMM.sh -m normal -r region -w 1 -b 'binsize' -g 'genome' -s 'sample' -c 'control' -o 'output'*. For MACS2, *macs2 callpeak –broad -g hs –broad-cutoff 0.1 -f BAM*. For MOSAiCS (R package), we used the following sequence of commands *constructBins(fileFormat = 'bam'), readBins(), mosaicsFit(analysisType="IO", bgEst="rMOM"), mosaicsFitHMM(), mosaicsPeakHMM(), extractReads(chipFileFormat='bam',controlFileFormat='bam'), findSummit(), adjustBoundary(), filterPeak()*, with read counts computed in fixed windows of size of 250bp, 500bp, 750bp, and 1000bp. For RSEG, *rseg-diff -verbose -mode 2 -out 'output' -chrom 'chromosome -deadzones 'deadzones' 'sample' 'control'* with read counts computed in fixed windows of size of 250bp, 500bp, 750bp, and 1000bp. For Zerone (R package), *zerone(.,returnall=T)*, with read counts computed in fixed windows of size of 250bp, 500bp, 750bp, and 1000bp.

Software

Regarding the implemented software, the ZIMHMM (Zero Inflated Mixed effects Hidden Mark Model) was implemented in a R package that is available on https://github.com/plbaldoni/ZIMHMM. ZIMHMM is a package with a peak caller to detect broad enrichment regions from multiple ChIP-seq experimental replicates. The main function of the package is *ZIMHMM*(), which models the zero-inflation of background counts, accounts for replicate-specific differences via a mixed effects model, and ensures that broad regions of enrichment are detected by fitting a hidden Markov model. The package also contains *ZIHMM*(), a fixed effects version of the peak caller.

The package allows the user to specify a set of parameters that control the Expectation-Maximization (EM) algorithm. These parameters include, for instance, the convergence (and termination) criteria of the algorithm and the threshold value for the rejection controlled EM algorithm. These parameters can be defined by the function *controlPeaks*(). Other auxiliary functions include *plotPeaks*(), which plots the

read counts from ChIP-seq replicates and the called peaks. Please refer to the package documentation (e.g. *?ZIMHMM::ZIMHMM*) for additional details.

The EM Algorithm

The presented EM algorithm can be summarized as follows:

1. Initialize $\boldsymbol{\pi}^{(0)}, \boldsymbol{\gamma}^{(0)}, \boldsymbol{\lambda}^{(0)}, \boldsymbol{\beta}_k^{(0)}, \boldsymbol{\phi}_k^{(0)}$ and $\sigma^{(0)}$ for $k \in \{1, 2\}$, such that $\sum_{k=1}^{2} \pi_k^{(0)} = 1$ and $\sum_{l=1}^{2} \gamma_{kl} = 1$.

2. E step ($s \geq 1$),

   (a) Calculate $\widehat{\mathbf{u}} = \underset{\mathbf{u} \in \mathbb{R}^N}{\arg\max} \left\{ \log \left(\mathbf{A}^{(s-1)}(\mathbf{u}) \prod_{j=2}^{M} \mathbf{C}_j^{(s-1)}(\mathbf{u}) \mathbb{I} \right) + \log(f(\mathbf{u})) \right\}$, as detailed in Appendix A

   (b) Calculate $P\left(Z_j = k | \mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Psi}^{(\mathbf{s-1})}\right)$ and $P\left(Z_{j-1} = l, Z_j = k | \mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Psi}^{(\mathbf{s-1})}\right)$ for all l and k in $\{1, 2\}$ and $j = 1, \ldots, N$ via Forward-Backward algorithm as detailed in Appendix A

3. M step ($s \geq 1$),

   (a) Maximize Equation 2.5 with respect to the initial and transition probabilities to obtain for all l and k in $\{1, 2\}$

$$\pi_k^{(s)} = P\left(Z_1 = k | \mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Psi}^{(\mathbf{s-1})}\right)$$

$$\gamma_{lk}^{(s)} = \frac{\sum_{j=2}^{M} P\left(Z_{j-1} = l, Z_j = k | \mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Psi}^{(\mathbf{s-1})}\right)}{\sum_{j=2}^{M} \sum_{r=1}^{2} P\left(Z_{j-1} = l, Z_j = r | \mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Psi}^{(\mathbf{s-1})}\right)}$$

   (b) Maximize Equation 2.5 with respect to $\boldsymbol{\beta}_k$ and $\boldsymbol{\phi}_k$ to obtain $\boldsymbol{\beta}_k^{(s)}$ and $\boldsymbol{\phi}_k^{(s)}$ (see Appendix A for partial derivatives),

   (c) Conditionally upon $\boldsymbol{\beta}_k^{(s)}$ and $\boldsymbol{\phi}_k^{(s)}$, maximize Equation 2.5 with respect to σ to obtain $\sigma^{(s)}$ (see Appendix A for partial derivatives),

   (d) Iterate between (b) and (c) until convergence.

4. Iterate between 2. and 3. until convergence.

Marginal Moments of the Mixed Effects HMM

The marginal moments of the random effects model are presented next. Let $\mathbf{x}_{ij}\boldsymbol{\beta}_k$ denote the linear predictor pertaining to the fixed effects of the mean model associated with the k^{th} emission distribution. One can show using basic properties of conditional expectations that

$$\begin{aligned}
\mathrm{E}(Y_{ij}) &= \exp(\sigma^2/2)\sum_k \pi_k \exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k), \\
\mathrm{Var}(Y_{ij}) &= \exp(\sigma^2/2)\sum_k \pi_k \exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k) + \exp(2\sigma^2)\sum_k \pi_k \exp(2\mathbf{x}_{ij}\boldsymbol{\beta}_k)\,(1/\phi_k + 1) - \\
&\quad - \exp(\sigma^2)\left\{\sum_k \pi_k \exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k)\right\}^2, \\
\mathrm{Cov}(Y_{is}, Y_{it}) &= \exp(2\sigma^2)\sum_k\sum_l P(Z_t = l|Z_s = k)\pi_k \exp(\mathbf{x}_{is}\boldsymbol{\beta}_k + \mathbf{x}_{it}\boldsymbol{\beta}_l) - \\
&\quad - \exp(\sigma^2)\left\{\sum_k \pi_k \exp(\mathbf{x}_{is}\boldsymbol{\beta}_k)\right\}\left\{\sum_k \pi_k \exp(\mathbf{x}_{it}\boldsymbol{\beta}_k)\right\}, \forall s < t.
\end{aligned}$$

Using a similar notation, the marginal moments of the fixed effects HMM with w_i as offsets are

$$\begin{aligned}
\mathrm{E}(Y_{ij}) &= w_i \sum_k \pi_k \exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k), \\
\mathrm{Var}(Y_{ij}) &= w_i \sum_k \pi_k \exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k) + w_i^2 \sum_k \pi_k \exp(2\mathbf{x}_{ij}\boldsymbol{\beta}_k)\,[1/\phi_k + 1], \\
\mathrm{Cov}(Y_{is}, Y_{it}) &= w_i^2 \sum_k\sum_l P(Z_t = l|Z_s = k)\pi_k \exp(\mathbf{x}_{is}\boldsymbol{\beta}_k + \mathbf{x}_{it}\boldsymbol{\beta}_l) - \\
&\quad - w_i^2 \left\{\sum_k \pi_k \exp(\mathbf{x}_{is}\boldsymbol{\beta}_k)\right\}\left\{\sum_k \pi_k \exp(\mathbf{x}_{it}\boldsymbol{\beta}_k)\right\}, \forall s < t.
\end{aligned}$$

Assuming convergence of $P(Z_t = l|Z_s = k) \to \pi_l$ for a fixed s and $t \to \infty$, the dominated convergence theorem holds under mild conditions and the asymptotic marginal covariance under the random effects model is non-negative and equal to zero if and only if the distribution of random effects is degenerate (Altman, 2007). For the fixed effects model, such a covariance converges to zero for a fixed s and $t \to \infty$.

Technical Derivations of the EM Algorithm

The algorithm for its fixed-effects version can be run in a similar fashion in the sense that the Laplace approximation used during the E-step of the algorithm is not necessary. All the remaining parts of the algorithm are similar. First, note that Q function presented in Section 2.4 is a N-dimensional integral of a

product between (1) the expectation of the complete data log likelihood function taken with respect to the distribution of $\mathbf{Z}$ conditional upon the observed data, the latent random effects $\mathbf{u}$ and the estimated parameters of the s^{th} EM iteration $\mathbf{\Psi}^{(s)}$, and (2) the distribution of random effects $\mathbf{U}$ conditional upon the observed data and the estimated parameters of the s^{th} EM iteration $\mathbf{\Psi}^{(s)}$. Also, note that the inner expectation of the Q function can be expressed as

$$
\begin{aligned}
&E\left(\sum_{k=1}^{2} I(Z_1=k)\log(\pi_k)+\sum_{j=1}^{M}\sum_{k=1}^{2}\sum_{i=1}^{N} I(Z_j=k)\log\left(f_k(y_{ij}|u_i,r_{ij},x_{ij};\boldsymbol{\psi}_k,\sigma)\right)+\right.\\
&\left.+\sum_{j=2}^{M}\sum_{k=1}^{2}\sum_{l=1}^{2} I(Z_{j-1}=l,Z_j=k)\log(\gamma_{lk})+\log(f(\mathbf{u}))|\mathbf{y},\mathbf{u},\mathbf{r},\mathbf{x};\mathbf{\Psi}^{(s)}\right)=\\
&=\sum_{k=1}^{2} P\left(Z_1=k|\mathbf{y},\mathbf{u},\mathbf{r},\mathbf{x};\mathbf{\Psi}^{(s)}\right)\log(\pi_k)+\\
&+\sum_{j=1}^{M}\sum_{k=1}^{2}\sum_{i=1}^{N} P\left(Z_j=k|\mathbf{y},\mathbf{u},\mathbf{r},\mathbf{x};\mathbf{\Psi}^{(s)}\right) f_k(y_{ij}|u_i,r_{ij},x_{ij};\boldsymbol{\psi}_k,\sigma)+\\
&+\sum_{j=2}^{M}\sum_{k=1}^{2}\sum_{l=1}^{2} P\left(Z_{j-1}=l,Z_j=k|\mathbf{y},\mathbf{u},\mathbf{r},\mathbf{x};\mathbf{\Psi}^{(s)}\right)\log(\gamma_{lk})+\log(f(\mathbf{u})),
\end{aligned}
\tag{B.18}
$$

where $P\left(Z_j=k|\mathbf{y},\mathbf{u},\mathbf{r},\mathbf{x};\mathbf{\Psi}^{(s)}\right)$ and $P\left(Z_{j-1}=l,Z_j=k|\mathbf{y},\mathbf{u},\mathbf{r},\mathbf{x};\mathbf{\Psi}^{(s)}\right)$ can be calculated by a standard Forward-Backward algorithm (see Appendix A). Conversely, the distribution of the random effects $\mathbf{U}$ conditional on the observed data and the estimated parameters of the s^{th} EM iteration $\mathbf{\Psi}^{(s)}$ can be re-expressed as

$$
\begin{aligned}
f\left(\mathbf{u}|\mathbf{y},\mathbf{r},\mathbf{x};\mathbf{\Psi}^{(s)}\right) &= \frac{f\left(\mathbf{u},\mathbf{y}|\mathbf{r},\mathbf{x};\mathbf{\Psi}^{(s)}\right)}{f\left(\mathbf{y}|\mathbf{r},\mathbf{x};\mathbf{\Psi}^{(s)}\right)}\\
&= \frac{\sum_{\mathbf{z}\in\mathcal{Z}} f\left(\mathbf{y},\mathbf{z}|\mathbf{u},\mathbf{r},\mathbf{x};\mathbf{\Psi}^{(s)}\right) f(\mathbf{u})}{\int_{\mathbf{u}\in\mathbb{R}^N}\sum_{\mathbf{z}\in\mathcal{Z}} f\left(\mathbf{y},\mathbf{z}|\mathbf{u},\mathbf{r},\mathbf{x};\mathbf{\Psi}^{(s)}\right) f(\mathbf{u})d\mathbf{u}}\\
&= \frac{\left(\mathbf{A}^{(s)}\prod_{j=2}^{M}\mathbf{C}_j^{(s)}\mathbb{I}\right) f(\mathbf{u})}{\int_{\mathbf{u}\in\mathbb{R}^N}\left(\mathbf{A}^{(s)}\prod_{j=2}^{M}\mathbf{C}_j^{(s)}\mathbb{I}\right) f(\mathbf{u})d\mathbf{u}},
\end{aligned}
\tag{B.19}
$$

where $\mathbf{A}^{(s)}=\left(A_1^{(s)},A_2^{(s)}\right)$, $A_k^{(s)}=\pi_k^{(s)}f_k\left(\mathbf{y}_{\cdot 1}|\mathbf{u},\mathbf{r}_{\cdot 1},\mathbf{x}_{\cdot 1};\boldsymbol{\psi}_k^{(s)},\sigma^{(s)}\right)$, $\mathbf{C}_j^{(s)}$ is a 2×2 matrix with elements $C_{j,lk}^{(s)}=\gamma_{lk}^{(s)}f_k\left(\mathbf{y}_{\cdot j}|\mathbf{u},\mathbf{r}_{\cdot j},\mathbf{x}_{\cdot j};\boldsymbol{\psi}_k^{(s)},\sigma^{(s)}\right)$ for all l and k in $\{1,2\}$, and $\mathbb{I}$ is a 2-dimensional vector of ones.

First, note that in Equation B.19, the denominator is the marginal distribution of $\mathbf{Y}$ and does not depend on $\mathbf{U}$. Therefore, we can incorporate this quantity into the function h since it will not affect the calculation of $\mathbf{J}$. Thus,

$$\begin{aligned} Q\left(\mathbf{\Psi}|\mathbf{\Psi}^{(s)}\right) &= \int_{\mathbf{u}\in\mathbb{R}^N} h\left(\mathbf{u};\mathbf{\Psi},\mathbf{\Psi}^{(s)}\right) \times \\ &\quad \times \exp\left(\log\left(\mathbf{A}^{(s)}\prod_{j=2}^{M}\mathbf{C}_j^{(s)}\mathbb{I}\right) + \log(f(\mathbf{u})) - \log\left(f\left(\mathbf{y};\mathbf{\Psi}^{(s)}\right)\right)\right) d\mathbf{u} \\ &= \int_{\mathbf{u}\in\mathbb{R}^N} \frac{h\left(\mathbf{u};\mathbf{\Psi},\mathbf{\Psi}^{(s)}\right)}{f\left(\mathbf{y};\mathbf{\Psi}^{(s)}\right)} \exp\left(\log\left(\mathbf{A}^{(s)}\prod_{j=2}^{M}\mathbf{C}_j^{(s)}\mathbb{I}\right) + \log(f(\mathbf{u}))\right) d\mathbf{u} \\ &= \int_{\mathbf{u}\in\mathbb{R}^N} h^*\left(\mathbf{u};\mathbf{\Psi},\mathbf{\Psi}^{(s)}\right) \times \exp\left(g^*\left(\mathbf{u};\mathbf{\Psi}^{(s)}\right)\right) d\mathbf{u}, \end{aligned} \tag{B.20}$$

where $g^*\left(\mathbf{u};\mathbf{\Psi}^{(s)}\right) = \log\left(\mathbf{A}^{(s)}\prod_{j=2}^{M}\mathbf{C}_j^{(s)}\mathbb{I}\right) + \log(f(\mathbf{u}))$. Denote $\widehat{\mathbf{u}}$ the value of $\mathbf{u}$ such that $\mathbf{J}_{g^*}|_{\mathbf{u}=\widehat{\mathbf{u}}} = \mathbf{0}$. Using a second order Taylor's series expansion of $g^*\left(\mathbf{u};\mathbf{\Psi}^{(s)}\right)$ around $\widehat{\mathbf{u}}$,

$$g^*\left(\mathbf{u};\mathbf{\Psi}^{(s)}\right) \approx g^*\left(\widehat{\mathbf{u}};\mathbf{\Psi}^{(s)}\right) - \frac{1}{2}(\mathbf{u}-\widehat{\mathbf{u}})'\left(-\mathbf{H}_{g^*}|_{\mathbf{u}=\widehat{\mathbf{u}}}\right)(\mathbf{u}-\widehat{\mathbf{u}}),$$

where $\mathbf{H}_{g^*}|_{\mathbf{u}=\widehat{\mathbf{u}}}$ denotes the Hessian matrix of $g^*\left(\mathbf{u};\mathbf{\Psi}^{(s)}\right)$ evaluated at $\mathbf{u}=\widehat{\mathbf{u}}$. Therefore, the Q function (Equation B.20) can be approximated by

$$\begin{aligned} Q\left(\mathbf{\Psi}|\mathbf{\Psi}^{(s)}\right) \approx \int_{\mathbf{u}\in\mathbb{R}^N} & h^*\left(\mathbf{u};\mathbf{\Psi},\mathbf{\Psi}^{(s)}\right) \times \\ & \exp\left\{g^*\left(\widehat{\mathbf{u}};\mathbf{\Psi}^{(s)}\right) - \frac{1}{2}(\mathbf{u}-\widehat{\mathbf{u}})'\left(-\mathbf{H}_{g^*}|_{\mathbf{u}=\widehat{\mathbf{u}}}\right)(\mathbf{u}-\widehat{\mathbf{u}})\right\} d\mathbf{u}. \end{aligned} \tag{B.21}$$

If we further expand $h^*\left(\mathbf{u};\mathbf{\Psi},\mathbf{\Psi}^{(s)}\right)$ linearly around $\widehat{\mathbf{u}}$ as

$$h^*\left(\mathbf{u};\mathbf{\Psi},\mathbf{\Psi}^{(s)}\right) \approx h^*\left(\widehat{\mathbf{u}};\mathbf{\Psi},\mathbf{\Psi}^{(s)}\right) + (\mathbf{u}-\widehat{\mathbf{u}})'\,\mathbf{J}_{h^*}|_{\mathbf{u}=\widehat{\mathbf{u}}},$$

then we can re-express the Q function as

$$
\begin{aligned}
Q\left(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(s)}\right) \approx & \int_{\mathbf{u}\in\mathbb{R}^N} \left(h^*\left(\widehat{\mathbf{u}};\boldsymbol{\Psi},\boldsymbol{\Psi}^{(s)}\right) + (\mathbf{u}-\widehat{\mathbf{u}})' \mathbf{J}_{h^*}|_{\mathbf{u}=\widehat{\mathbf{u}}}\right) \times \\
& \times \exp\left\{g^*\left(\widehat{\mathbf{u}};\boldsymbol{\Psi}^{(s)}\right) - \frac{1}{2}(\mathbf{u}-\widehat{\mathbf{u}})' \left(-\mathbf{H}_{g^*}|_{\mathbf{u}=\widehat{\mathbf{u}}}\right)(\mathbf{u}-\widehat{\mathbf{u}})\right\} d\mathbf{u} \\
& = h^*\left(\widehat{\mathbf{u}};\boldsymbol{\Psi},\boldsymbol{\Psi}^{(s)}\right) \exp\left\{g^*\left(\widehat{\mathbf{u}};\boldsymbol{\Psi}^{(s)}\right)\right\} \times \\
& \times \int_{\mathbf{u}\in\mathbb{R}^N} \exp\left\{-\frac{1}{2}(\mathbf{u}-\widehat{\mathbf{u}})' \left(-\mathbf{H}_{g^*}|_{\mathbf{u}=\widehat{\mathbf{u}}}\right)(\mathbf{u}-\widehat{\mathbf{u}})\right\} d\mathbf{u} + \\
& + \underbrace{\left(\mathbf{J}_{h^*}|_{\mathbf{u}=\widehat{\mathbf{u}}}\right) \exp\left\{g^*\left(\widehat{\mathbf{u}};\boldsymbol{\Psi}^{(s)}\right)\right\} \int_{\mathbf{u}\in\mathbb{R}^N} (\mathbf{u}-\widehat{\mathbf{u}})' \exp\left\{-\frac{1}{2}(\mathbf{u}-\widehat{\mathbf{u}})' \left(-\mathbf{H}_{g^*}|_{\mathbf{u}=\widehat{\mathbf{u}}}\right)(\mathbf{u}-\widehat{\mathbf{u}})\right\} d\mathbf{u}}_{=0} \\
& = h^*\left(\widehat{\mathbf{u}};\boldsymbol{\Psi},\boldsymbol{\Psi}^{(s)}\right) \exp\left\{g^*\left(\widehat{\mathbf{u}};\boldsymbol{\Psi}^{(s)}\right)\right\} (2\pi)^{n/2} \left|\left(\left(-\mathbf{H}_{g^*}|_{\mathbf{u}=\widehat{\mathbf{u}}}\right)^{-1}\right)\right|^{1/2} \times \\
& \times \underbrace{\int_{\mathbf{u}\in\mathbb{R}^N} (2\pi)^{-n/2} \left|\left(\left(-\mathbf{H}_{g^*}|_{\mathbf{u}=\widehat{\mathbf{u}}}\right)^{-1}\right)\right|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{u}-\widehat{\mathbf{u}})' \left(-\mathbf{H}_{g^*}|_{\mathbf{u}=\widehat{\mathbf{u}}}\right)(\mathbf{u}-\widehat{\mathbf{u}})\right\} d\mathbf{u}}_{=1} \\
& = h^*\left(\widehat{\mathbf{u}};\boldsymbol{\Psi},\boldsymbol{\Psi}^{(s)}\right) \exp\left\{g^*\left(\widehat{\mathbf{u}};\boldsymbol{\Psi}^{(s)}\right)\right\} (2\pi)^{n/2} \left|\left(\left(-\mathbf{H}_{g^*}|_{\mathbf{u}=\widehat{\mathbf{u}}}\right)^{-1}\right)\right|^{1/2}.
\end{aligned}
$$

Recall that neither the function $g^*\left(\widehat{\mathbf{u}};\boldsymbol{\Psi}^{(s)}\right)$ nor its Hessian matrix $\mathbf{H}_{g^*}$ depends on $\boldsymbol{\Psi}$. Therefore, for the purposes of obtaining parameter estimates in the M-step, the quantity to be maximized is

$$
\begin{aligned}
Q^*\left(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(s)}\right) = h^*\left(\widehat{\mathbf{u}};\boldsymbol{\Psi},\boldsymbol{\Psi}^{(s)}\right) = \\
= \frac{h\left(\widehat{\mathbf{u}};\boldsymbol{\Psi},\boldsymbol{\Psi}^{(s)}\right)}{f\left(\mathbf{y};\boldsymbol{\Psi}^{(s)}\right)} = \\
\propto h\left(\widehat{\mathbf{u}};\boldsymbol{\Psi},\boldsymbol{\Psi}^{(s)}\right) = \\
= \sum_{k=1}^{2} P\left(Z_1 = k|\mathbf{y},\widehat{\mathbf{u}},\mathbf{r},\mathbf{x};\boldsymbol{\Psi}^{(s)}\right)\log(\pi_k) + \\
+ \sum_{j=1}^{M}\sum_{k=1}^{2}\sum_{i=1}^{N} P\left(Z_j = k|\mathbf{y},\widehat{\mathbf{u}},\mathbf{r},\mathbf{x};\boldsymbol{\Psi}^{(s)}\right) f_k(y_{ij}|\widehat{u}_i, r_{ij}, x_{ij};\boldsymbol{\psi}_k,\sigma) + \\
+ \sum_{j=2}^{M}\sum_{k=1}^{2}\sum_{l=1}^{2} P\left(Z_{j-1} = l, Z_j = k|\mathbf{y},\widehat{\mathbf{u}},\mathbf{r},\mathbf{x};\boldsymbol{\Psi}^{(s)}\right)\log(\gamma_{lk}).
\end{aligned}
$$

The Forward-Backward Algorithm

In the E-step, the forward and backward probabilities can be calculated as

Forward probabilities f_{jk}^{p}

$$f_{1k}^{p} = \pi_k \left(\prod_{i=1}^{N} f_k(y_{i1}|u_i, r_{i1}, x_{i1}; \boldsymbol{\psi}_k, \sigma) \right), \qquad \forall k = 1, 2.$$

$$f_{jk}^{p} = \sum_{l=1}^{2} \gamma_{lk} f_{(j-1)l}^{p} \left(\prod_{i=1}^{N} f_k(y_{ij}|u_i, r_{ij}, x_{ij}; \boldsymbol{\psi}_k, \sigma) \right), \qquad \forall j = 2, \ldots, M \quad \textbf{and} \quad k = 1, 2.$$

Backward probabilities b_{jk}^{p}

$$b_{Mk}^{p} = 1, \qquad \forall k = 1, 2.$$

$$b_{jk}^{p} = \sum_{l=1}^{2} \gamma_{kl} b_{(j+1)l}^{p} \left(\prod_{i=1}^{N} f_l(y_{i(j+1)}|u_i, r_{i(j+1)}, x_{i(j+1)}; \boldsymbol{\psi}_k, \sigma) \right),$$

$$\forall j = 1, \ldots, (M-1) \quad \textbf{and} \quad k = 1, 2.$$

The forward and backward probabilities can be used to calculate marginal and posterior probabilities as

$$P\left(Z_j = k|\mathbf{y}, \mathbf{u}, \mathbf{r}, \mathbf{x}; \boldsymbol{\psi}\right) = \frac{f_{jk}^{p} b_{jk}^{p}}{\sum_{l=1}^{2} f_{Ml}^{p}}, \qquad \forall j = 1, \ldots, M \quad \text{and} \quad k = 1, 2.$$

$$P\left(Z_{j-1} = l, Z_j = k|\mathbf{y}, \mathbf{u}, \mathbf{r}, \mathbf{x}; \boldsymbol{\psi}\right) = \frac{f_{(j-1)l}^{p} \gamma_{lk} \left(\prod_{i=1}^{N} f_k(y_{ij}|u_i, r_{ij}, x_{ij}; \boldsymbol{\psi}_k, \sigma) \right) b_{jk}^{p}}{\sum_{l=1}^{2} f_{Ml}^{p}},$$

$$\forall j = 2, \ldots, M \quad \text{and} \quad l, k = 1, 2.$$

Parameter Estimates and Derivatives

In the M-step, we maximize the Q function with respect to the unknown parameters. In order to avoid constrained numerical maximization, we reparametrize the dispersion parameters of the NB distribution and variance component as $\phi_k = \exp(\phi_k)$, for $k \in \{1, 2\}$, and $\sigma^2 = \exp(2\sigma)$. For a Zero-Inflated Mixed Effects HMM, the partial derivatives of Q with respect to the model parameters $\boldsymbol{\psi}_1 = (\boldsymbol{\lambda}^{'}, \boldsymbol{\beta}_1^{'}, \phi_1)^{'}$, $\boldsymbol{\psi}_2 = (\boldsymbol{\beta}_2^{'}, \phi_2)^{'}$, and σ are given by

$$\frac{\partial Q}{\partial \boldsymbol{\psi}_1^{'}} = \sum_{j=1}^{M} P\left(Z_j = 1 | \mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Psi}^{(\mathbf{s})}\right) \sum_{i=1}^{N} \left\{ \frac{\frac{\partial f_1(y_{ij}|\widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\psi_1}, \sigma)}{\partial \boldsymbol{\psi}_1^{'}}}{f_1(y_{ij}|\widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\psi_1}, \sigma)} \right\},$$

$$\frac{\partial Q}{\partial \boldsymbol{\psi}_2^{'}} = \sum_{j=1}^{M} P\left(Z_j = 2 | \mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Psi}^{(\mathbf{s})}\right) \sum_{i=1}^{N} \left\{ \frac{\frac{\partial f_2(y_{ij}|\widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\psi_2}, \sigma)}{\partial \boldsymbol{\psi}_2^{'}}}{f_2(y_{ij}|\widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\psi_2}, \sigma)} \right\},$$

$$\frac{\partial Q}{\partial \sigma} = \sum_{j=1}^{M} P\left(Z_j = 1 | \mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Psi}^{(\mathbf{s})}\right) \sum_{i=1}^{N} \left\{ \frac{\frac{\partial f_1(y_{ij}|\widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\psi_1}, \sigma)}{\partial \sigma}}{f_1(y_{ij}|\widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\psi_1}, \sigma)} \right\} +$$
$$+ \sum_{j=1}^{M} P\left(Z_j = 2 | \mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Psi}^{(\mathbf{s})}\right) \sum_{i=1}^{N} \left\{ \frac{\frac{\partial f_2(y_{ij}|\widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\psi_2}, \sigma)}{\partial \sigma}}{f_2(y_{ij}|\widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\psi_2}, \sigma)} \right\}.$$

Let p_{ij} denote the zero-inflation probability associated with the zero-inflation part of the f_1 model. In addition, let $\text{NB}_k(y_{ij})$ denote the NB emission distribution associated to the k^{th} model component evaluated at the integer y_{ij}. This notation implicitly assumes that the mean and dispersion are $\mu_{k,i,j}$ and $\exp(\phi_k)$, respectively. In the following derivatives, we will omit the conditional part of the component-specific emission distributions f_k for $k = \{1, 2\}$. It is assumed that f_k is conditional on the quantities $(\widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\psi_k}, \sigma)^{'}$. The

derivatives of the Q function with respect to the parameters of the component-specific densities are:

$$
\begin{aligned}
\frac{\partial f_1(y_{ij})}{\partial \lambda_0} &= I(y_{ij}=0)(p_{ij}(1-p_{ij})(1-\text{NB}_1(0)))+ \\
&\quad + I(y_{ij}>0)(-p_{ij}(1-p_{ij})\text{NB}_1(y_{ij})), \\
\frac{\partial f_1(y_{ij})}{\partial \lambda_1} &= I(y_{ij}=0)(p_{ij}(1-p_{ij})x_{ij}(1-\text{NB}_1(0)))+ \\
&\quad + I(y_{ij}>0)(-p_{ij}(1-p_{ij})x_{ij}\text{NB}_1(y_{ij})), \\
\frac{\partial f_1(y_{ij})}{\partial \beta_{11}} &= I(y_{ij}=0)(1-p_{ij})\frac{e^{\phi_1}\mu_{1ij}(1-\mu_{1ij})}{(\mu_{1ij}+e^{\phi_1})^2}\left(\frac{e^{\phi_1}}{\mu_{1ij}+e^{\phi_1}}\right)^{e^{\phi_1}} + \\
&\quad + I(y_{ij}>0)(1-p_{ij})\text{NB}_1(y_{ij})(y_{ij}-\mu_{1ij})\frac{e^{\phi_1}}{\mu_{1ij}+e^{\phi_1}}, \\
\frac{\partial f_1(y_{ij})}{\partial \beta_{12}} &= I(y_{ij}=0)(1-p_{ij})\frac{e^{\phi_1}\mu_{1ij}(1-\mu_{1ij})}{(\mu_{1ij}+e^{\phi_1})^2}\left(\frac{e^{\phi_1}}{\mu_{1ij}+e^{\phi_1}}\right)^{e^{\phi_1}} x_{ij}+ \\
&\quad + I(y_{ij}>0)(1-p_{ij})\text{NB}_1(y_{ij})(y_{ij}-\mu_{1ij})\frac{e^{\phi_1}}{\mu_{1ij}+e^{\phi_1}}x_{ij}, \\
\frac{\partial f_1(y_{ij})}{\partial \phi_1} &= I(y_{ij}=0)(1-p_{ij})\mu_{1ij}\left(\phi_1-\log(\mu_{1ij}+\phi_1)-\frac{1-\mu_{1ij}}{\mu_{1ij}+e^{\phi_1}}\right)\left(\frac{e^{\phi_1}}{\mu_{1ij}+e^{\phi_1}}\right)^{e^{\phi_1}+1} + \\
&\quad + I(y_{ij}>0)(1-p_{ij})\left(\frac{e^{\phi_1}}{\mu_{1ij}+e^{\phi_1}}\right)\text{NB}_1(y_{ij})\times \\
&\quad \times((\mu_{1ij}+e^{\phi_1})(\phi_1-\log(\mu_{1ij}+e^{\phi_1})+\varphi(y_{ij}+e^{\phi_1})+\varphi(e^{\phi_1}))+\mu_{1ij}-y_{ij}), \\
\frac{\partial f_2(y_{ij})}{\partial \beta_{21}} &= \text{NB}_2(y_{ij})(y_{ij}-\mu_{2ij})\frac{e^{\phi_2}}{\mu_{2ij}+e^{\phi_2}}, \\
\frac{\partial f_2(y_{ij})}{\partial \beta_{22}} &= \text{NB}_2(y_{ij})(y_{ij}-\mu_{2ij})\frac{e^{\phi_2}}{\mu_{2ij}+e^{\phi_2}}x_{ij}, \\
\frac{\partial f_2(y_{ij})}{\partial \phi_2} &= \left(\frac{e^{\phi_2}}{\mu_{2ij}+e^{\phi_2}}\right)\text{NB}_2(y_{ij})\times \\
&\quad \times((\mu_{2ij}+e^{\phi_2})(\phi_2-\log(\mu_{2ij}+e^{\phi_2})+\varphi(y_{ij}+e^{\phi_2})+\varphi(e^{\phi_2}))+\mu_{2ij}-y_{ij}),
\end{aligned}
$$

in which $\varphi(.)$ denotes the Digamma function.

Closed formulas for the initial and transition probabilities can be calculated in the M-step of the algorithm as

$$\pi_1^{(s+1)} = P\left(Z_1 = 1|\mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Psi}^{(s)}\right) \quad \text{and} \quad \pi_2^{(s+1)} = 1 - \pi_1^{(s+1)},$$

$$\gamma_{kk}^{(s+1)} = \frac{\sum_{j=2}^{M} P\left(Z_{j-1} = k, Z_j = k|\mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Psi}^{(s)}\right)}{\sum_{j=2}^{M}\sum_{l=1}^{2} P\left(Z_{j-1} = l, Z_j = k|\mathbf{y}, \widehat{\mathbf{u}}, \mathbf{r}, \mathbf{x}; \boldsymbol{\Psi}^{(s)}\right)} \quad \text{and} \quad \gamma_{kl}^{(s+1)} = 1 - \gamma_{kk}^{(s+1)}.$$

APPENDIX B: TECHNICAL DETAILS FOR CHAPTER 3

Data

The data utilized in Chapter 2 pertaining to the ENCODE Consortium are listed in Table B.10.

Table B.10: GEO sample accession codes of the analyzed data from the ENCODE Consortium in Chapter 2.

Cell Line	H3K27me3	H3K36me3	EZH2	H3K4me3	H3K27ac	CTCF	RNA-seq
H1hesc	GSM733748	GSM733725	GSM1003524	GSM733657	GSM733718	GSM733672	GSM758566
HelaS3	GSM733696	GSM733711	GSM1003520	GSM733682	GSM733684	GSM733785	GSM765402
Hepg2	GSM733754	GSM733685	GSM1003487	GSM733737	GSM733743	GSM733645	GSM758575
Huvec	GSM733688	GSM733757	GSM1003518	GSM733673	GSM733691	GSM733716	GSM758563

The following steps were conducted to process the data. First, we removed PCR duplicates from the BAM files using SAMTools (Li et al., 2009) and converted the resulting indexed and sorted files to BED format using BEDTools (Quinlan and Hall, 2010), as RSEG only accepts such a format as input. Then, the fragment length of each ChIP-seq experiment was estimated using csaw and its functions *correlateReads* and *maximizeCcf*. Finally, using the estimated fragment length, read counts from all cell lines were tabulated for their ChIP replicates using fixed-step and non overlapping windows of size 250bp, 500bp, 750bp, and 1000bp through the R package *bamsignals* (Mammana and Helmuth, 2016). For all methods using window-based approaches (csaw, ChIPComp, diffReps, RSEG, THOR, and mixNBHMM), we assessed their performance with different window sizes. See Section 3.3.2 and Baldoni et al. 2019a for a discussion about results with different window sizes.

All the methods considered in the data applications and simulation study output a set of differential genomic regions/windows that were used for benchmark purposes. THOR output a list of differential peaks in BED6+4 format (*narrowPeak*) with adjusted p-values. RSEG output a WIG file with genomic windows and their posterior probabilities for differential enrichment. diffReps output an annotated TXT file with differential regions of enrichment and their adjusted p-values. DiffBind output a TXT file with differential regions of enrichment and their respective multiple testing corrected FDR. diffReps output a TXT file with differential regions of enrichment and their p-values. csaw output a TSV file with differential regions of enrichment and their FDR adjusted p-values. For a fair FDR thresholding comparison, we control the total FDR and output the differential regions of enrichment based on the set of posterior probabilities as described in Section 3.3.2. For a comparison between the Viterbi and the FDR thresholding approach, see Section 3.3.2 and Baldoni et al. 2019a.

The following parametrization was used when calling peaks from the benchmarked methods. For THOR, *rgt-THOR 'config' –name 'name' -b 'bp' –pvalue 1.0 –output-dir 'output'*. For RSEG, *rseg-diff -verbose -mode 3 -out 'output' -score 'score' -chrom 'chrom' -bin-size 'bp' -deadzones 'deadzonee' -duplicates 'sample1' 'sample2'*.

For ChIPComp, *ChIPComp(makeCountSet(conf,design,filetype="bam",species="hg19",binsize=bp))*. For diffReps, *diffReps.pl –gname hg19 –report 'output' –treatment 'sample1' –control 'sample2' –btr 'control1' –bco 'control2' –window 'bp' –pval 1 –nsd 'marktype' –meth 'nb'*.

For DiffBind, *dba.report(dba.analyze(dba.contrast(dba.count(dba(sampleSheet = conf)), categories = DBA_CONDITION, minMembers=2)), th=1)*. In this parametrization, bp $= \{250, 500, 750, 100\}$ and $marktype =' broad'$ if H3K27me3, H3K36me3, or EZH2, or $marktype =' sharp'$ otherwise. For csaw, we followed the authors's recommended settings and the details are presented in (Baldoni et al., 2019a).

For DiffBind under 3 conditions (Figure 3.6), the set of differential peaks included all peaks deemed to be differential by DiffBind under an FDR control of 0.05 simultaneously for all three pairwise contrast tests between the cell lines Helas3, Hepg2, and Huvec. In the particular genomic position shown in Figure 3.6B, no differential peaks were reported by DiffBind.

For ChIPComp and DiffBind, candidate peaks were called in advance using MACS with the syntax *macs2 callpeak -f BAM -g 2.80e+09 -B 'options' -t 'sample' -c 'control' –outdir 'output' -n 'filename'*, such that $options = \{$–broad –broad-cutoff 0.1$\}$ if H3K27me3, H3K36me3, or EZH2, or $options = \{$-q 0.01$\}$ otherwise.

Software

mixNBHMM is available on https://github.com/plbaldoni/mixNBHMM as an R package.

mixNBHMM is a package with a differential peak caller to detect differential enrichment regions from multiple ChIP-seq experiments with replicates. The main function of the package is *mixNBHMM*(). The package allows the user to specify a set of parameters that control the Expectation-Maximization (EM) algorithm. These parameters include, for instance, the convergence (and termination) criteria of the algorithm and the threshold value for the rejection controlled EM algorithm. These parameters can be defined by the function *controlEM*(). Please refer to the package documentation (e.g. *?mixNBHMM::mixNBHMM*) for additional details and the complete help manual.

Code

The necessary code to replicate the results presented in the main article and in the supplementary material can be downloaded from https://github.com/plbaldoni/mixNBHMMPaper.

The EM Algorithm

A pseudo code of the presented EM algorithm is below.

1. Initialize $\boldsymbol{\pi}^{(0)}, \boldsymbol{\gamma}^{(0)}, \boldsymbol{\delta}^{(0)}, \beta_1^{(0)}, \beta_3^{(0)}, \lambda_1^{(0)}, \lambda_3^{(0)}$, such that $\sum_{r=1}^{3} \pi_r^{(0)} = 1$ and $\sum_{s=1}^{3} \gamma_{rs} = 1$.

2. E step ($t \geq 1$),

   (a) Calculate $Pr\left(Z_j = r|\mathbf{y}, \mathbf{x}; \boldsymbol{\Psi}^{(t-1)}\right)$ and $Pr\left(Z_{j-1} = r, Z_j = s|\mathbf{y}, \mathbf{x}; \boldsymbol{\Psi}^{(t-1)}\right)$ for all r and s in $\{1, 2, 3\}$ and $j = 1, \ldots, M$ via Forward-Backward algorithm as detailed in Appendix B of the main article

   (b) Calculate $Pr(W_{jl} = 1|Z_j = 2, \mathbf{y}_{..j}, \mathbf{x}; \boldsymbol{\Psi}^{(t-1)})$ for all l $\{1, \ldots, L\}$ and $j = 1, \ldots, M$ as $f_{(2,l)}(\mathbf{y}_{..j}|\mathbf{x}_l; \boldsymbol{\psi}_{(2,l)}^{(t-1)})\delta_l^{(t-1)} / \sum_{k=1}^{L} f_{(2,k)}(\mathbf{y}_{..j}|\mathbf{x}_k; \boldsymbol{\psi}_{(2,k)}^{(t-1)})\delta_k^{(t-1)}$

3. M step ($t \geq 1$),

   (a) Maximize Equation 3.10 with respect to the initial and transition probabilities to obtain for all r and s in $\{1, 2, 3\}$

$$\pi_r^{(t)} = Pr\left(Z_1 = r|\mathbf{y}, \mathbf{x}; \boldsymbol{\Psi}^{(t-1)}\right)$$
$$\gamma_{rs}^{(t)} = \sum_{j=2}^{M} Pr(Z_{j-1} = r, Z_j = s|\mathbf{y}, \mathbf{x}; \boldsymbol{\Psi}^{(t-1)}) / \sum_{j=2}^{M} Pr(Z_{j-1} = r|\mathbf{y}, \mathbf{x}; \boldsymbol{\Psi}^{(t-1)})$$

   (b) Maximize Equation 3.10 with respect to $\boldsymbol{\delta}$ to obtain $\boldsymbol{\delta}^{(t)}$ such that $\sum_{l=1}^{L} \delta_l^{(t)} = 1$.

   (c) Conditionally upon $\boldsymbol{\delta}^{(t)}$, maximize Equation 3.10 with respect to $\beta_1, \beta_3, \lambda_1, \lambda_3$ to obtain $\beta_1^{(t)}$, $\beta_3^{(t)}$, $\lambda_1^{(t)}$, $\lambda_3^{(t)}$,

   (d) Iterate between (b) and (c) until convergence.

4. Iterate between 2. and 3. until convergence.

Adjustments for nuisance effects

Normalization for non-linear biases via model offsets

In our analyses, we observed that the magnitude of the local differences in read counts between conditions changed with the average of local read coverage. Here, we accounted for these trended differences to avoid calling spurious differential peaks due to the different magnitude of library sizes across groups. Specifically, we implemented an approach similar to the non-linear normalization method used by csaw as follows (Lun and Smyth, 2015). First, we create a reference sample of read counts formed by the geometric mean of read counts from all replicates and conditions. Then, we fitted a loess curve on the difference between the read counts of each sample and the reference on the average of those two quantities. A similar approach was first implemented by (Lun and Smyth, 2015) and is available in their software. Here, we add a continuity correction of 1 to avoid discarding genomic windows with zero counts. Using the smoothed curve as the model offset, we observed better results than a simple correction via either the total sum of read counts or cell-specific median log ratio. The rationale behind this approach is to create a reference library in which each genomic window is the geometric mean of counts across all conditions and replicates, and then read counts are properly adjusted by accounting for the smoothed differences between each individual library and the reference library. A useful way to evaluate the performance of this normalization method is to compare samples with respect to their adjusted read counts. For example, plotting the ratios between counts and the calculated offsets $y_{hij}/exp(u_{hij})$ for all samples in the study. In Figure B.21 we show an example of a genomic region from three analyzed cell lines and their respective MA plot, unadjusted ChIP counts, and offset-adjusted ChIP counts. After accounting for the offset, the read counts from Helas3 are adjusted to its larger library size with respect to the other under sequenced cell lines.

Input control adjustment in differential peak calling

Our implementation allows the optional inclusion of continuous covariates in the model with state-specific parametrization. The main purpose of the inclusion of such covariates in the model is the adjustment for input control (or any other continuous variable, such as autoregressive counts) that can be helpful in distinguishing background from enrichment signal. Several methods for differential peak calling allow the inclusion of input control in their computational framework (Stark and Brown, 2011; Shen et al., 2013; Chen et al., 2015). However, Lun and Smyth (2015) point out that "(...) controls are mostly irrelevant when testing for DB

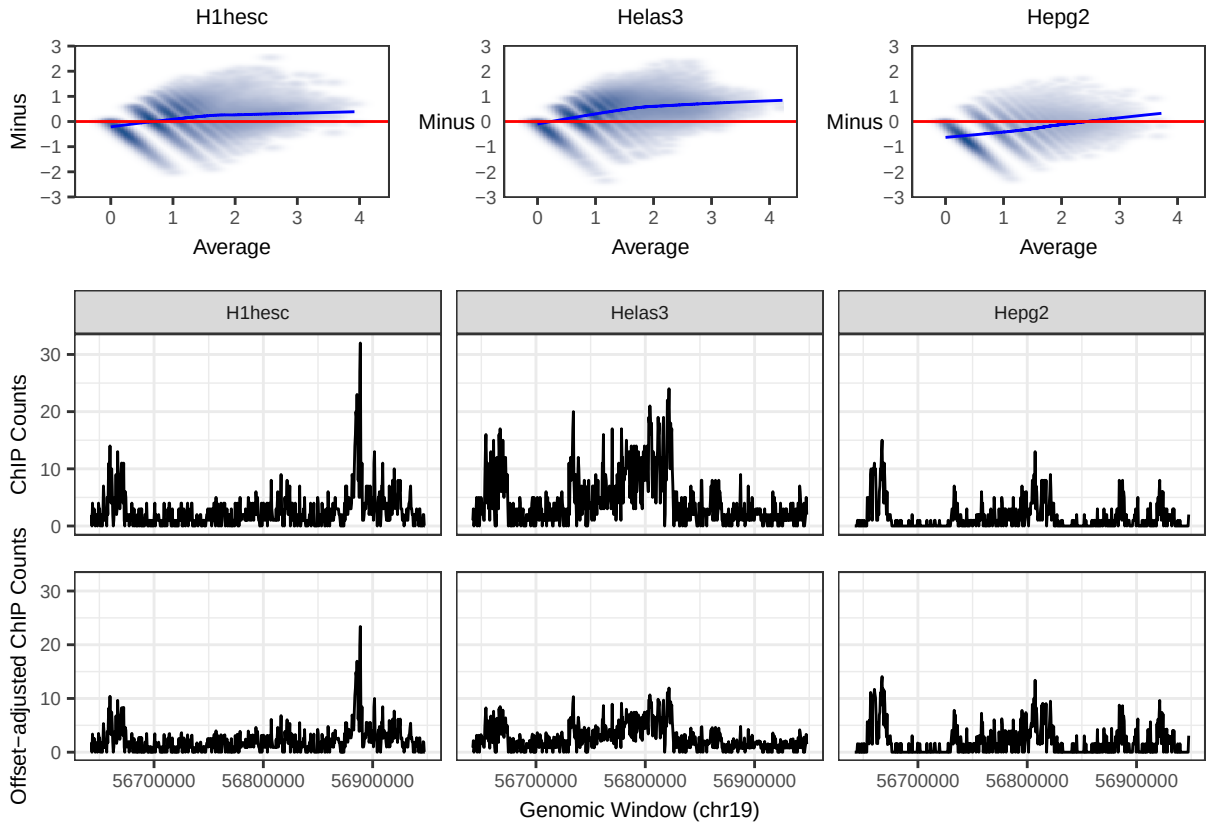

Figure B.21: MA plot of read counts from three distinct analyzed cell lines (top), unadjusted ChIP read counts (center), and offset-adjusted ChIP read counts (bottom) from a given genomic region on chromosome 19. The blue line in the MA plots shows the offset created via loess smoothing.

between ChIP samples.". To evaluate this claim, we ran an analysis of real data and simulated data while accounting for the input control effect.

To asses whether accounting for input control effect leads to an improvement in performance, we utilized the smoothing technique proposed by Chen et al. (2015) to account for input controls and autoregressive counts. Specifically, we fitted generalized additive models (GAM, instead of loess smoothing) in the data normalization step while accounting for input control (or autoregressive counts) as a covariate. The resulting fitted curve was then used in the analysis as model offsets.

First, we analyzed real data by smoothing the input control effect and autoregressive counts with a two-step approach. Specifically, we first called peaks without the inclusion of extra covariates in the model, and then utilized the called differential peaks from the first step to smooth the covariates for each HMM predicted state. Predicted smoothing curves from the GAM approach were then passed as model offsets in a

second step of analysis. As claimed by Lun and Smyth (2015), we observed minor differences in the results that would justify their inclusion in the analysis. Results from the histone modification mark H3K36me3 are presented in Figure B.22.

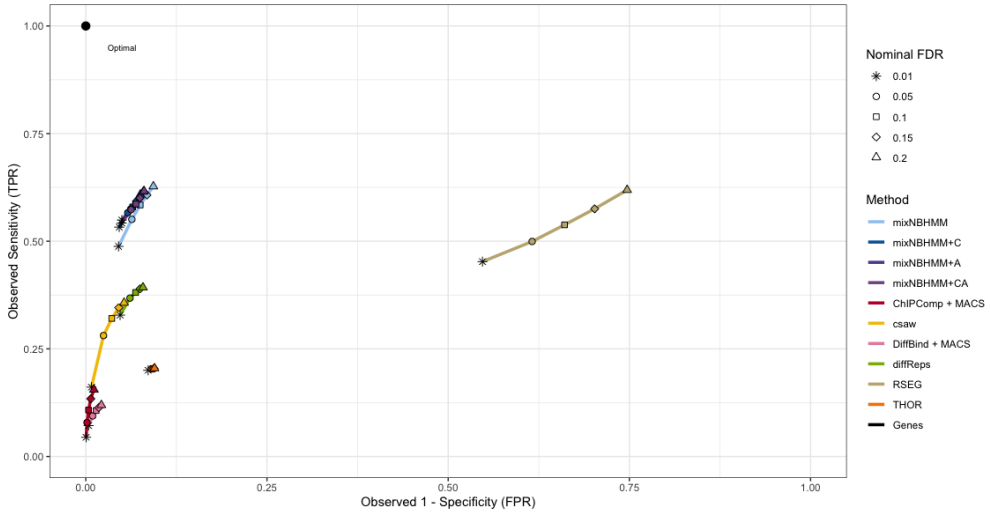

Figure B.22: ROC curves for H3K36me3 utilizing no input controls (mixNBHMM), input control only (mixNBHMM + C), autoregressive counts only (mixNBHMM + A), and smoothing of both input controls and autoregressive counts (mixNBHMM + CA)

Next, we reasoned that our approach of modeling input control effect with state-specific parametrization could not be ideal, since independent controls were available for every sample and there could exist sample-specific effects not captured by our model. We then attempted to verify the utility of including input control into the differential binding analysis by simulating data where ChIP-counts were generated such that their log-mean had a linear relationship with input controls (Figure B.23). We then fitted three different models that differed regarding the inclusion of input control: a model without control, a model with control, and a model with controls where the smoothing was calculated separately for each latent HMM state. Again, results did not show significant improvement by including the effect of control in the analysis.

Overall, we did not observe a significant improvement in performance by including input control in differential peak calling. Although several methods do offer the option of including controls in their analysis pipeline, we did not find that their inclusion was justifiable under our modelling assumptions. Our findings are in agreement with Lun and Smyth (2015).

Bayesian Information Criterion (BIC) for Hidden Markov Models

The BIC for hidden Markov models has been discussed by Zucchini et al. (2017). For the presented three-state HMM, one can calculate the BIC as

$$BIC = -2\log\left(\sum_{r=1}^{3} f_{Mr}^{p}\right) + (11 + L)\log\left(M\sum_{h=1}^{G} n_h\right), \tag{C.22}$$

where f_{Mr}^{p} is the forward probability pertaining to the r^{th} state calculated at the (last) M^{th} genomic window (as detailed in the Appendix of the main text), L is the number of mixture components, G is the number of conditions, and n_h is the number of replicates pertaining to condition h. The number of model parameters to be estimated is $(11 + L)$: 6 transition probabilities, 2 initial probabilities, 4 model coefficients pertaining to the emission distributions, and $L - 1$ prior probabilities from the mixture model.

As shown in the main text, the proposed HMM is robust to situations where certain combinatorial patters are rare. However, if pruning rare combinatorial patterns is still of interest, such a task can be performed by making use of the BIC. For the analysis of G experimental conditions with a given BIC threshold ϵ, say $\epsilon = 0.01$, and $L = 2^G - 2$ mixture components, one can prune rare combinatorial patterns by the following algorithm.

1. Fit the three-state HMM with L mixture components (model 0) and compute the model BIC, BIC_0, as in Equation C.22.

2. Fit a reduced three-state HMM with $L - 1$ mixture components (model 1) by excluding the component associated with the rarest combinatorial pattern of enrichment. Compute its BIC, BIC_1.

3. Calculate $\Delta\text{BIC} = (\text{BIC}_1 - \text{BIC}_0)/\text{BIC}_0$. If $|\Delta\text{BIC}| \leq \epsilon$, set $L \leftarrow L - 1$ and return to 1.. If $|\Delta\text{BIC}| > \epsilon$, stop and set the model 0 as the final model.

In scenarios where the number of mixture components is smaller than $2^G - 2$, the implemented method initializes the EM algorithm by clustering genomic windows with respect to the posterior probabilities of enrichment obtained from a initial run of a two-state HMM to classify genomic windows into background and enrichment windows. Such an initialization improves the overall computation time by reducing the time to convergence of the presented EM algorithm.

We applied the above approach in real data where the goal was to reduce the number of rare mixture components. Specifically, we reanalyzed the data presented in Section 5.3 of the main text by refitting

the presented model with reduced number of combinatorial patterns. Figure B.24 presents the BIC from various models regarding the number of mixture components on data from epigenomic marks H3K37me3, H3K36me3, and EZH2. As shown, models with more than 2 mixture components exhibited values of BIC quite close to each other. Conversely, the model with a single differential component had an excessively large BIC. These results suggest that, according to the BIC, the parsimonious model with only 2 mixture components would be the one chosen. As detailed in the main text, the analyzed data sets are characterized by only 2 combinatorial patterns of enrichment, which are associated with the enrichment of H3K36me3 alone, and the enrichment of H3K27me3 and EZH2 in consensus. Hence, choosing the model with 2 components via BIC agrees with the biological roles of the analyzed marks.

The Forward-Backward Algorithm and Posterior Probabilities

The Q-function of the EM algorithm is defined as $Q\left(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(t)}\right) = Q_0\left(\boldsymbol{\pi},\boldsymbol{\gamma}|\boldsymbol{\Psi}^{(t)}\right) + Q_1\left(\boldsymbol{\psi}_1|\boldsymbol{\Psi}^{(t)}\right) + Q_2\left(\boldsymbol{\delta},\boldsymbol{\psi}_2|\boldsymbol{\Psi}^{(t)}\right) + Q_3\left(\boldsymbol{\psi}_3|\boldsymbol{\Psi}^{(t)}\right)$, such that

$$
\begin{aligned}
Q_0\left(\boldsymbol{\pi},\boldsymbol{\gamma}|\boldsymbol{\Psi}^{(t)}\right) &= \sum_{r=1}^{3}\left\{Pr\left(Z_1 = r|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)\log(\pi_r)\right\} + \\
&\quad + \sum_{j=2}^{M}\sum_{r=1}^{3}\sum_{s=1}^{3}\left\{Pr\left(Z_{j-1} = r, Z_j = s|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)\log(\gamma_{rs})\right\}, \\
Q_1\left(\boldsymbol{\psi}_1|\boldsymbol{\Psi}^{(t)}\right) &= \sum_{j=1}^{M} Pr\left(Z_j = 1|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)\log f_1\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_1\right), \\
Q_2\left(\boldsymbol{\delta},\boldsymbol{\psi}_2|\boldsymbol{\Psi}^{(t)}\right) &= \sum_{j=1}^{M} Pr\left(Z_j = 2|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)\sum_{l=1}^{L} Pr(W_{jl} = 1|Z_j = 2,\mathbf{y}_{..j},\mathbf{x};\boldsymbol{\Psi}^{(t)})\times \\
&\quad \times\left\{\log f_{(2,l)}\left(\mathbf{y}_{..j}|\mathbf{x}_l;\boldsymbol{\psi}_{(2,l)}\right) + \log(\delta_l)\right\}, \quad \text{and} \\
Q_3\left(\boldsymbol{\psi}_3|\boldsymbol{\Psi}^{(t)}\right) &= \sum_{j=1}^{M} Pr\left(Z_j = 3|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}^{(t)}\right)\log f_3\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_3\right),
\end{aligned}
$$

in which $f_1\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_1\right)$, $f_{(2,l)}\left(\mathbf{y}_{..j}|\mathbf{x}_l;\boldsymbol{\psi}_{(2,l)}\right)$, and $f_3\left(\mathbf{y}_{..j}|\boldsymbol{\psi}_3\right)$ are defined in Equations 3.7 and 3.8, respectively.

Define, for $j = 1,\ldots,M$, the forward probabilities as

$$
\begin{aligned}
&f_{11}^p = \pi_1 f_1(\mathbf{y}_{..1}|\boldsymbol{\psi}_1), f_{12}^p = \pi_2 f_2(\mathbf{y}_{..1}|\mathbf{x};\boldsymbol{\psi}_2), f_{13}^p = \pi_3 f_3(\mathbf{y}_{..1}|\boldsymbol{\psi}_3),\\
&f_{j1}^p = \sum_{l=1}^{3} \gamma_{l1} f_{(j-1)l}^p f_1(\mathbf{y}_{..j}|\boldsymbol{\psi}_1),\\
&f_{j2}^p = \sum_{l=1}^{3} \gamma_{l2} f_{(j-1)l}^p f_2(\mathbf{y}_{..j}|\mathbf{x};\boldsymbol{\delta},\boldsymbol{\psi}_2),\\
&f_{j3}^p = \sum_{l=1}^{3} \gamma_{l3} f_{(j-1)l}^p f_3(\mathbf{y}_{..j}|\boldsymbol{\psi}_3).
\end{aligned}
$$

Conversely, for $j = 1, \ldots, M$, define the backward probabilities as

$$
\begin{aligned}
b_{Mk}^p &= 1, \qquad \forall k = 1, 2, 3,\\
b_{j1}^p &= \sum_{l=1}^{3} \gamma_{1l} b_{(j+1)l}^p f_1(\mathbf{y}_{..(j+1)}|\boldsymbol{\psi}_1),\\
b_{j2}^p &= \sum_{l=1}^{3} \gamma_{2l} b_{(j+1)l}^p f_2(\mathbf{y}_{..(j+1)}|\mathbf{x};\boldsymbol{\delta},\boldsymbol{\psi}_2),\\
b_{j3}^p &= \sum_{l=1}^{3} \gamma_{3l} b_{(j+1)l}^p f_3(\mathbf{y}_{..(j+1)}|\boldsymbol{\psi}_3).
\end{aligned}
$$

Then, we have the following posterior probabilities

$$
\begin{aligned}
P\left(Z_j = k|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}\right) &= \frac{f_{jk}^p b_{jk}^p}{\sum_{l=1}^{3} f_{Ml}^p}, \qquad \forall j = 1, \ldots, M \quad \text{and} \quad k = 1, 2, 3,\\
P\left(Z_{j-1} = l, Z_j = 1|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}\right) &= \frac{f_{(j-1)l}^p \gamma_{l1} f_1(\mathbf{y}_{..j}|\boldsymbol{\psi}_1) b_{j1}^p}{\sum_{l=1}^{3} f_{Ml}^p},\\
P\left(Z_{j-1} = l, Z_j = 2|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}\right) &= \frac{f_{(j-1)l}^p \gamma_{l2} f_2(\mathbf{y}_{..j}|\mathbf{x};\boldsymbol{\delta},\boldsymbol{\psi}_2) b_{j2}^p}{\sum_{l=1}^{3} f_{Ml}^p},\\
P\left(Z_{j-1} = l, Z_j = 3|\mathbf{y},\mathbf{x};\boldsymbol{\Psi}\right) &= \frac{f_{(j-1)l}^p \gamma_{l3} f_3(\mathbf{y}_{..j}|\boldsymbol{\psi}_3) b_{j3}^p}{\sum_{l=1}^{3} f_{Ml}^p}, \forall j = 2, \ldots, M \text{ and } l = 1, 2, 3.
\end{aligned}
$$

HMM Emission Distributions

For the consensus background ($r = 1$) and consensus enrichment ($r = 3$) states, the emission distribution function is

$$\begin{aligned} f_r(\mathbf{y}_{..j}|\boldsymbol{\psi}_r) &= \prod_{h=1}^{G}\prod_{i=1}^{n_h} f_r(y_{hij}|\boldsymbol{\psi}_r), \quad r \in \{1,3\} \quad \text{and} \quad y_{hij} \in \{0,1,2,\ldots\}, \\ &= \prod_{h=1}^{G}\prod_{i=1}^{n_h} \text{Pr}(Y_{hij} = y_{hij}|Z_j = r;\boldsymbol{\psi}_r), \\ &= \prod_{h=1}^{G}\prod_{i=1}^{n_h} \frac{\Gamma(y_{hij}+\phi_r)}{y_{hij}!\Gamma(\phi_r)} \left(\frac{\phi_r}{\mu_{(r,hij)}+\phi_r}\right)^{\phi_r} \left(\frac{\mu_{(r,hij)}}{\mu_{(r,hij)}+\phi_r}\right)^{y_{hij}}. \end{aligned} \tag{C.23}$$

For the differential state ($r = 2$), the emission distribution is

$$\begin{aligned} f_2(\mathbf{y}_{..j}|\mathbf{x};\boldsymbol{\delta},\boldsymbol{\psi}_2) &= \sum_{l=1}^{L} \delta_l f_{(2,l)}\left(\mathbf{y}_{..j}|\mathbf{x}_l;\boldsymbol{\psi}_{(2,l)}\right), \quad y_{hij} \in \{0,1,2,\ldots\}, \\ &= \sum_{l=1}^{L} \delta_l \prod_{h=1}^{G}\prod_{i=1}^{n_h} Pr\left(Y_{hij} = y_{hij}|Z_j = 2,\mathbf{x}_l;\boldsymbol{\psi}_{(2,l)}\right), \\ &= \sum_{l=1}^{L} \delta_l \prod_{h=1}^{G}\prod_{i=1}^{n_h} \frac{\Gamma\left(y_{hij}+\phi_{(2,l,h)}\right)}{y_{hij}!\Gamma\left(\phi_{(2,l,h)}\right)} \left(\frac{\phi_{(2,l,h)}}{\mu_{(2,l,hij)}+\phi_{(2,l,h)}}\right)^{\phi_{(2,l,h)}} \times \\ &\quad \times \left(\frac{\mu_{(2,l,hij)}}{\mu_{(2,l,hij)}+\phi_{(2,l,h)}}\right)^{y_{hij}}. \end{aligned} \tag{C.24}$$

Apart from the offset u_{hij}, we will assume that replicates from the same (different) condition share common (distinct) mean and dispersion parameters under every mixing probability distribution $f_{(2,l)}$. To define all possible combinations of background and enrichment across G conditions, we consider the following sets of singletons A_1, pairs A_2, …, and $(G-1)$-tuples A_{G-1} such that

$$\begin{aligned} A_1 &= \left\{a^{(1)} \mid a^{(1)} \in \mathbb{G}_+ \text{ and } a^{(1)} \leq G\right\}, \\ A_2 &= \left\{(a_1^{(2)}, a_2^{(2)}) \mid (a_1^{(2)}, a_2^{(2)}) \in \mathbb{G}_+^2 \text{ and } a_1^{(2)} < a_2^{(2)} \leq G\right\}, \\ &\vdots \\ A_{G-1} &= \left\{(a_g^{(G-1)})_{g=1}^{G-1} \mid (a_g^{(G-1)})_{g=1}^{G-1} \in \mathbb{G}_+^{G-1} \text{ and } a_1^{(G-1)} < \ldots < a_{G-1}^{(G-1)} \leq G\right\}. \end{aligned}$$

The union of all sets $A = \cup_{k=1}^{G-1} A_k$ contains an exhaustive list of $L = 2^G - 2$ elements that determines the differential pattern across G conditions such that each element of A indicates which of the G conditions are enriched. For instance, if $G = 3$, $A_1 = \{1, 2, 3\}$ and $A_2 = \{(1, 2), (1, 3), (2, 3)\}$ define the six possible combinations of enrichment and background across three conditions. Then, we define a bijective mapping $A \rightarrow S_1, \ldots, S_L$ and let $x_{hl} = I(h \in S_l)$ indicate whether the read count of genomic window j from replicate i of condition h is enriched in the mixture component l. We model the log-mean $\mu_{(2,l,hij)}$ and log-dispersion $\phi_{(2,l,h)}$ of mixture l from the emission distribution of Equation C.24 as

$$\log(\mu_{(2,l,hij)}) = \beta_1 + \beta_3 x_{hl} + u_{hij}, \quad \text{and}$$
$$\log(\phi_{(2,l,h)}) = \lambda_1 + \lambda_3 x_{hl}.$$

According to this parametrization, β_1 and λ_1 are the baseline log-mean and log-dispersion parameters of the read count distribution from replicates of conditions that are not enriched under the mixing distribution l. Conversely, $\beta_1 + \beta_3$ and $\lambda_1 + \lambda_3$ are the baseline log-mean and log-dispersion parameters of the read count distribution from replicates of conditions enriched under the mixing distribution l. This choice of parametrization ensures that windows exhibiting differential enrichment across conditions share means and dispersions that are common between the remaining non differential HMM states.

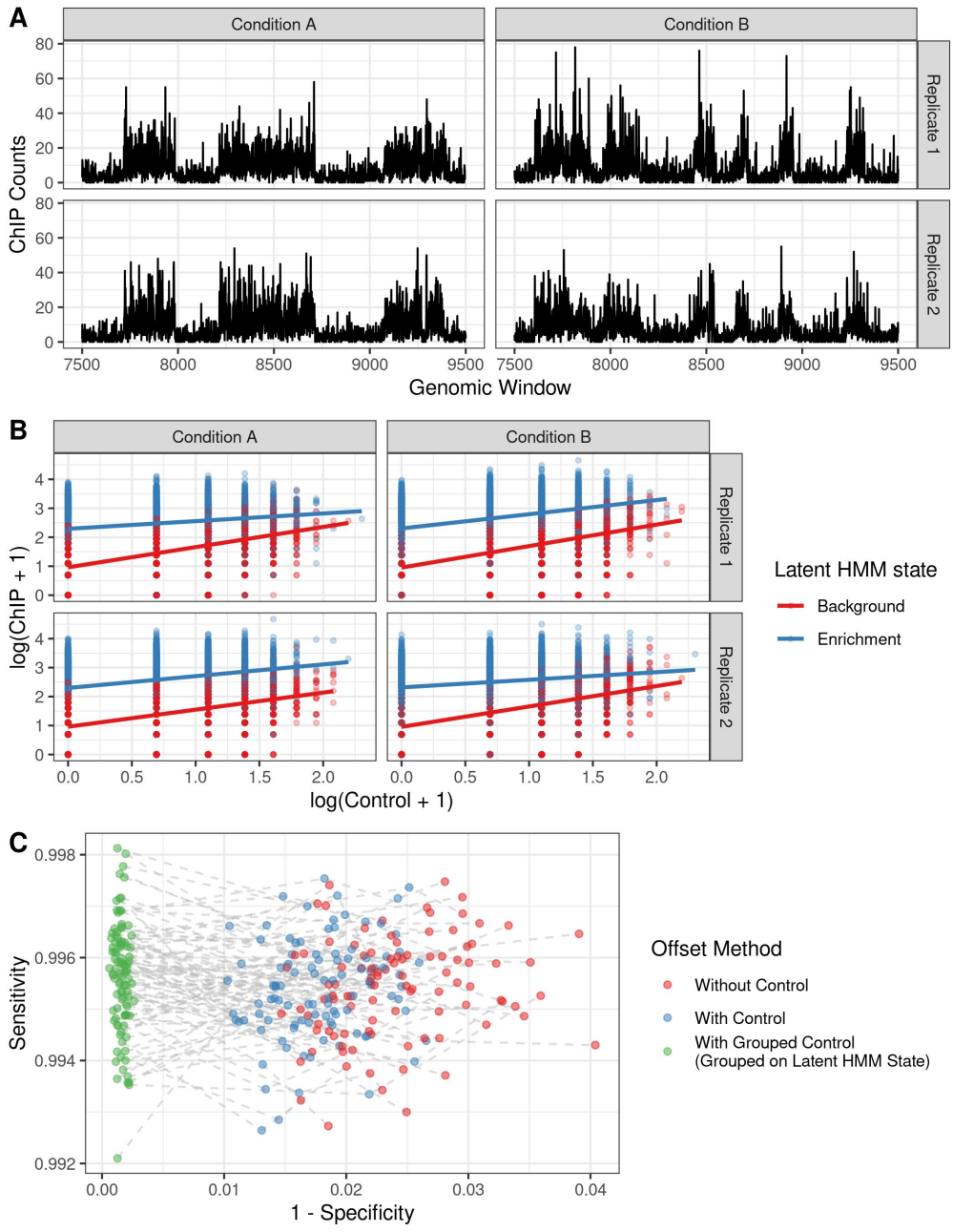

Figure B.23: Results from simulated data (A) where the log-means of ChIP-seq counts were generated as a linear function of input controls (B). Sensitivity/specificity analyses did not show significant improvement by including the effect of control in the offset scheme.

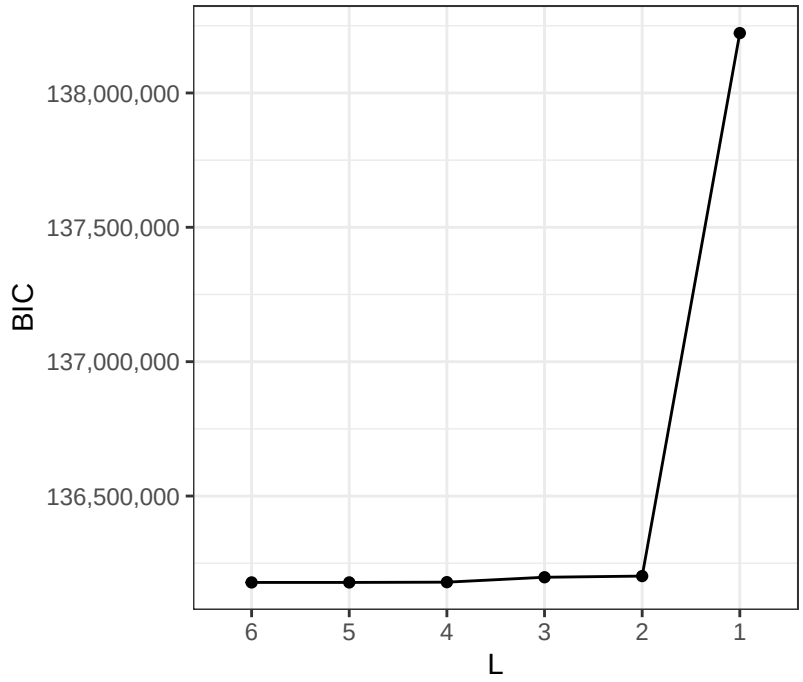

Figure B.24: BIC from various models regarding their number of mixture components on epigenomic marks H3K36me3, H3K27me3, and EHZ2 (Section 3.5.3)

BIBLIOGRAPHY

Abdelsamed, H. A., Zebley, C. C., Nguyen, H., Rutishauser, R. L., Fan, Y., Ghoneim, H. E., Crawford, J. C., Alfei, F., Alli, S., Ribeiro, S. P., et al. (2020). Beta cell-specific cd8+ t cells maintain stem cell memory-associated epigenetic programs during type 1 diabetes. *Nature Immunology*, 21(5):578–587.

Allhoff, M., Seré, K., Chauvistré, H., Lin, Q., Zenke, M., and Costa, I. G. (2014). Detecting differential peaks in chip-seq signals with odin. *Bioinformatics*, 30(24):3467–3475.

Allhoff, M., Seré, K., F. Pires, J., Zenke, M., and G. Costa, I. (2016). Differential peak calling of chip-seq signals with replicates with thor. *Nucleic acids research*, 44(20):e153–e153.

Altman, R. M. (2007). Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102(477):201–210.

Baker, S. M., Rogerson, C., Hayes, A., Sharrocks, A. D., and Rattray, M. (2019). Classifying cells with scasat, a single-cell atac-seq analysis tool. *Nucleic acids research*, 47(2):e10–e10.

Baldoni, P. L., Rashid, N. U., and Ibrahim, J. G. (2019a). Efficient detection and classification of epigenomic changes under multiple conditions. *bioRxiv*, page 864124.

Baldoni, P. L., Rashid, N. U., and Ibrahim, J. G. (2019b). Improved detection of epigenomic marks with mixed-effects hidden markov models. *Biometrics*, 75(4):1401–1413.

Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell research*, 21(3):381.

Bardet, A. F., Steinmann, J., Bafna, S., Knoblich, J. A., Zeitlinger, J., and Stark, A. (2013). Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics*, 29(21):2705–2713.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., et al. (2010). The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, 28(10):1045.

Bottomly, D., Kyler, S. L., McWeeney, S. K., and Yochum, G. S. (2010). Identification of β-catenin binding regions in colon cancer cells using chip-seq. *Nucleic acids research*, 38(17):5735–5745.

Buenrostro, J. D., Corces, M. R., Lareau, C. A., Wu, B., Schep, A. N., Aryee, M. J., Majeti, R., Chang, H. Y., and Greenleaf, W. J. (2018). Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, 173(6):1535–1548.

Bulyk, M. L., Gentalen, E., Lockhart, D. J., and Church, G. M. (1999). Quantifying dna–protein interactions by double-stranded dna arrays. *Nature biotechnology*, 17(6):573.

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R. S., and Zhang, Y. (2002). Role of histone h3 lysine 27 methylation in polycomb-group silencing. *Science*, 298(5595):1039–1043.

Chantalat, S., Depaux, A., Héry, P., Barral, S., Thuret, J.-Y., Dimitrov, S., and Gérard, M. (2011). Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. *Genome research*, 21(9):1426–1437.

Chen, H., Lareau, C., Andreani, T., Vinyard, M. E., Garcia, S. P., Clement, K., Andrade-Navarro, M. A., Buenrostro, J. D., and Pinello, L. (2019). Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome biology*, 20(1):1–25.

Chen, L., Wang, C., Qin, Z. S., and Wu, H. (2015). A novel statistical method for quantitative comparison of multiple chip-seq datasets. *Bioinformatics*, 31(12):1889–1896.

Chen, X., Miragaia, R. J., Natarajan, K. N., and Teichmann, S. A. (2018). A rapid and robust method for single cell chromatin accessibility profiling. *Nature Communications*, 9(1):1–9.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117.

Chen, Y., Negre, N., Li, Q., Mieczkowska, J. O., Slattery, M., Liu, T., Zhang, Y., Kim, T.-K., He, H. H., Zieba, J., et al. (2012). Systematic evaluation of factors influencing chip-seq fidelity. *Nature methods*, 9(6):609.

Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G., and Reik, W. (2016). Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome biology*, 17(1):72.

Clouaire, T., Webb, S., and Bird, A. (2014). Cfp1 is required for gene expression-dependent h3k4 trimethylation and h3k9 acetylation in embryonic stem cells. *Genome biology*, 15(9):451.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., et al. (2010). Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936.

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., et al. (2018). A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, 174(5):1309–1324.

Cuscò, P. and Filion, G. J. (2016). Zerone: a ChIP-seq discretizer for multiple replicates with built-in quality control. *Bioinformatics*, 32(19):2896–2902.

Dunham, I., Kundaje, A., Aldred, S., Collins, P., Davis, C., Doyle, F., Epstein, C., Frietze, S., Harrow, J., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160.

ElTanbouly, M. A., Zhao, Y., Nowak, E., Li, J., Schaafsma, E., Le Mercier, I., Ceeraz, S., Lines, J. L., Peng, C., Carriere, C., et al. (2020). Vista is a checkpoint regulator for naïve t cell quiescence and peripheral tolerance. *Science*, 367(6475).

Ernst, J. and Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–216.

Fang, R., Preissl, S., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A. K., Mukamel, E. A., Zhang, Y., Behrens, M. M., et al. (2019). Fast and accurate clustering of single cell epigenomes reveals cis-regulatory elements in rare cell types. *bioRxiv*, page 615179.

Feng, J., Wilkinson, M., Liu, X., Purushothaman, I., Ferguson, D., Vialou, V., Maze, I., Shao, N., Kennedy, P., Koo, J., et al. (2014). Chronic cocaine-regulated epigenomic changes in mouse nucleus accumbens. *Genome biology*, 15(4):R65.

Fletcher, R. (2013). *Practical methods of optimization*. John Wiley & Sons.

Gill, G. (2001). Regulation of the initiation of eukaryotic transcription. *Essays in biochemistry*, 37:33–44.

González-Blas, C. B., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., and Aerts, S. (2019). cistopic: cis-regulatory topic modeling on single-cell atac-seq data. *Nature methods*, 16(5):397–400.

Granja, J. M., Klemm, S., McGinnis, L. M., Kathiria, A. S., Mezger, A., Corces, M. R., Parks, B., Gars, E., Liedtke, M., Zheng, G. X., et al. (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nature Biotechnology*, 37(12):1458–1465.

Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., Dahmani, A., Lameiras, S., Reyal, F., Frenoy, O., et al. (2019). High-throughput single-cell chip-seq identifies heterogeneity of chromatin states in breast cancer. *Nature genetics*, 51(6):1060–1066.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589.

Hocking, T. D., Goerner-Potvin, P., Morin, A., Shao, X., Pastinen, T., and Bourque, G. (2016). Optimizing chip-seq peak detectors using visual labels and supervised machine learning. *Bioinformatics*, 33(4):491–499.

Huang, T., Lin, C., Zhong, L. L., Zhao, L., Zhang, G., Lu, A., Wu, J., and Bian, Z. (2017). Targeting histone methylation for colorectal cancer. *Therapeutic advances in gastroenterology*, 10(1):114–131.

Ibrahim, M. M., Lacadie, S. A., and Ohler, U. (2014). JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics*, 31(1):48–55.

Jebara, T., Song, Y., and Thadani, K. (2007). Spectral clustering and embedding with hidden markov models. In *European Conference on Machine Learning*, pages 164–175. Springer.

Ji, H., Li, X., Wang, Q.-f., and Ning, Y. (2013). Differential principal component analysis of chip-seq. *Proceedings of the National Academy of Sciences*, 110(17):6789–6794.

Jia, G., Preussner, J., Chen, X., Guenther, S., Yuan, X., Yekelchyk, M., Kuenne, C., Looso, M., Zhou, Y., Teichmann, S., et al. (2018). Single cell rna-seq and atac-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nature communications*, 9(1):1–17.

Jones, P. A., Issa, J.-P. J., and Baylin, S. (2016). Targeting the cancer epigenome for therapy. *Nature reviews Genetics*, 17(10):630.

Jung, Y. L., Luquette, L. J., Ho, J. W., Ferrari, F., Tolstorukov, M., Minoda, A., Issner, R., Epstein, C. B., Karpen, G. H., Kuroda, M. I., et al. (2014). Impact of sequencing depth in ChIP-seq experiments. *Nucleic acids research*, 42(9):e74–e74.

Kagohara, L. T., Zamuner, F., Davis-Marcisak, E. F., Sharma, G., Considine, M., Allen, J., Yegnasubramanian, S., Gaykalova, D. A., and Fertig, E. J. (2020). Integrated single-cell and bulk gene expression and atac-seq reveals heterogeneity and early changes in pathways associated with resistance to cetuximab in hnscc-sensitive cell lines. *British Journal of Cancer*, pages 1–13.

Kim, J., Lee, Y., Lu, X., Song, B., Fong, K.-W., Cao, Q., Licht, J. D., Zhao, J. C., and Yu, J. (2018). Polycomb- and methylation-independent roles of ezh2 as a transcription activator. *Cell reports*, 25(10):2808–2820.

Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282.

Koues, O. I., Kowalewski, R. A., Chang, L.-W., Pyfrom, S. C., Schmidt, J. A., Luo, H., Sandoval, L. E., Hughes, T. B., Bednarski, J. J., Cashen, A. F., et al. (2015). Enhancer sequence variants and transcription-factor deregulation synergize to construct pathogenic regulatory circuits in b-cell lymphoma. *Immunity*, 42(1):186–198.

Kuan, P. F., Chung, D., Pan, G., Thomson, J. A., Stewart, R., and Keleş, S. (2011). A statistical framework for the analysis of chip-seq data. *Journal of the American Statistical Association*, 106(495):891–903.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., et al. (2012). Chip-seq guidelines and practices of the encode and modencode consortia. *Genome research*, 22(9):1813–1831.

Latchman, D. S. (1997). Transcription factors: an overview. *The international journal of biochemistry & cell biology*, 29(12):1305–1312.

Lauberth, S. M., Nakayama, T., Wu, X., Ferris, A. L., Tang, Z., Hughes, S. H., and Roeder, R. G. (2013). H3k4me3 interactions with taf3 regulate preinitiation complex assembly and selective gene activation. *Cell*, 152(5):1021–1036.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, J., Moazed, D., and Gygi, S. P. (2002). Association of the histone methyltransferase set2 with rna polymerase ii plays a role in transcription elongation. *Journal of Biological Chemistry*, 277(51):49383–49388.

Liang, K. and Keleş, S. (2011). Detecting differential binding of transcription factors with chip-seq. *Bioinformatics*, 28(1):121–122.

Liu, X., Wang, C., Liu, W., Li, J., Li, C., Kou, X., Chen, J., Zhao, Y., Gao, H., Wang, H., et al. (2016). Distinct features of h3k4me3 and h3k27me3 chromatin domains in pre-implantation embryos. *Nature*, 537(7621):558.

Lu, C., Jain, S. U., Hoelper, D., Bechet, D., Molden, R. C., Ran, L., Murphy, D., Venneti, S., Hameed, M., Pawel, B. R., et al. (2016). Histone h3k36 mutations promote sarcomagenesis through altered histone methylation landscape. *Science*, 352(6287):844–849.

Lun, A. T. and Smyth, G. K. (2014). De novo detection of differentially bound regions for chip-seq data using peaks and windows: controlling error rates correctly. *Nucleic acids research*, 42(11):e95–e95.

Lun, A. T. and Smyth, G. K. (2015). csaw: a bioconductor package for differential binding analysis of chip-seq data using sliding windows. *Nucleic acids research*, 44(5):e45–e45.

Ma, P., Castillo-Davis, C. I., Zhong, W., and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic acids research*, 34(4):1261–1269.

Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27(12):1696–1697.

Mammana, A. and Helmuth, J. (2016). bamsignals: Extract read count signals from bam files. *R package version*, 1(3).

Margueron, R. and Reinberg, D. (2011). The polycomb complex prc2 and its mark in life. *Nature*, 469(7330):343.

Nelson, D. L., Lehninger, A. L., and Cox, M. M. (2008). *Lehninger principles of biochemistry*. Macmillan.

Ngollo, M., Lebert, A., Dagdemir, A., Judes, G., Karsli-Ceppioglu, S., Daures, M., Kemeny, J.-L., Penault-Llorca, F., Boiteux, J.-P., Bignon, Y.-J., et al. (2014). The association between histone 3 lysine 27 trimethylation (H3K27me3) and prostate cancer: relationship with clinicopathological parameters. *BMC cancer*, 14(1):994.

Niu, W., Lu, Z. J., Zhong, M., Sarov, M., Murray, J. I., Brdlik, C. M., Janette, J., Chen, C., Alves, P., Preston, E., et al. (2011). Diverse transcription factor binding features revealed by genome-wide chip-seq in c. elegans. *Genome research*, 21(2):245–254.

Park, P. J. (2009). Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417.

Pfister, S. X., Markkanen, E., Jiang, Y., Sarkar, S., Woodcock, M., Orlando, G., Mavrommati, I., Pai, C.-C., Zalmas, L.-P., Drobnitzky, N., et al. (2015). Inhibiting wee1 selectively kills histone h3k36me3-deficient cancers by dntp starvation. *Cancer cell*, 28(5):557–568.

Portela, A. and Esteller, M. (2010). Epigenetic modifications and human disease. *Nature biotechnology*, 28(10):1057.

Powell, M. J. (2009). The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, pages 26–46.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W., and Lieb, J. D. (2011). ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome biology*, 12(7):R67.

Rashid, N. U., Sun, W., and Ibrahim, J. G. (2014). Some statistical strategies for DAE-seq data analysis: variable selection and modeling dependencies among observations. *Journal of the American Statistical Association*, 109(505):78–94.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, 4(8):651.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):R25.

Robinson, M. D. and Smyth, G. K. (2007). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332.

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., Brown, G. D., Gojis, O., Ellis, I. O., Green, A. R., et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481(7381):389.

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., and Bernstein, B. E. (2015). Single-cell chip-seq reveals cell subpopulations defined by chromatin state. *Nature biotechnology*, 33(11):1165–1172.

Sanderson, C. and Curtin, R. (2016). Armadillo: a template-based c++ library for linear algebra. *Journal of Open Source Software*, 1(2):26.

Schep, A. N., Wu, B., Buenrostro, J. D., and Greenleaf, W. J. (2017). chromvar: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature methods*, 14(10):975–978.

Shen, L., Shao, N.-Y., Liu, X., Maze, I., Feng, J., and Nestler, E. J. (2013). diffreps: detecting differential chromatin modification sites from chip-seq data with biological replicates. *PloS one*, 8(6):e65598.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). Ctcf-promoted rna polymerase ii pausing links dna methylation to splicing. *Nature*, 479(7371):74.

Smyth, P. (1997). Clustering sequences with hidden markov models. In *Advances in neural information processing systems*, pages 648–654.

Soneson, C., Love, M. I., and Robinson, M. D. (2015). Differential analyses for rna-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4.

Song, Q. and Smith, A. D. (2011). Identifying dispersed epigenomic domains from chip-seq data. *Bioinformatics*, 27(6):870–871.

Spyrou, C., Stark, R., Lynch, A. G., and Tavaré, S. (2009). Bayespeak: Bayesian analysis of chip-seq data. *BMC bioinformatics*, 10(1):299.

Stark, R. and Brown, G. (2011). Diffbind: differential binding analysis of chip-seq peak data. *R package version*, 100:4–3.

Starmer, J. and Magnuson, T. (2016). Detecting broad domains and narrow peaks in chip-seq data with hiddendomains. *BMC bioinformatics*, 17(1):144.

Steinhauser, S., Kurzawa, N., Eils, R., and Herrmann, C. (2016). A comprehensive comparison of tools for differential chip-seq analysis. *Briefings in bioinformatics*, 17(6):953–966.

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nature methods*, 5(9):829.

Varambally, S., Dhanasekaran, S. M., Zhou, M., Barrette, T. R., Kumar-Sinha, C., Sanda, M. G., Ghosh, D., Pienta, K. J., Sewalt, R. G., Otte, A. P., et al. (2002). The polycomb group protein ezh2 is involved in progression of prostate cancer. *Nature*, 419(6907):624.

Vermunt, J. K., Tran, B., and Magidson, J. (2008). Latent class models in longitudinal research. *Handbook of longitudinal research: Design, measurement, and analysis*, pages 373–385.

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269.

Wei, Y., Xia, W., Zhang, Z., Liu, J., Wang, H., Adsay, N. V., Albarracin, C., Yu, D., Abbruzzese, J. L., Mills, G. B., et al. (2008). Loss of trimethylation at lysine 27 of histone h3 is a predictor of poor outcome in breast, ovarian, and pancreatic cancers. *Molecular Carcinogenesis: Published in cooperation with the University of Texas MD Anderson Cancer Center*, 47(9):701–706.

Westbrook, T. F., Martin, E. S., Schlabach, M. R., Leng, Y., Liang, A. C., Feng, B., Zhao, J. J., Roberts, T. M., Mandel, G., Hannon, G. J., et al. (2005). A genetic screen for candidate tumor suppressors identifies rest. *Cell*, 121(6):837–848.

Wheaton, K., Atadja, P., and Riabowol, K. (1996). Regulation of transcription factor activity during cellular aging. *Biochemistry and cell biology*, 74(4):523–534.

Wu, H. and Ji, H. (2014). PolyaPeak: detecting transcription factor binding sites from ChIP-seq using peak shape information. *PLoS One*, 9(3):e89694.

Xing, H., Mo, Y., Liao, W., and Zhang, M. Q. (2012). Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. *PLoS computational biology*, 8(7):e1002613.

Xiong, L., Xu, K., Tian, K., Shao, Y., Tang, L., Gao, G., Zhang, M., Jiang, T., and Zhang, Q. C. (2019). Scale method for single-cell atac-seq analysis via latent feature extraction. *Nature communications*, 10(1):1–10.

Xu, H., Handoko, L., Wei, X., Ye, C., Sheng, J., Wei, C.-L., Lin, F., and Sung, W.-K. (2010). A signal–noise model for significance analysis of chip-seq with negative control. *Bioinformatics*, 26(9):1199–1204.

Yang, Y., Fear, J., Hu, J., Haecker, I., Zhou, L., Renne, R., Bloom, D., and McIntyre, L. M. (2014). Leveraging biological replicates to improve analysis in ChIP-seq experiments. *Computational and structural biotechnology journal*, 9(13):e201401002.

Young, M. D., Willson, T. A., Wakefield, M. J., Trounson, E., Hilton, D. J., Blewitt, M. E., Oshlack, A., and Majewski, I. J. (2011). Chip-seq analysis reveals distinct h3k27me3 profiles that correlate with transcriptional activity. *Nucleic acids research*, 39(17):7415–7427.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., et al. (2017). Ensembl 2018. *Nucleic acids research*, 46(D1):D754–D761.

Zhang, Y., Lin, Y.-H., Johnson, T. D., Rozek, L. S., and Sartor, M. A. (2014). Pepr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated chip-seq data. *Bioinformatics*, 30(18):2568–2575.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9):R137.

Zucchini, W., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC.