Solving the Differential Peak Calling Problem in ChIP-seq Data

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

M.Sc. Manuel Allhoff

aus Soest

Berichter: Universitätsprofessor Dr. rer. nat. Thomas Berlage
Universitätsprofessor Dr. rer. nat. Martin Zenke
Universitätsprofessor Dr. rer. pol. Matthias Jarke

Tag der mündlichen Prüfung: 7. Juni 2016

Acknowledgement

First of all, I thank Dr. Ivan G. Costa for giving me the opportunity to work with him. I could not imagine a better supervisor. He supported me with helpful ideas and always found time for profitable conversations.

Furthermore, I want to thank Prof. Dr. Thomas Berlage, Prof. Dr. Martin Zenke and Prof. Dr. Matthias Jarke for instantly agreeing to become the examiners of this thesis, as well as Prof. Dr Martin Grohe and Prof. Dr. Horst Lichter for completing the examination committee.

I very much enjoyed the time at the Computational Biology Group of the university clinic of Aachen and at the graduate school AICES of the RWTH Aachen. Many thanks go to my colleagues, in particular to Eduardo G. Gusmao, Joseph Kuo, Fabio Ticconi and Marcus Schmidt for sharing thoughts, providing feedback and having fun in the office.

Moreover, I am thankful to Janine Mergel, Angela Knappstein, Dawid Kopetzki, René Grzeszick, Robert O'Connor and Iris Tewes for proof-reading parts of my thesis.

I want to thank my family and my friends for the moral support. In particular, I am very grateful to my parents who give me the best possible start in life and support me in every situation. Finally, I want to thank my girlfriend Christina for all her support and love.

Selbstständigkeitserklärung

Ich versichere hiermit an Eides statt, dass ich die vorliegende Doktorarbeit selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich habe die Grundsätze zur Sicherung guter wissenschaftlicher Praxis der RWTH Aachen zur Kenntnis genommen und eingehalten.

Aachen, 9.2.2016 ______________________________
(Manuel Allhoff)

Publications

As required by §5(3) of the

> *Promotionsordnung für die Fakultät für Mathematik, Informatik und Naturwissenschaften der Rheinisch-Westfälischen Technischen Hochschule Aachen vom 27.09.2010 in der Fassung der zweiten Ordnung zur Änderung der Promotionsordnung vom 30.06.2014 (veröffentlicht als Gesamtfassung),*

a declaration of results that are published by the author as well as particular contributions to co-authored articles follows.

I am the main author of the following articles which are co-authored with Ivan G. Costa and our experimental collaborators Kristin Seré, Heike Chauvistré, Qiong Lin and Martin Zenke from the cell biology group of the university clinic Aachen. As my adviser, Ivan G. Costa supported me in all stages of research.

> Manuel Allhoff, Kristin Seré, Heike Chauvistré, Qiong Lin, Martin Zenke, and Ivan G. Costa. *Detecting differential peaks in ChIP-seq signals with ODIN*. Bioinformatics, Volume 30, Issue 24, 3467-3475, 2014

> Manuel Allhoff, Juliana F. Pires, Kristin Seré, Martin Zenke, and Ivan G. Costa. *Differential Peak Calling of ChIP-seq Signals with Replicates with THOR*. Genome Biology, *under review*

I implemented ODIN, THOR as well as the simulation algorithm. Moreover, I performed all data analysis experiments. All chapters of this thesis have been written by me. Only Section 4.6.1 and Section 5.2.6 (Identifying rSNPs) are equally shared by Juliana F. Pires and me.

I am also (co-)author of the following articles, which did not directly contribute to this thesis, but shaped a general understanding of next-generation sequencing analysis.

> Manuel Allhoff, Alexander Schönhuth, Marcel Martin, Ivan G. Costa, Sven Rahmann, and Tobias Marschall. *Discovering motifs that induce sequencing errors*. BMC Bioinformatics, 14(Suppl 5):S1, 2013

> Eduardo G. Gusmao, Manuel Allhoff, Martin Zenke and Ivan G. Costa. *Analysis of computational footprinting methods for DNase sequencing experiments*. Nature Methods, Volume 13, Issue 4, 303-309, 2016

List of Figures

List of Tables

Abbreviations

AUC	area under the curve
bp	base pair
CB	combined binomial
CC	proliferative centroblast
cDC	classical DC
CDP	common DC progenitors
ChIP-seq	chromatin immunoprecipitation followed by sequencing
CO	epigenomics effects of cocaine study
DAGE	differential average gene expression
DC	dendritic cell
DC	dendritic cell differentiation study
DCA	differential correlation analysis
DP	differential peak
DPC	differential peak caller
EGA	european genome-phenome archive
FL	follicular lymphoma
FPR	false positive rate
FRiP	fraction of reads in peaks
GMM	gaussian mixture model
HK	housekeeping genes
HMM	hidden Markov models
IDR	irreproducible discovery rate
LYMP	B cell lymphoma study
MHC	major histocompatibility complex
MM	monocyte and macrophages study
MMP	M-value of peak-associated bins
MPP	multipotent progenitors
NGS	next-generation sequencing
PBBA	peripheral blood B
pDC	plasmacytoid DC
ROC	receiver operating characteristic
rSNPs	regulatory single nucleotide polymorphisms
SDS	sequencing depth scaling
SPC	single signal peak caller
TF	transcription factor
TFBS	transcription factor binding site
TMM	trimmed mean of M-values
TPR	true positive rate

Contents

CHAPTER 1

Introduction

Gene expression is the process of selectively reading genetic information and it describes a life-essential mechanism in all known living organisms. Key players in the regulation of gene expression are proteins that interact with DNA. For example, histone proteins can be modified (histone modifications) and locally influence the DNA's accessibility. The DNA's accessibility is a crucial feature for transcription factor proteins, as they bind to the DNA. DNA-protein interaction sites are nowadays analyzed in a genome wide manner with chromatin immunoprecipitation followed by sequencing (ChIP-seq).

ChIP is a complex multistep protocol that provides millions of short DNA fragments covering the regions around the protein-DNA interaction sites. The subsequent sequencing step produces DNA strings (reads) of the beginning or the end of these fragments. The information of the reads' positions, which is associated with the positions of the protein-DNA interaction sites, gets lost during the sequencing process. Sophisticated string search algorithms have to be applied to reconstruct these positions by mapping the reads back to a reference genome. As the mapped reads only partially describe the original DNA fragments, the exact positions of the DNA-protein interaction sites have to be determined.

To each genomic location, a discrete value is assigned, which corresponds to the number of reads that cover this position. The number of reads corresponds to the strength of the protein binding event. Peaks, that is, regions with a signal higher than expected by chance, describe to the protein-DNA interaction sites. Detecting such peaks is the fundamental computational challenge in the ChIP-seq analysis. The great majority of published computational tools have concentrated on the detection of peaks in a single ChIP-seq signal. As in every complex wet lab protocol, ChIP-seq contains a wide range of potential biases. To reduce the effect of unwanted biases, ChIP-seq experiments are often replicated, which helps to distinguish between biological and random events and to verify the reliability of all experimental steps. Complex ChIP-seq based studies emphasize the demand of methods to compare replicated ChIP-seq signals which are associated with distinct biological conditions. For example, the detection of histone changes for distinct cellular conditions is an outstanding crucial problem in current biological and medical research which leads to a deeper understanding of gene expression regulation. For example,

- cancer can exhibit histone changes which affect gene expression. Koues et al. (2015) for instance analyze regulatory features including histone changes between lymphoma patients and a control group with healthy individuals.

- histone states have a high impact on cell differentiation and play therefore a key role in biological processes. Lin et al. (2015) investigate regulatory changes during the development of antigen-presenting dendritic cells.

- changes in histones lead to cell activation. Saeed et al. (2014) for instance explore how monocyte cells are activated to macrophages cells which play a key role in the defence system of the organism.

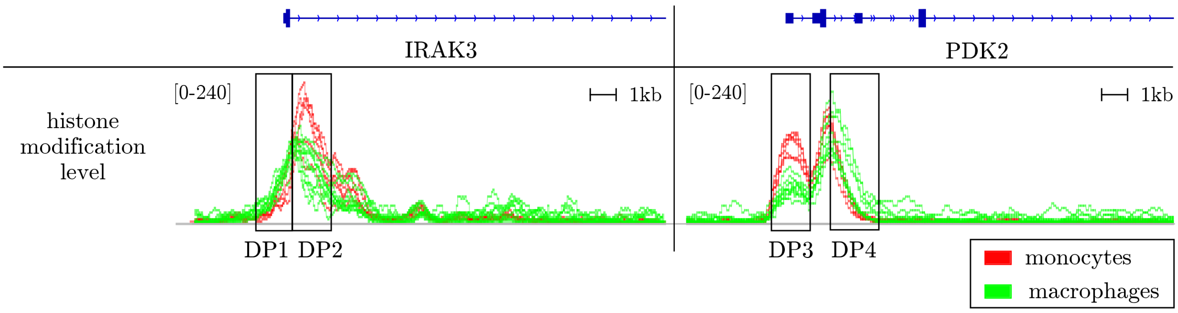


Figure 1.1.: Example for changes in histone modification levels of monocyte and macrophage cells from the study of Saeed et al. (2014). With ChIP-seq it becomes possible to assign to each genomic position a discrete value that reflects the strength of the DNA-protein binding. Replicated ChIP-seq signals are shown as line plots for the monocytes (red) and macrophages (green) cells for the regions where the genes IRAK3 and PDK2 are located in the human genome. Differential peaks between the profiles of monocyte and macrophage cells are indicated by black boxes: DP2 and DP3 gain monocyte and DP1 and DP4 gain macrophage signal. An increasing signal in this particular histone modification (H3K4me3) usually leads to an increase in the gene expression.

All these studies are about the comparison of histone modification levels from distinct experimental conditions. Figure 1.1 gives an example for the replicated ChIP-seq signal based on histone modifications of monocyte and macrophage cells from the study of Saeed et al. (2014). Differential peaks (DPs) between the green and red signal correspond to changes of certain histone modifications. Several computational challenges arise when detecting DPs:

- the shape of ChIP-seq peaks depends on the underlying protein of interest. For ChIP-seq data of histone modifications, the DNA-protein interactions occur in mid-size to large domains. Here, domains can span several hundreds of base pairs and may have intricate patterns of gains and losses of ChIP-seq signals within the same domain. In contrast, ChIP-seq from transcription factors mostly happens in small isolated peaks.

- artefacts, which arise due to the complexity of the ChIP-seq protocol, produce signals with distinct signal-to-noise ratios, even when they are produced in the same lab and follow the same protocols (Furey, 2012; Meyer and Liu, 2014). Furthermore, different sequencing depths between samples aggravate the comparison of their ChIP-seq signal. Hence, a robust normalization method for the ChIP-seq signals is required.

- clinical samples, where patients have a distinct genetic background, introduce further variation to the distinct ChIP-seq signals (Ashoor et al., 2013). Moreover, replicated ChIP-seq experiments introduce further complexity which has to be reflected by the use of sophisticated statistical models.

Current differential peak calling methods fail to cover all listed challenges. They apply heuristic signal segmentation strategies, such as window-based approaches, to identify DPs. There are only a few attempts to normalize ChIP-seq data. Furthermore, most methods do not support replicates. Hence, there is a clear need for computational methods that address all challenges.

In this thesis, we propose ODIN and THOR, algorithms to determine changes of protein-DNA complexes for distinct cellular conditions in ChIP-seq experiments without and with replicates. In particular, our algorithms detect DPs in the above described studies from Lin et al. (2015), Koues et al. (2015) and Saeed et al. (2014). Our methods address all described challenges. We apply a statistical model (hidden Markov model) to call DPs and to handle replicates. We also introduce a novel normalization strategy which is based on control regions. These features lead to comprehensive algorithms that accurately call DPs in ChIP-seq signals.

Moreover, the evaluation of differential peak calling algorithms is an open problem. The research community lacks both a direct metric to rate the algorithms and data sets with a genome wide map of DNA-protein interaction sites which can serve as gold standards. We propose two alternative approaches for the evaluation. First, we present indirect metrics to quantify DPs by taking advantage of gene expression data and second, we use simulation to customize artificial gold standards.

1.1 Organization of the Thesis

In Chapter 2, we first explain the biological and technical background. Second, we formalize the differential peak calling problem. Finally, we discuss the previous computational work done on this field and formulate specific goals of this thesis. In Chapter 3, we explain our method to call differential peaks in ChIP-seq signals. First, we give details about the preprocessing pipeline which is necessary for the ChIP-seq analysis. Next, we explain the differential peak calling procedure. We conclude with the postprocessing pipeline of our method. In Chapter 4, we describe the experiments performed in this thesis. First, we propose an algorithm to simulate ChIP-seq reads that contain differential peaks. Second, we describe our evaluation approach which is based on gene expression data. We also list the biological data used for our experiments and give a short introduction to the statistical test for the evaluation. Finally, we detail the experiments we perform to evaluate differential peak calling methods. In Chapter 5, we give an overview of the results we have achieved with our experiments. We distinguish between our methods THOR that takes replicates into account and ODIN that does not take replicates into account. The final Chapter 6 contains a concluding discussion and an outlook to future work.

CHAPTER 2

Background

We first introduce the fundamental biological concepts of DNA, gene expression and epigenetics. We then explain ChIP-seq, a method to investigate epigenetics by identifying DNA-protein complexes in a genome-wide manner. Next, we detail computational aspects of the ChIP-seq data analysis. We introduce the peak calling problem on a single ChIP-seq signal. The extension of this problem is the differential peak calling problem, which is extensively addressed in this thesis. We motivate the differential peak calling with current biological and medical studies and point out arising challenges. Next, we discuss previous work that is related to this thesis. Finally, we formulate the aims of the thesis.

2.1 Biology

We give an overview of the biological concepts that are necessary to understand the thesis. See Alberts et al. (2002) for a more detailed description of molecular biology and, in particular, for a gentle introduction to DNA. Lodish et al. (2007) give a detailed introduction to gene expression and Allis et al. (2007) explain epigenetics in detail.

2.1.1. DNA

Deoxyribonucleic acid (DNA) is the carrier of genetic information of living organisms. DNA is a chain molecule with nucleotides as elements. While the phosphate group and the sugar molecule are similar, the third element of a nucleotide, the nucleobase (or base), varies. We therefore can describe a DNA strand by its bases adenine (A), cytosine (C), guanine (G) and thymine (T). DNA is directional, that is, it has a 5′ and a 3′ end. The nucleotides adenine and thymine as well as cytosine and guanine can pair and form a double-stranded structure. Both strands are coiled around each other and build the typical double helix. The strands are reverse complements of each other.

2.1.2. Gene Expression

Gene expression is the process of selectively reading genetic information contained in the DNA. Processing the genetic information works in two steps: first, DNA is translated into RNA molecules, and second, the RNA is translated into proteins. In the first step, a protein complex called RNA polymerase II binds to the DNA and successively reads the genetic information of a DNA molecule (transcription). Figure 2.1 shows the concept of gene transcription. The RNA polymerase complex (1) attaches to the promoter of gene X, (2) locally separates the two DNA strands, (3) creates an RNA molecule by reading one nucleotide at a time and (4) finally reconnects the two DNA strands. In the second step, a certain protein translates the RNA to protein molecules (translation). Proteins are life-essential molecules which contribute to the structural components of a cell and perform all activities within a cell. Various control mechanisms of a cell facilitate production of proteins on demand. For example, in Figure 2.1 certain proteins, so called transcription factors (TFs), attach to the DNA and may effect the rate of transcription initiation.

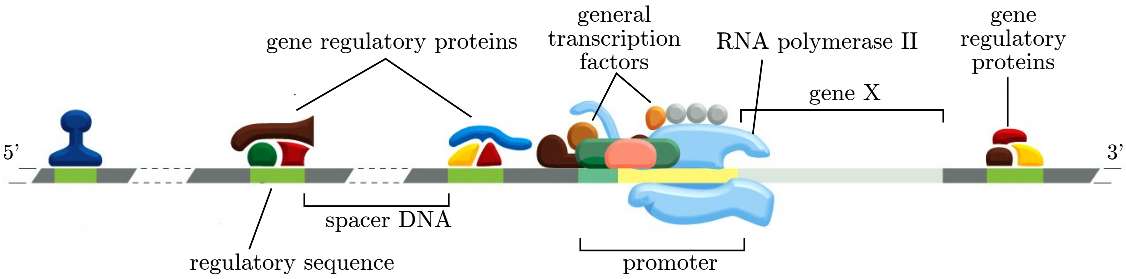


Figure 2.1.: Gene transcription. In this example the transcription of gene X is shown. The RNA polymerase II attaches to the gene promoter, a DNA sequence that is upstream located to the gene. Also, general transcription factors bind to the promoter and help the polymerase to position properly at the promoter. Several control mechanisms determine the gene transcription. For example, gene regulatory proteins such as TFs bind to regulatory sequences and effect the rate of transcription initiation. Regulatory sequences, also called enhancer regions, may either be located close to the promoter, far upstream or close downstream of the gene. The figure is based on Alberts et al. (2002).

2.1.3. Epigenetics

Epigenetics investigates changes in gene expression by mechanisms other than variation in the DNA sequence such as the chromatin organization. Chromatin is a macromolecule that helps packaging DNA and proteins to make them fit within the cell. It also serves as an index system to organise the genome. Figure 2.2 depicts the concept of chromatin. There are two chromatin states, open and closed chromatin, which facilitate the DNA to be more or less compact. Hence, the states effect the gene expression, because for genomic regions with close chromatin the DNA is less accessible for TFs compared to regions with open chromatin. Chromatin states play a key role for example in cell differentiation by allowing the selective expression of particular genes.

A nucleosome is the fundamental core unit of chromatin and consists of eight histones. Histones are proteins where the DNA is wrapped around to be spatially organised in a maximal condensed way. There are two types of histones: core histones form the nucleosome and linker histones bind the nucleosome to the DNA. Amino-terminal histone tails drive through the nucleosome core and make contact with adjacent nucleosomes to build the chromatin structure.

Enzymes may chemically modify the histone tails which then affect the overall chromatin structure. The enzyme adds a chemical flag to the histone tails comparable to DNA methylation or the chromatin remodelling processes. We refer to a histone with a particular chemical flag as histone modification. Importantly, some histone modifications effect the chromatin which effects the gene expression. Particular histone modifications are therefore associated for example to activation or deactivation of gene expression (see Figure 2.2).

The naming of a histone modifications follow the following structure: the histone number in a nucleosome whose tail is modified, the single-letter amino acid abbreviation, the amino acid position in the protein, the type of the modification and the number of modifications. For instance, the histone modification H3K4me3 describes a chemical modification of histone three (H3), where the amino acid lysine at the fourth position (K4) is changed by adding three methyl groups (me3). In the following we list histone modifications which are considered in this thesis. Histone modifications H3K79me2 and H3K36me3 are associated with transcription (Nguyen and Zhang, 2011; Sims Iii and Reinberg, 2009). Histone modification H3K4me3, typically located in the promoter, H3K4me1, typically located in enhancer regions, as well as H3K27ac and H3K9ac are associated with activation (Briggs et al., 2001; Creyghton et al., 2010; Grant et al., 1999). Histone modification H3K9me3 and H3K27me3 are associated with repression (Cao et al., 2002).

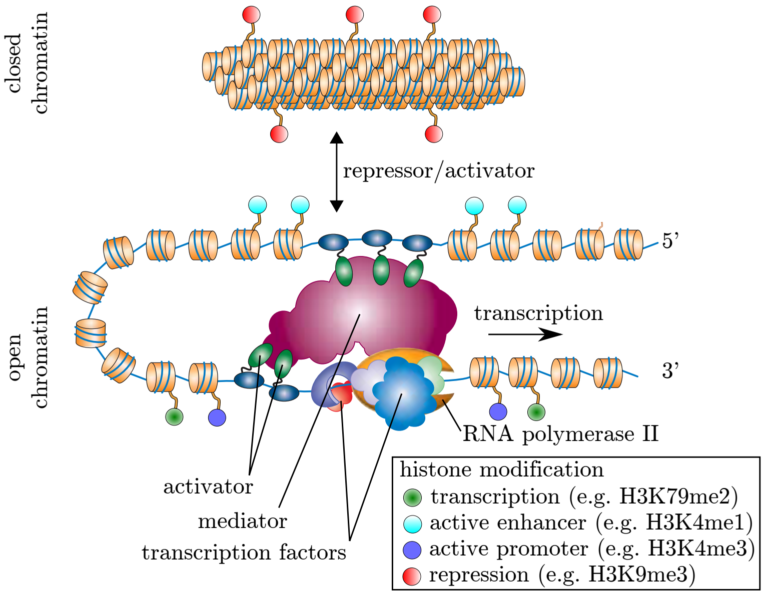


Figure 2.2.: Epigenetic concept. Among others, gene expression is regulated by closed and open chromatin. Closed chromatin exhibits repressive histone modifications and inhibits RNA ploymerases II to attach to the DNA. In contrary, open chromatin has active histone marks as well as certain activator proteins. Activator proteins are gene regulatory proteins which are also shown in Figure 2.1. They interact with mediator proteins and enable transcription factors as well as RNA polymerase to bind to the DNA. RNA polymerase reads the DNA from 5′ to 3′ end. Thereby, RNA molecules are produced which eventually are translated to proteins. Histone modifications are also associated with DNA transcription. The figure is based on Lodish et al. (2007).

2.2 ChIP-seq to Analyze Epigenetics

We introduce chromatin immunoprecipitation followed by sequencing (ChIP-seq), a method to identify DNA-protein complexes in a genome-wide manner. First, we explain ChIP which is a method to isolate DNA fragments that are attached to certain proteins of interest. We then explain DNA sequencing which is used to determine the nucleotide sequence of a DNA molecule and its position with regard to a reference genome. Finally, we describe ChIP-seq which combines ChIP with DNA sequencing. Further, we emphasize important challenges that have to be considered in the analysis of ChIP-seq experiments.

2.2.1. ChIP

Chromatin immunoprecipitation (ChIP) is used to investigate protein-DNA interactions inside a cell. In particular, ChIP enables localization of posttranslational modifications of the histone tails (see Section 2.1.3) as well as DNA target sites of TFs in the genome. The basic idea of ChIP was first reported in the 1960s, while applications of ChIP in studies of histone-DNA interactions go back to the late 1970s (Jackson, 1978; Collas, 2010).

The ChIP protocol has several steps. First, DNA and proteins in a cell are cross-linked with formaldehyde. These DNA-protein complexes are fragmented for example by sonication into fragments of $200 - 1000$bp. Specific antibodies are then used to pull out (immunoprecipitate) protein-DNA complexes that contain the proteins of interest. Finally, the cross-link is reversed (formaldehyde is heat-reversible) and the DNA that was bound to the protein is purified. To identify the DNA sequences associated with the proteins of interest, downstream analysis, such as DNA sequencing (see Figure 2.3), is required.

2.2.2. DNA sequencing

DNA sequencing methods determine the base sequence of a DNA sample. The first approach to sequence DNA was the chain-termination method invented by Sanger et al. (1977). However, routine studies of mammalians became possible by high-throughput sequencing technologies which are also called next-generation sequencing (NGS) or second-generation sequencing. NGS takes advantage of parallelization: nucleotides are read in parallel and a large number of DNA fragments is considered at the same time (Shendure and Ji, 2008). Parallelization is often done by cloning DNA fragments which is usually performed by PCR (Mullis and Faloona, 1987). Each DNA fragment can be sequenced from one end, resulting in single-end reads, or from both ends, resulting in paired-end reads. In most cases, the reads are shorter than the DNA fragments and therefore give only partially the base sequence of the fragment. Compared to Sanger sequencing, the costs of NGS are much lower, while the base calls are less accurate and the reads are smaller. NGS is the current method for large-scale sequencing applications. In many cases, a reference genome of the organism is known. Then, sequencing of organisms results in the computational problem to determine the position of each read in the genome. We refer to the process of estimating the read positions as aligning or mapping the reads to the reference genome. There are various technologies for DNA sequencing. All sequencing experiments analyzed in this thesis were performed with Illumina devices.

2.2.3. ChIP-seq Method

ChIP-seq combines the ChIP protocol with high-throughput sequencing technologies and thereby offers a low-cost way to identify DNA-protein interactions in a genome-wide manner (Park, 2009). ChIP-seq was one of the earliest applications of NGS (Johnson et al., 2007). Figure 2.3 gives an overview of the ChIP-seq workflow. First, the DNA obtained by ChIP (see Figure 2.3, step 1) and associated with the proteins of interest is sequenced by NGS methods (see Figure 2.3, step 2). NGS sequencing typically produces short ($\sim 50 - 100$bp) single-ended reads. The reads are then aligned to the reference genome (Park, 2009) (see Figure 2.3, step 3). Finally, a genomic signal is created based on the aligned reads (see Figure 2.3, step 4). Genomic regions where reads accumulate more than by chance (peaks) are identified within the ChIP-seq signal (peak calling). Peaks represent regions in the genome where the proteins of interest are localized in the original DNA sample (Collas, 2010).

The size of DNA fragments is larger than the protein-DNA interaction site. Therefore, reads derived from the DNA fragments map to different genomic locations which results in fuzzy peak shapes. Furthermore, peaks usually have different shapes due to the underlying proteins of interest taken into account by the ChIP protocol. TFs usually attach to small DNA regions without any further TF in the vicinity. The ChIP-seq landscape of TFs therefore tends to exhibit sharp, isolated peaks. In contrary, histone modifications are organised in groups where all histones are close to each other. In general, this leads to a complex ChIP-seq landscape with several peaks in close vicinity. Histone modifications at active regulatory elements exhibit relatively small protein domains. Broad domains are caused by a large number of proteins, which typically occur for histone modifications that repress genomic regions.

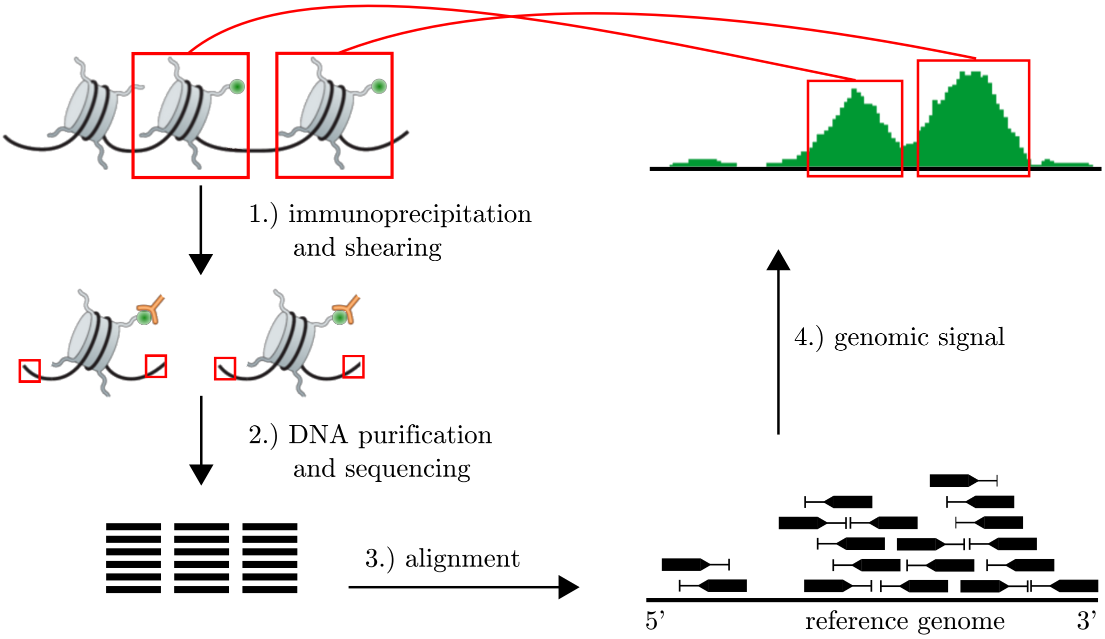


Figure 2.3.: ChIP-seq workflow. We show DNA that is wrapped around histones which may or may not be modified. The histone modifications are indicated by a green signal at their tails. First, ChIP is used to fetch out the proteins of interest with specific antibodies and to shear the DNA. Antibodies are represented as elements that are attached to the modified histones. Next, NGS methods are used to create reads which are mapped to a given reference genome. As the DNA fragment is not entirely sequenced, the reads stem either from the beginning or the end of the fragment. Reads can be mapped to the forward strand, that is, left oriented reads are mapped from the 5′ to 3′ end of the genome; or to the reverse strand, that is, right oriented reads are mapped from 3′ to 5′ end of the genome. Typically, a discrete signal is then derived from the set of aligned reads for the entire genome. It is a common procedure to extend the ChIP-seq reads to match the fragment size. Peaks indicated by red boxes in the ChIP-seq signal refer to positions in the genome where DNA interacts with the proteins of interest.

2.2.4. Control Sample

Each experimental step in the ChIP-seq protocol potentially involves various sources of artifacts. For example, DNA shearing usually does not result in a uniform fragment distribution of the genome, as open chromatin regions tend to be fragmented more easily than closed regions. Also, repetitive regions in the sample DNA may incorrectly seem to be enriched due to differences in the reference and sample genome (Park, 2009). Moreover, the DNA to be analyzed may be contaminated with DNA that was not bound by the chosen antibody. Therefore, it is highly recommended to compare a peak in a ChIP-seq profile to a control sample in the same cell in order to determine its significance (Meyer and Liu, 2014). There are three different ways to obtain control DNA:

- input-DNA: a fraction of the DNA sample is removed prior to the immunoprecipitation step;
- mock IP DNA: DNA is obtained from immunoprecipitation without an antibody; and

- DNA from non-specific immunoprecipitation, that is, the immunoprecipitation step is performed for a fraction of the sample DNA with an antibody that is known to not be involved in DNA bindings.

Input-DNA is the most widely used method as it is assumed to test against the most common artifacts introduced by the ChIP-seq protocol such as bias in the DNA fragmentation process (Park, 2009; Furey, 2012).

2.2.5. Arising Challenges

The ChIP-seq protocol is frequently refined to improve its accuracy (Meyer and Liu, 2014). Various challenges arising from both the technical and computational side have to be overcome to improve the peak detection. Here, we list the most important challenges associated with ChIP-seq.

Antibody

The antibody chosen for the ChIP experiment directly affects how well defined a peak in the ChIP-seq signal appears. Thus, the antibody's sensitivity and specificity is crucial for the analysis. High quality antibodies precisely pull out the proper protein-DNA complexes and thereby ensure a high level of enriched signal compared to the background noise. For some proteins there is no proper antibody, such that these proteins cannot be examined by ChIP-seq. Furthermore, the quality of the antibody may also depend on the manufacturer.

Cell Population

A typical ChIP experiment needs approximately 10^7 cells and thereby limits the number of ChIP experiments that can be performed on a biological sample. The number of cells depends on the quality of the antibody as well as on the abundance of the target protein (Furey, 2012). Some techniques (Acevedo et al., 2007; Adli and Bernstein, 2011) have been developed to decrease the number of required cells. However, fewer cells generally produce less well defined peaks in the resulting ChIP-seq signal.

Sequencing

DNA fragments obtained from ChIP are sequenced and mapped to a reference genome. The sequencing depth is crucial for the success of the ChIP-seq experiment. It is recommended to have approximately $2 \cdot 10^7$ reads for ChIP-seq experiments with the human genome and target proteins that lead to relatively isolated ChIP-seq peaks such as active histone marks. Apart from the absolute number of reads, it is recommended that the amount of reads mapping to distinct genomic location is higher than 80% (Furey, 2012).

There are genomic regions that cannot be captured in the ChIP-seq analysis (ENCODE Project Consortium, 2012). These regions comprise unstructured, high signals and occur independently of the type of NGS experiment. We refer to such regions as blacklisted regions. They are typically ignored in the ChIP-seq analysis.

Several studies address sequencing-specific bias (Khrameeva and Gelfand, 2012; Allhoff et al., 2013) which have to be taken into account by the mapping algorithm and the downstream analysis. Particularly important is the bias due to GC-content. Benjamini and Speed (2012) describe a dependency between the fragment count and the fragment's GC-content which may lead to artificial, non-biological high signals in the ChIP-seq experiment.

PCR Duplicates

PCR duplicates lead to artificially induced reads that impose unwanted bias on the downstream analysis. There are two ways how these duplicates are created. First, during PCR (see Section 2.2.2) of the sequencing procedure, duplicates may be created by accidentally considering several times the same fragment. Second, PCR duplicates may be created in the picture analyzing step during the sequencing process. That is, one DNA feature is mistaken as two or more features (Meyer and Liu, 2014; Maze et al., 2014). Both types of PCR duplicates lead to reads which are mapped to the same genomic location. However, because of sonication-based fragmentation, it is highly unlikely that two DNA fragments will stem from the same genomic location. Hence, reads with identical mapping positions indicate that they are PCR duplicates.

Fragment Size Estimation

ChIP-seq experiments typically comprise single-ended reads. For all DNA fragments, these reads are expected to come from on average uniform ratio of both DNA strands. Furthermore, the reads only partially cover one end of the fragments. Hence, the read distribution exhibits two peaks up- and downstream of the proteins of interest (see Figure 2.4). Typically, the reads are extended to the original fragment length, such that the read distribution provides a single peak that correlates with the protein position.

The fragment length can be obtained from the ChIP-seq protocol or be computationally estimated with the aligned ChIP-seq reads. Due to some expected difficulties in the protocol's accuracy, the fragment size is usually derived from the reads (Pepke et al., 2009). In Figure 2.3 and Figure 2.4, the extension size of reads is indicated by a distance arrow for each read. The extended reads are then used to generate the genomic signal (Step 4 in Figure 2.3).

2.2.6. Quality Measures of ChIP-seq Experiments

Successfully calling peaks in a ChIP-seq signal highly depends on the signal's signal-to-noise ratio (Landt et al., 2012). The ChIP-seq signal correlates to DNA-protein interaction sites. Signal that is not correlated to the these interaction sites is called background noise. Background noise may stem from various sources such as the fragmentation step or poor antibodies in the ChIP protocol. A high signal-to-noise ratio is highly desired for all downstream analyses of ChIP-seq data as it positively effects the accuracy of the peak calling step.

Landt et al. (2012) introduced several widely used measures to evaluate ChIP-seq data. These metrics give indications about the quality of the ChIP-seq experiment. First, the fraction of reads in peaks (FRiP) is an indicator for the signal-to-noise ratio in the data. The FRiP is estimated by calling peaks in a ChIP-seq signal and computing the ratio of reads within the called peaks and the overall number of reads. The higher the FRiP, the better the signal-to-noise ratio. See Figure 2.5 for an illustration of FRiP. Second, the non-redundant fraction (NRF) of reads denotes the ratio between the number of positions in the genome that uniquely mappable reads map to and the total number of uniquely mappable reads. NRF is associated with the number of PCR duplicates (see Section 2.2.5) and measures the entropy of the set of aligned reads. NRF decreases with sequencing depth as at some point PCR-amplified DNA fragment will be sequenced repeatedly.

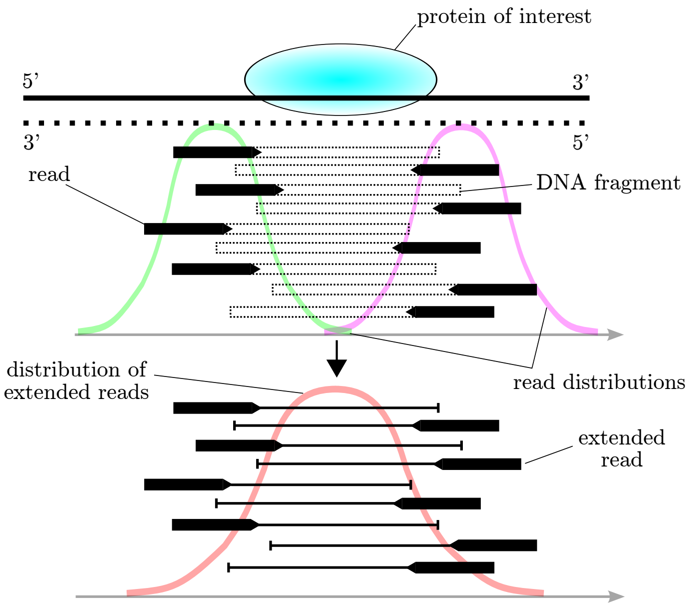


Figure 2.4.: Fragment size estimation. Reads are mapped to the forward strand (black line) or reverse strand (dotted black line) of the reference genome. The distributions of the reads (red and pink) build two peaks at the left and right side of the protein of interest, as only the beginning or the end of the DNA fragments is sequenced. The original DNA fragments are indicated as dotted lines which extend the reads. Because of the shearing process of the ChIP-seq protocol, the fragments slightly differ in their start positions. The fragmentation size is computed and the reads are artificially extended to obtain the original fragment length. The distribution of the extended reads (red) exhibits one peak whose position correlates with the position of the protein of interest. The figure is based on Park (2009).

2.3 Computational Analysis of ChIP-seq

Computational analysis is necessary to derive the positions of DNA-protein complexes from ChIP-seq data. First, we introduce the single signal peak calling problem. Next, we formalize the differential peak calling problem and motivate it by presenting related studies of current biological and medical research. Finally, we point out arising challenges.

2.3.1. Single Signal Peak Calling Problem

A common goal in ChIP-seq data analysis is the genome-wide detection of protein-DNA interactions in a single biological condition. The protein-DNA interaction positions are associated with peaks in a ChIP-seq profile. We refer to the detection of peaks in a ChIP-seq profile as the single signal peak calling problem.

Definition 2.1 (Single Signal Peak Calling Problem)**.** *For a given vector X describing a ChIP-seq profile, find genomic positions (peaks), where the signal is significantly enriched.*

To each genomic location, we assign a discrete value, which corresponds to the number of reads that cover this position. The number of reads corresponds to the strength of the protein binding event. Hence, we describe a ChIP-seq profile with a vector X.

Single signal peak callers (SPCs) typically work in two phases. First, they segment the genomic signal into background regions and regions with potential peaks. The segmentation is either performed with a window-based approach or more sophisticated methods like hidden Markov models (HMMs) (Rabiner, 1989). Then, they perform a statistical test to check

	#read within peaks	#read outside peaks	FRiP
S1	1018	9775	0.09
S2	4656	13025	0.26

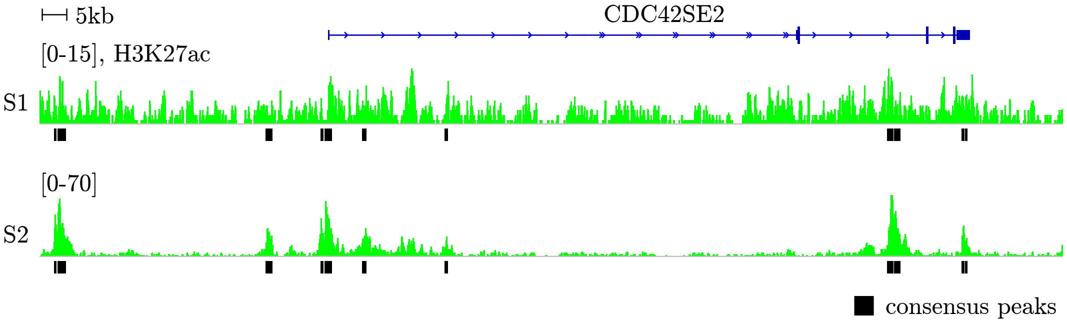


Figure 2.5.: Example for different signal-to-noise ratios. The figure shows two ChIP-seq profiles of histone modification H3K27ac based on two cancer patients of the same cell. We therefore assume that the ChIP-seq profiles have similar peaks. ChIP-seq profile S1 has a low and profile S2 a high signal-to-noise ratio (please note the different y-axis scales). To compute FRiP, peaks (black bars below the signal) are first called for S1 and S2. Next, the number of reads falling into these peaks is divided by the total number of reads. The ChIP-seq data stem from Koues et al. (2015).

whether the potential peaks significantly differ to the background signal. SPC provide a list of peaks where each peak is typically assigned to a p-value.

The single signal peak calling problem has already been addressed by several research groups. Wilbanks and Facciotti (2010) as well as Chen et al. (2012) review and evaluate various SPCs. Single signal peak callers with good evaluation performance in transcription factor binding sites (TFBS) studies are for example PeakSeq (Rozowsky et al., 2009), QuEST (Valouev et al., 2008) and MACS (Zhang et al., 2008). Moreover, sophisticated segmentation methods like HMMs are used for example by HMCan (Ashoor et al., 2013) and BayesPeaks (Spyrou et al., 2009). Figure 2.6 gives an example for the peak prediction of a SPC. Light blue stripes below the ChIP-seq profiles indicate genomic regions, where the SPC calls a peaks.

ChIP-seq experiments are often replicated to avoid considering peaks resulting from variability by random chance. Replication is therefore desired to distinguish between biological and random events as well as to verify the reliability of experimental steps (Park, 2009). The majority of SPCs is not able to handle replicates; and only recently, strategies have been developed for this purpose. For example, the ENCODE project proposes the use of the irreproducible discovery rate (IDR). IDR finds common peaks of a set of candidate peaks that are separately called by SPCs on individual replicates (Landt et al., 2012; Li et al., 2011). Also, Ibrahim et al. (2015) propose a method for the joint analysis of ChIP-seq replicates for the single signal peak calling problem. Their method detects peak boundaries with higher precision than identifying common peaks in replicates with IDR or pooling ChIP-seq reads of replicates.

2.3.2. Differential Peak Calling Problem

Differential peak calling is an important problem in current medical and biological research that investigates changes in protein-DNA interactions of distinct cellular conditions. In contrary to the single signal peak calling problem, this computationally challenge has not been extensively addressed. The differential peak calling problem is defined as follows:

Definition 2.2 (Differential Peak Calling Problem)**.** *Given two experimental conditions* $X_1 = \{X_{11},\ldots,X_{1k}\}$ *and* $X_2 = \{X_{21},\ldots,X_{2k}\}$ *containing a set of genomic ChIP-seq signals, find genomic positions (differential peaks) where* X_1 *and* X_2 *significantly differ.*

We are interested in significant differential peaks (DPs) between two biological conditions X_1 and X_2 which can or cannot contain replicates.

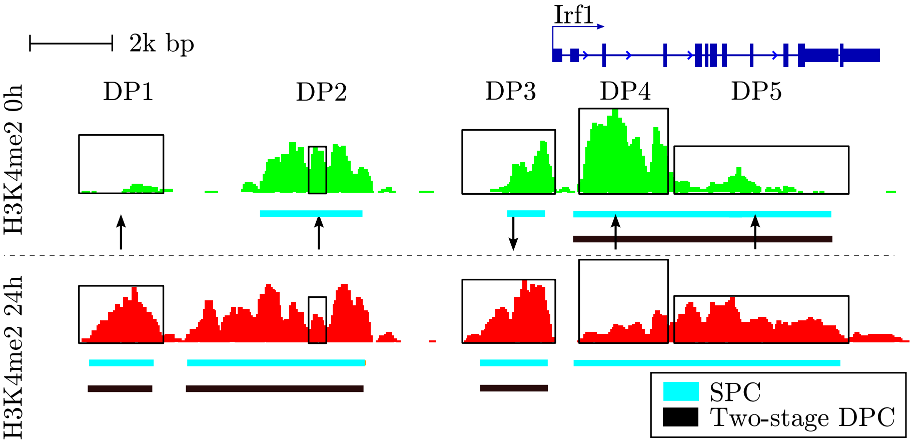


Figure 2.6.: Differential peak calling example. We show an example of two distinct ChIP-seq signals for the histone modification H3K4me2 before (0h, upper signal) and 24 hours after (24h, lower signal) induction of TLR4 signaling of macrophages around the gene Irf1 (Kaikkonen et al., 2013). We indicate with squares examples of regions, which are putative DPs with gain (or loss) of ChIP-seq signal after 24 hours of TLR4 treatment. The height of the squares indicates the size of the highest ChIP-seq signal for a DP. We display results from the SPC PeakSeq (grey bars) and a two-stage peak caller based on applying DESeq on PeakSeq peaks (black bars). PeakSeq successfully detects broad peaks describing ChIP-seq signal for each cell. The two-stage peak caller can detect DP1 and DP3, but cannot detect changes within the broad candidate peaks such as DP2 or complex changes in the signal within the Irf1 gene body (DP4 and DP5).

Initially, differential peak calling was performed by peak calling on individual ChIP-seq signals. Peaks detected in only one of the conditions were then defined as cell-specific peaks (Heinz et al., 2010). However, such methods are not able to detect cases where peaks were presented (and called) in both cell types, but exhibit a significant increase (decrease) of the DNA-protein signal in one of the cells. In the example of Figure 2.6, which is based on peaks from PeakSeq (Rozowsky et al., 2009), only DP1 would be detected as cell-specific. Moreover, most SPCs do not provide any functionality to normalize ChIP-seq profiles. Thus, it is likely that they show bias in experiments with distinct number of reads.

A more sophisticated strategy to detect DPs is the combination of peaks from SPCs with statistical methods for the analysis of differential gene expression of RNA-seq data. These two-stage differential peak callers (DPCs) first combine peaks that are called on individual ChIP-seq conditions using SPCs. Next, they count the number of reads for each candidate peak, perform signal normalization and apply statistical tests assuming a differential count model. Therefore, they can detect candidate peaks where the number of read counts is significantly higher or lower in one of the ChIP-seq conditions. While this approach allows the detection of significant changes in ChIP-seq data within candidate peaks, it is highly dependent on the initial peak calling step as well as on the strategy used to create the set of candidate peaks. For example, histone modifications associated with active regulatory regions occur in domains spanning several hundreds of base pairs and may have intricate patterns of gain/loss of ChIP-seq signals within the same domain. SPCs tend to call the domains as single peaks and consequently the differential analysis is only able to evaluate the differential counts of the complete called peaks. In Figure 2.6, the SPCs calls one peak for the gene body of Irf4. The DPC is therefore not able to distinguish between DP4 and DP5.

A further strategy to detect DPs is to first segment the genome with a fixed window. Next, it is tested whether the windows contain differential counts. Heuristic methods are applied to merge windows in close vicinity to each other and with similar counts (Shen et al., 2013). The performance of such methods depends on the window merging strategy as well as the window size. Too large windows tend to omit small peaks in the ChIP-seq signal. For example, DP2 in Figure 2.6 is not detectable with a window size of 1000bp, which is the default parameter of the DPC Diffreps (Shen et al., 2013).

2.3.3. Example of Studies Comparing ChIP-seq Signals

There are several studies of current biological and medical research that compare ChIP-seq signals under distinct conditions. These studies investigate for example

- cell differentiation: Lin et al. (2015) investigate regulatory changes in a mouse model during the development of antigen-presenting dendritic cells with regard to the histone modifications H3K4me1 and H3K27ac.
- cell activation: Saeed et al. (2014) perform ChIP-seq experiments in humans for monocytes that are activated to macrophages. The differentiation from monocytes to macrophages plays a key role in the host's defence system. The study describes epigenetic differences with regard to several histone modifications. Biological replicates based on different donors are used for the study.
- comparison of healthy and diseased individuals: Koues et al. (2015) analyze the difference of regulatory genomic features between healthy individuals and lymphomas patients. They investigate the histone modification H3K27ac.
- the activation of signaling pathways: Kaikkonen et al. (2013) describe the response of macrophages after the time dependent activation of the TLR4 pathway which plays an important role in the immune system. Their study comprises a mouse model and does not provide replicates.

Calling DPs in these studies can give findings that generally lead to a deeper understanding of epigenetics. Depending on the application, DP predictions may exhibit starting points for drug discovery and epigenetic biomarker detection (Koues et al., 2015; Saeed et al., 2014), give new insights into cell differentiation steps (Lin et al., 2015) or unravel mechanisms for the immune system activation (Kaikkonen et al., 2013).

2.3.4. Arising Challenges

The differential peak calling problem leads to computational challenges which arise additionally to the ChIP-seq specific tasks described in Section 2.2.5.

Replicates

Replicated ChIP-seq experiments can be used to reduce the effect of unwanted technical bias. There are two kinds of replicates of ChIP-seq experiments. Biological replicates stem from independent cell cultures or tissue samples to ensure reproducibility. Technical replicates are based on measuring a single biological sample and can therefore only be used to estimate the variability of the sequencing step (Yang et al., 2014). These two kinds exhibit different characteristics. For instance, the variance between biological replicates is supposed to be higher than for technical replicates, as they stem from various biological samples.

If replicates are available, the problem becomes computationally more complex as more information has to be taken into account. In particular, count data derived from NGS data usually exhibit overdispersion, that is, the variance in the data exceeds the mean (Anders and Huber, 2010; Cameron and Trivedi, 2001; Ismail and Jemain, 2007). To ensure accurate DP estimates on must take into account overdispersion. This typically requires the use of complex statistical models.

Normalization

For the differential peak calling problem, we compare several ChIP-seq profiles which typically exhibit different sequencing depths as well as different signal-to-noise ratios. ChIP-seq profiles may be over- or underrepresented when comparing them, and therefore normalization against different sequencing depths is necessary. The signal-to-noise ratios should also be considered in the normalization. Even in the case without replicates, normalization of samples associated with distinct conditions is important.

Evaluation

There is no direct metric to systematically quantify DP predictions. Furthermore, due to the biological complexity, there is no genome-wide map of DNA-protein interactions which could be used as a gold standard. Consequently, evaluating solutions for the differential peak calling problem is still an open problem. However, indirect metrics can be used to quantify the DP predictions. For example, as gene expression correlates well to certain histone modifications (Karlić et al., 2010), the validation of DP predictions with gene expression is possible. Furthermore, the simulation of ChIP-seq reads is an effective strategy to produce artificial gold standards with various data characteristics (Humburg, 2011; Zhang et al., 2008; Lun and Smyth, 2014).

2.4 Related Work

The differential peak calling problem for ChIP-seq data has been addressed only in a few studies. Here, we first review normalization approaches. Second, we give an overview of existing simulation algorithms for ChIP-seq data. Finally, we list existing methods to solve the differential peak calling problem and give a short description of their working procedure. We explain how the tools address the challenges described in Section 2.3.4 and Section 2.2.5.

2.4.1. Normalization

The majority of normalization approaches multiplies the ChIP-seq signals by a constant factor. One example is the normalization by library sizes. Here, the normalization factor denotes the ratio between the total number of counts of (1) the signal with the highest total number of counts; and (2) the total number of counts of the signal which has to be normalized. Thereby, signals with lower total counts are raised to the level of the signal with maximal total counts.

We demonstrate the normalization by library sizes with an example. We resort to Figure 2.5 which shows two replicates of the same condition, but with different read counts and signal-to-noise ratios. As the replicates stem from the same condition, the consensus peaks should have similar counts across the conditions after normalization. MA plots visualize count distributions in two genomic signals and we use them to picture the counts of the consensus peaks in Figure 2.5. We divide the genome into consecutive bins and count the signal specific reads falling into these bins, that is, for each bin, we obtain the number of reads for both signals. The MA-plot assigned for each bin the M-value, that is, the logarithmic ratio of the counts, to the A-value, that is, the logarithmic mean of the counts. Figure 2.7A gives the MA-plot for the signals shown in Figure 2.5. The rationale for using MA-plots is that, after signal normalization, the bins associated with consensus peaks (indicated by red points) should give low absolute M-values, as they stem from two replicates of the same condition. Without any normalization, the mean M-value of peak-associated bins (MMP) is 2.4 in the example of Figure 2.5.

In this example, the normalization by library sizes gives a factor of 1.6 for signal S1. S1's low signal-to-noise ratio inhibits a higher normalization factor, as the entire signal of S1, including the noise, is used for the calculation. Figure 2.7B gives the corresponding MA-plot. Compared to the case without normalization, the peak-associated bins yield a lower MMP value (1.54), which demonstrates the advantage of the normalization.

Robinson and Oshlack (2010) propose a strategy to normalize RNA gene expression data under the assumption that the majority of the genes are not differentially expressed. For given signals, they first compute the M- and A-values. Next, they estimate a quantile based range of the values which are used for the normalization. The rationale of not considering outliers of M- and A-values is that they may have a strong influence to the results (Meyer and Liu, 2014). The normalization factor is computed by the product of the resulting M- and A-values normalized against the A-values. They refer to their normalization approach as trimmed mean of M-values (TMM). Anders and Huber (2010) implemented a similar approach by using the geometric mean of the gene expression data. The majority of DPCs dealing with replicates use a TMM-based normalization strategy.

Figure 2.7C depicts the MA-plot after normalizing with a TMM-based factor of 1.29 for S1 and 0.98 for S2. Trimming M- and A-values does not exclude the noise signal which is still comprehensively considered for the estimation of the normalization factors. Hence, the TMM normalization leads to an MMP of 1.79 which is in this example even higher than the MMP of the simple normalization by library sizes (1.54).

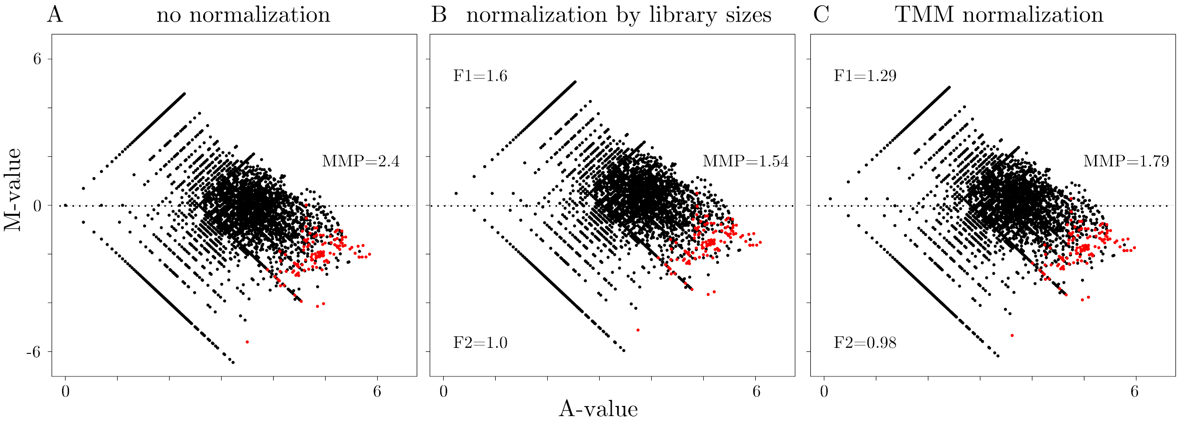


Figure 2.7.: We show MA-plots for different normalization approaches for the region shown in Figure 2.5, where we apply IDR to find common peaks within the replicates. Bins associated with IDR-peaks are highlighted in red. We also give the mean M-value of peak-associated bins (MMP) which is supposed to be low after normalization. We depict MA-plots without (A), after the normalization by library sizes (B) and after the TMM (C) normalization.

2.4.2. Evaluation

There is neither a direct metric to rate nor a gold standard to compare DP predictions. However, the simulation of ChIP-seq profiles is an effective strategy to evaluate DPCs. The majority of ChIP-seq simulation algorithms is based on the single signal peak calling problem. Zhang et al. (2008) developed a strategy to model TF-based ChIP-seq signals that contain sharp peaks. They simulate the background noise with a Gamma distribution which determines the impact of noise on particular genomic regions. Zhang et al. (2008) do not provide NGS reads and consequently they lack to model the bias based on the sequencing process, such as the GC-content. Humburg (2011) followed Zhang et al. (2008) and extended their model to make it capable of producing NGS reads. He enhanced the model by making it more flexible in terms of the number of reads and the number of binding events. However, the model of Humburg (2011) has no parameter to directly set the number of binding events that occur in the simulated ChIP-seq data. None of these approaches can be directly used for the differential peak calling problem as they are restricted to exactly one ChIP-seq profile.

Lun and Smyth (2014) developed a method to simulate ChIP-seq profiles with DPs between two conditions. In this method, the reads of an enriched regions are sampled from a Negative Binomial distribution. DPs are included by adjusting the parameters of the Negative Binomial distribution such that one condition is expected to gain more reads than the other. Next, the reads' positions are determined. The simulation algorithm lacks to model crucial parameters such as the background noise and the variability of peaks in ChIP-seq profiles associated with the same condition. Also, their simulation algorithm is not publicly available.

2.4.3. Two-Stage Differential Peak Caller

DPCs can be roughly categorized in two-stage and one-stage DPCs. Two-stage approaches are based on separate candidate peaks for each ChIP-seq profile. These candidate peaks are pre-computed by SPCs and used as input for sophisticated differential count models. In general, two-stage DPCs merge the candidate peaks with regard to the conditions they stem from, count the number of reads for each candidate peak, perform signal normalization and apply statistical tests assuming particular count models. Some two-stage DPCs use count models that are tailored for the differential expression analysis of RNA-seq data. Such models are for instance implemented in DESeq (Anders and Huber, 2010) and edgeR (Robinson et al., 2010). We list all two-stage DPCs that are, to our best knowledge, available. Table 2.1 gives an overview of the tools and their supported features.

DiffBind

DiffBind (Ross-Innes et al., 2012) is a two-stage differential peak method based on SPC candidate peaks. First, the peak lists are merged to obtain consensus peaks. The number of reads falling into these consensus peaks are counted and a statistical model based on edgeR (Robinson et al., 2010) is estimated to call DPs. The count data is modeled by a Negative Binomial distribution to take into account overdispersion induced by potential replicates. DiffBind normalizes data by following the TMM approach after input-control is subtracted from ChIP-seq profiles. DiffBind can also account for replicates. However, neither the fragmentation size nor GC-content is estimated by DiffBind. Also, the input-DNA is not normalized and no postprocessing step is implemented to filter artefacts.

MACS2

MACS2 (unpublished, available at `https://github.com/taoliu/MACS/`, last access October 14th, 2015) works in two steps. First, all ChIP-seq profiles are pooled together and MACS2's SPC (*callpeak*) is executed for each condition. Second, we apply MACS2's algorithm *bdgdiff* to identify DPs within these peaks. The SPC normalizes against input-DNA, considers GC-content and estimates the fragmentation size. MACS2's differential peak calling method works with a sliding window approach on candidate regions (personal communication). There is no formal description of its parameters and the strategy for normalization.

DESeq

One can combine DEseq with SPCs to make it applicable to the differential peak calling problem. First, a SPC computes a list of candidate peaks for each condition and second, DESeq is used to determine DPs. DESeq (Anders and Huber, 2010) is a tool to analyze differential gene expression. The observed counts are normalized with the geometric mean and the count data is modeled with a Negative Binomial distribution. DESeq uses the Negative Binomial distribution to compute a p-value for each estimated differential gene. By using the Negative Binomial distribution, DESeq is capable to take overdispersion into account. In general, combinations of DESeq and a SPC do not apply any filtering steps to avoid strand bias.

If replicates are available, we have to consider proper SPCs. We use JAMM (Ibrahim et al., 2015), a peak caller that considers replicates, to define a peak list and refer to this method as DESeq-JAMM. JAMM takes input-DNA into account and subtracts it from ChIP-seq profiles. Also, we apply IDR (Li et al., 2011) which is a method to define for a set of replicates a list of peaks with high consistency within the replicates. We follow the framework of ENCODE for the IDR computation (see `https://sites.google.com/site/anshulkundaje/projects/idr`, last access on 21th November 2014). We refer to this method as DESeq-IDR.

DBChIP

The two-stage DPC DBChIP (Liang and Keleş, 2012) receives as input the summit (position with maximal count within a peak) information of peaks from SPCs. The peaks' summits are clustered to obtain consensus peaks. Then, edgeR (Robinson et al., 2010) is applied to derive DPs from the consensus peaks. If available, input-DNA is subtracted from the ChIP-seq profiles. DBChIP focuses on the analysis of transcription factor peaks and therefore uses predefined short regions of 200 bp around the peak summits as candidates for DPs. DBChIP is not able to take into account replicates, to compute the GC-content and to normalize the input-DNA before subtraction. The fragmentation size can be computed by the SPC that is used.

MAnorm

MAnorm (Shao et al., 2012) receives as input the candidate peaks from SPCs. MAnorm normalizes the peak counts between two samples with a local robust regression approach and computes for each candidate peak a p-value. The p-value is used to check whether a DP has been found. The fragmentation size can be computed by the used SPC. MAnorm is not able to consider replicates and does not take advantage of input-DNA or GC-content. Furthermore, no post-precessing steps are performed.

2.4.4. One-Stage Differential Peak Caller

One-stage DPC methods are based on segmentation methods, such as hidden Markov models (HMMs) or sliding window-based approaches. While two-stage DPC work in two phases, one-stage DPCs analyze ChIP-seq profiles and perform DP calling in a single step. We list all one-stage DPC that are, to our best knowledge, available. Table 2.1 gives an overview of the tools and their supported features.

ChIPDiff

To our knowledge, the earliest published method proposed for the differential peak calling problem is ChIPDiff (Xu et al., 2008). ChIPDiff uses a three state HMM to distinguish between DPs and background signal. The HMM emission is based on an approximation of a Beta-Binomial distribution, which is fixed after the initialization of the model. The Baum-Welch algorithm is used to estimate transition parameters. ChIPDiff exhibits some limitations. Instead of a *p*-value , an empirical fold-change criterion is used to determine whether a DP is significant. Moreover, the fragmentation size of a ChIP-seq experiment is fixed to 200bp. ChIPDiff does not take advantage of input-DNA and does not perform any GC-content normalization. Also, replicates are not supported.

Csaw

Csaw (Lun and Smyth, 2014) main method is a window-based approach to segment ChIP-seq profiles. Replicates can be taken into account. A modified version of the TMM method is applied to normalize the CHIP-seq signal on 10kbp bins. EdgeR (Robinson et al., 2010), which is based on a Negative Binomial distribution test, is used to assign a *p*-value to each DP. Consecutive significant bins are merged to form final DPs. Input-DNA is not used to normalize ChIP-seq signals, but only in a postprocessing step to filter out potential false positive DPs. Furthermore, csaw does not normalize against GC-content and does not estimate the fragmentation size.

PePr

PePr (Zhang et al., 2014) follows a window-based strategy to detect DPs. The windows size is computed automatically and equals the estimated average width of initially called peaks. PePr normalizes the input-DNA to the mean of all ChIP-seq signals, computes the fold change of input-DNA and ChIP-seq signal and follows the TMM approach to globally normalize across different ChIP-seq profiles. PePr requires input-DNA to run. To check for DPs, first read counts are modeled by a Negative Binomial distribution and second Wald's test is applied to check for significance in read counts. Furthermore, PePr provides estimation of fragment size, input subtraction, filtering of peaks with strand bias, but does not correct for GC-content. PePr can handle replicates of ChIP-seq profiles.

DiffReps

DiffReps (Shen et al., 2013) performs a sliding window approach to identify potential DPs. It globally normalizes by the geometric mean for each sample and also takes into account input-DNA. A pre-screening test ensures that only bins with a sufficient number of reads are considered for the analysis. DiffReps can deal with replicates and uses a Negative Binomial test based on Anders and Huber (2010) to detect DPs.

RSEG

RSEG (Song and Smith, 2011) is specialized for the single signal peak calling problem based on repressive histones, which are distributed in large sequenced genomic domains. However, it has an option to call DPs with a three-state HMM. RSEG's HMM uses Difference Negative Binomial distribution as emission distribution. As it is tailored for broad histone marks, we do not take RSEG into account for this thesis.

	characteristics					pre- and postprocessing						
	replicates	One-Stage DPC	Segmentation Strategy	statistical model DP	peak size	frag. size estimation	input-DNA norm.	Subtracting input-DNA	normalization strategy	GC-content	input-DNA not required	strand bias
PePr	×	×	win	W	s/m	×	×	×	TMM			×
diffReps	×	×	win	NB	s/m		×		GM		×	
csaw	×	×	win	NB	s/m				TMM		×	
DiffBind	×		SPC	NB	s/m			×	TMM		×	
DESeq-IDR	×		SPC	NB	s/m				MR		×	
DESeq-JAMM	×		SPC	G,NB	s/m	×		×	MR		×	
MACS2			SPC	*NA*	s/m	×	*NA*	*NA*	*NA*	*NA*	×	*NA*
DBChIP			SPC	NB	s/m			×	MR		×	
MAnorm			SPC	-	s/m				MA		×	
ChIPDiff		×	HMM	-	s/m	fixed			LS		×	
RSEG		×	HMM	NBD	l						×	

Table 2.1.: *Tool characteristics. Differential peak callers can be categorized in one-stage or two-stage approaches using either an HMM or a window-based approach to segment the ChIP-seq profiles. They perform a statistical test based on a Negative Binomial (NB) distribution, a Difference Negative Binomial distribution (NBD), Wald's test (w) or Gaussian mixture model (G) to identify DPs. The tools are specialized in different domain sizes in the ChIP-seq signal. ChIP-seq experiments with small (s) domains are based on TFs, with medium (m) domains on active histone marks, and large (l) domains on repressive histone marks. In this thesis, we investigate differential peak callers that concentrate on small and medium size domains. Input-DNA can be normalized and may be used to subtract it from ChIP-seq profiles (see Section 2.2.4). The normalization strategies are based on TMM, geometric means (GM), median ratios (MR) from Anders and Huber (2010), MA plots (MA) or library sizes (LS) (see Section 2.4.1). Also, normalizing against GC-content may prohibit bias in profiles (see Section 2.2.5). For DESeq-JAMM, JAMM uses GMM to detect peaks and DESeq uses NB to detect DPs. JAMM subtracts the input-DNA from ChIP-seq profiles. MAnorm does not model counts of DPs, but normalizes them and assigns directly a p-value to them.*

2.5 Discussion and Conclusion

Two challenges naturally arise from ChIP-seq data. SPCs call peaks on a single ChIP-seq signal and DPCs identify differences in ChIP-seq signals that are associated with two biological conditions. We divide the DPCs into two classes: two- and one-stage DPCs. Two-stage DPCs have clear conceptional disadvantages. First, their DPs are restricted to their initial candidate regions as well as to the strategy used to create the set of candidate peaks. These candidate regions depend on the SPC and its concrete parametrization. Some SPC are specialized in calling broader regions, while some SPC show advantages in calling sharp peaks (Wilbanks and Facciotti, 2010). Consequently, prior knowledge of the data is required to obtain accurate peak predictions. While two-stage DPCS can detect DPs where the number of read counts is significantly higher or lower in one of the ChIP-seq conditions, two-stage DPC fail to detect subtle changes within these candidate regions (Allhoff et al., 2014; Maze et al., 2014). This is particularly crucial for ChIP-seq data of histone modifications, where DNA-protein interactions occur in mid-size to large domains. Histone modifications associated with active regulatory regions occur in domains spanning several hundreds of base pairs and may have intricate patterns of gain or loss of ChIP-seq signals within the same domain. In contrary, ChIP-seq from transcription factors mostly happens in small isolated peaks. Figure 2.6 shows an example for the predictions of a SCP which are merged by a two-stage DPC to identify DPs. The two-stage DPC fails to detect DP2, as the SCP predicts a too broad peak in this region such that the signal change of DP2 is not detectable for the DPC. Furthermore, the SCP calls a domain that contain both DP4 and DP5. Consequently, it is impossible for the DPC to distinguish between these DPs. Second, two-stage DPC methods usually do not provide any preprocessing steps crucial for ChIP-seq analysis, such as fragment size estimation, GC-bias correction and input-DNA subtraction (see Table 2.1).

The majority of DPCs, namely DiffBind, csaw and PePr, uses TMM or an approach similar to TMM (median ratio (MR), see DESeq and DBChIP in Table 2.1) to normalize ChIP-seq profiles. However, TMM was devised for gene expression experiments which assumes that counts of most observations (genes or peaks) do not change. This is not necessarily the case for protein interactions, as two distinct cells can have distinct amounts of proteins or histone modifications bound to their DNA (Meyer and Liu, 2014). Particularly problematic in this normalization approaches is the effect of replicate specific background noise. Background noise does not reflect the protein-DNA interaction sites and therefore induces bias in the normalization strategy.

Moreover, all one-stage DPCs that solve the differential peak calling problem with replicates use window-based approaches to identify DPs and apply heuristic strategies to merge peaks (DiffReps, PePr and csaw, see Table 2.1). HMM-based approaches are more appropriate to segment a signal, as they intrinsically detect peaks with variable size through the use of posterior decoding algorithms. Hence, HMMs represent a robust alternative to windows-based segmentation approaches.

There is no method that addresses all pre- and postprocessing steps listed in Table 2.1. For instance, only PePr, MACS2 and DESeq-JAMM are able to estimate the fragmentation size of a ChIP-seq experiment. Furthermore, input-DNA may help to identify technical artifacts and therefore to avoid false positive DPs. PePr, DiffBind, DESeq-JAMM and DBChIP take advantage of input-DNA, but only PePr also normalizes the control DNA. No method takes GC-content into account to improve the DP predictions. Also, only PePr provides postprocessing steps to get rid of implausible DP candidates. The main disadvantage of PePr is that is requires input-DNA to predict DP which is not always available. Furthermore, PePr uses a window-based approach to detect DPs. There is no method that takes into account blacklisted genomic regions.

The systematic evaluation of DPCs is still an open problem. An indirect metric, such as the combination of gene expression data with DP estimates, can determine the quality of DPs. Moreover, simulated data can help to investigate solutions for the differential peak calling problem in a systematic way as gold standards can be customized. In particular, it is crucial to have methodologies exploring the performance of DPCs on data with distinct characteristics: from ChIP-seq samples with low variability and high signal-to-noise ratio to samples with high variability and low signal-to-noise ratios.

2.6 Aims of the Thesis

In the previous sections, we pointed out that various methods have been developed to solve the differential peak calling problem. We discussed that two-stage DPCs have conceptual disadvantages and that therefore one-stage DPCs are the favourable method of choice. However, there is no method that covers all challenges that have to be considered in the ChIP-seq analysis (see Table 2.1). Hence, the aims of this thesis are the following:

- we want to propose one-stage DPCs using HMMs that take into account all challenges associated with ChIP-seq. We restrict our analysis to TFs and activating histone marks resulting in small to medium sized peaks in the ChIP-seq signal. ChIP-seq experiments are typically replicated to reduce the effect of technical bias. Hence, our methods have to account for replicates and properly consider overdispersion in their statistical models (see Section 2.3.4).

- we pointed out that normalization of ChIP-seq profiles is a crucial step to identify DPs (see Section 2.3.4). In this thesis, we want to propose a novel normalization strategy that is more robust to background noise. The background noise is problematic for TMM and the normalization by library size.

- the evaluation of differential peak calling solutions is still an open problem. As described in Section 2.3.4, there is neither a gold standard nor a direct metric to check the quality of a differential peak calling solution. Evaluation strategies can assess DP estimates in a systematic way. In this thesis, we want to propose an indirect metric to quantify DPCs. Moreover, we will develop a simulation algorithm to be able to produce customized gold standards. With regard to these evaluation strategies our methods should give best results.

- we listed several challenges that either arise from the ChIP-seq protocol itself (see Section 2.2.5) or in particular from the differential peak calling problem (see Section 2.3.4). Our proposed methods will address all of them: the GC-content, to compensate the correlation between the number of reads and the underlying GC-content; PCR duplicates, to avoid signal in ChIP-seq profiles which is based on PCR duplicates rather than biological events; the input-DNA, to get rid of bias in the ChIP-seq data for example due to the shearing process; the fragment size estimation, to compute the precise location of the DNA-protein complexes in the genome; and blacklisted genomic regions, to get rid of DPs that lie within regions that are not properly covered by the sequencing process. None of the competing methods listed in Table 2.1 addresses all of these issues.

CHAPTER 3

Methods

In the previous chapter, we introduced the fundamental biological concepts as well as the ChIP-seq technique. We also formalized and motivated the differential peak calling problem. The aim of the thesis is to develop algorithms to call differential peaks in ChIP-seq profiles. In this chapter we explain our differential peak calling methods to achieve this goal. Algorithm 3.1 gives an overview of the methods. We first introduce the notation and conventions that are necessary to formalize our solution. We then explain the preprocessing steps that are required to make the ChIP-seq signal applicable for our methods (see Algorithm 3.1, Step 1). In particular, we propose a novel normalization strategy for ChIP-seq signals which is based on control regions. After a brief introduction to HMMs, we describe how to use them to estimate potential DP candidates (see Algorithm 3.1, Step 2). HMMs represent a convenient strategy to segment the signal and are not considered by the majority of the competing methods (see Table 2.1). Next, we explain which postprocessing steps are performed to obtain the final DPs (see Algorithm 3.1, Step 3). Here, we propose a novel p-value estimation strategy which is based on an HMM. Finally, we describe the implementation of our solutions.

3.1 Notations and Conventions

We denote an alphabet Σ and typically use $\Sigma = \{A, C, G, T\}$, where the nucleotides are given as capital letters. Character N is used as a wildcard that represents any element in Σ. A string is an element of Σ^* and is denoted a lower case character. Let $s = \langle s_1, \ldots, s_m \rangle$ be a string. Then, string s has length $|s| = m$ and the substring $\langle s_i, s_{i+1}, \ldots, s_j \rangle$ is written as $[s_i, s_j]$. A genome is a string which can be divided into a sequence $\langle b_1, \ldots, b_L \rangle$ of bins. A bin is assigned to the number of reads covering this bin. The number of reads for a bin is the genomic signal for that bin. We use genomic signal, ChIP-seq experiment and ChIP-seq profile as synonyms. The index i of a genome is called a genomic position.

The matrix $\mathbf{X}$ represents a genomic signal

$$\mathbf{X} = \{x_{ij}\}^{D \times L},$$

where D is the number of genomic signals and L the number of bins. The ith genomic signal is represented by the vector $x_{i\cdot} = (x_{i1}, \ldots, x_{iL})$ and the genomic signals for bin j is represented by the vector $x_{\cdot j} = (x_{1j}, \ldots, x_{Dj})$. Moreover, each ChIP-seq experiment belongs to one of K biological conditions. The set of experiments associated with condition k is given as

$$G_k = \{i \mid i \in \{1, \ldots, D\}, i \text{ belongs to } k\},$$

and the set of all experiment as

$$G = \{G_1, ..., G_K\}.$$

In this thesis we investigate the case $K = |G| = 2$, that is, we are interested in two biologi-

Algorithm 3.1 Differential peak calling algorithm

Input: reference genome g, two sets of aligned reads S_1, S_2
Output: list of DPs $\langle d_i \rangle,\ i \in \mathbb{N}$

```
1. employ preprocessing pipeline to $S_1$ and $S_2$:                         ▷ Section 3.2
   1.1 filter reads in $S_1, S_2$                                              ▷ Section 3.2.1
   1.2 estimate fragmentation size $\hat{f}$                                   ▷ Section 3.2.2
   1.3 create signal matrix $\mathbf{X}^{D \times L}$                          ▷ Section 3.2.3
   1.4 normalize $\mathbf{X}$ against GC-content                               ▷ Section 3.2.4
   1.5 if $x^{input}$ available:
           normalize with input-DNA $x_{i\cdot} = x'_{i\cdot} - \alpha \cdot x_i^{input}$   ▷ Section 3.2.5
   1.6 normalize* $\mathbf{X}$ among ChIP-seq profiles                         ▷ Section 3.2.6
2. identify candidate DPs $\langle d'_i \rangle$ with HMM* $\delta$:
   if $D > 2$:                                                                 ▷ with replicates
       use Negative Binomial as emission of HMM $\delta$                       ▷ Section 3.3.3
   else:                                                                       ▷ without replicates
       use Binomial or mixture of Poissons as emission of HMM $\delta$         ▷ Section 3.3.3
3. postprocess* candidate DPs $\langle d'_i \rangle$ and output final DPs $\langle d_i \rangle$   ▷ Section 3.4.2
```

* novel features proposed in this thesis

cal conditions. In particular, in case $D = 2$ without replicates, we have $|G| = 2$ and the two genomic signals are associated with different conditions. Further, $x_{G_k j}$ represents the genomic signal restricted to experiments belonging to G_k and $\bar{x}_{G_k j}$ is the mean read count for all experiments in condition k, that is,

$$\bar{x}_{G_k j} = \frac{\sum_{i \in G_k} x_{ij}}{|G_k|}.$$

Moreover, ChIP-seq experiments often have input-DNA for each cell type analyzed (see Section 2.2.4). We will refer to input-DNA as $x^{input} = \{x_1^{input}, ..., x_L^{input}\}$.

The probabilities measure is given by Pr. We use bp (base pair) as length unit for nucleotide sequences.

3.2 Preprocessing Pipeline

We employ a pipeline to preprocess data obtained by ChIP-seq. The aim of the pipeline is to construct and to improve the genomic signal represented by matrix $\mathbf{X}$. Signal improvement is necessary since the data contains bias as it is described in Section 2.2.5. The first step of Algorithm 3.1 gives an overview of the pipeline. The rationale for steps $1.1 - 1.5$ are well described by the research community. Step 1.6, the normalization of ChIP-seq profiles, is crucial for a successful peak calling step. We propose a novel normalization strategy which is based on control regions. We assume that the data are given as reads that are aligned to a genome g.

3.2.1. Filtering of Reads

Reads are mapped to the reference genome g and serve as input for our method. We ignore reads mapping to genomic regions which are either unassembled (denoted by Ns in the genome) or that exhibit a poor mappability (see Section 2.2.5). Regions with poor mappability stem from the fact that short reads cannot be uniquely mapped to repetitive regions that exhibit a higher length than the reads themselves (Song and Smith, 2011). Reads aligned completely to a region with poor mappability are ignored. Moreover, we ignore all reads but one that are mapped to a same coordinate as it is likely that these reads stem from PCR duplicates (see Section 2.2.5).

3.2.2. Fragment Size Estimation

Figure 2.4 (see Section 2.2.5) depicts the situation where reads are aligned to the forward and reverse genome and lie half the fragment size away from the ChIP-seq source. As only the beginning of the sample's DNA fragments is sequenced in the ChIP-seq protocol, we have to reconstruct the unknown fragment size f. With the reconstructed fragment size, peaks in the read distribution correlate to protein position in the ChIP-seq signal.

Given set F of the leftmost positions of all reads aligned to the forward strand and given set R of the rightmost positions of all reads aligned to the reverse strand, we follow Mammana et al. (2013) and define the strand cross-correlation function

$$c(f) = \sum_{p \in F \cup R} h(p) \cdot h(f+p), \quad \text{with}$$

$$h(x) = \begin{cases} 0, & x \notin F \cup R, \\ 2, & x \in F \cap R, \\ 1, & \text{else.} \end{cases}$$

The convolution c gives the correlation between counts on the forward and reverse strands for a given fragment size f. The fragment size f corresponds to the value with the maximum correlation between both strands, that is,

$$\hat{f} = \text{arg max}_{\text{f} \in \text{G}}\, c(f),$$

with $G \subseteq \mathbb{N}$. In other words, we shift the coordinates and compute the overlap. The shifting distance resulting in a maximum overlap corresponds to the size of the fragments covering the target DNA-protein complexes (Kharchenko et al., 2008).

3.2.3. Signal Profile

We create the genomic profile of a ChIP-seq experiment by fragmenting the genome into bins and counting the reads falling in these bins. First, we extend all forward (reverse) reads from the leftmost (rightmost) position to the $3'$ ($5'$) direction by the estimated read fragment size $\hat{f}$. Second, we divide the genome into consecutive bins $\langle b_1, \ldots, b_L \rangle$ by using a sliding window approach. Each bin b_j covers the genomic positions $[j \cdot s - 0.5 \cdot w, j \cdot s + 0.5 \cdot w]$, where s and w are the step size and the window size. The genomic positions are restricted to a range from 0 to the genome's length. The value of the genomic profile x_{ij} is the number of extended reads of ChIP-seq signal i aligned to regions overlapping bin b_j. If a read lies entirely in a filtered region, it will be ignored (see Section 3.2.1).

3.2.4. GC-Content

Sequencing technologies usually exhibit an undesired correlation between the number of reads and the GC-content of the regions where the reads are located (see Section 2.2.5). To model and correct this effect, we use an histogram-based approach inspired by Ashoor et al. (2013). Let $g_j \in [0,1]$ indicate the GC-content of the genomic bin b_j, that is, the proportion of Gs and Cs in the bin's underlying genomic sequence. We want to measure the average number of reads from input signal x^{input} assigned to bins on a particular GC-content interval. For an interval $[v, v+\delta] \subseteq [0,1]$ with resolution parameter δ and genomic control signal x^{input}, we have

$$h(v) = \frac{\sum_{j=1}^{L} x_j^{input} \mathbf{1}(g_j \in [v, v+\delta])}{\sum_{j=1}^{L} \mathbf{1}(g_j \in [v, v+\delta])},$$

where $\mathbf{1}$ is an indicator function and $v \in \{0, \delta, \ldots, 1-\delta\}$. We sum over all bins j to compute the average input-DNA signal for a particular GC-content interval. Next, we define the expected signal value of function h as

$$T = \delta \cdot \sum_{v} h(v).$$

We then correct the genomic signal x_{ij} for given $g_j \in [v, v+\delta]$ with

$$x_{ij}^{GC} = x_{ij} \cdot \frac{T}{h(v)}.$$

Loosely speaking, we increase (decrease) the genomic signal of a bin, if the average GC-dependent signal is lower (higher) than expected. We use input-DNA as no immunoprecipitate step has taken place (see Section 2.2.1) and therefore no signal induced by antibodies may influence the correlation between the number of reads and the GC-content.

3.2.5. Control Normalization

To avoid bias associated with the DNA shearing process, the input-DNA is usually subtracted from the ChIP-seq genomic signals (rows of matrix $\mathbf{X}$). We follow the sequencing extraction scaling approach of Diaz et al. (2012), which performs a signal normalization previous to the subtraction. The rationale is that while input-DNA and ChIP-seq libraries usually have similar number of reads, the majority of ChIP-seq reads are concentrated in protein-DNA interaction sites. Therefore, a simple subtraction tends to over-penalize the ChIP-seq signal. For a scaling factor α and input-DNA x^{input}, we perform for the ith signal

$$x_{i\cdot} = x'_{i\cdot} - \alpha \cdot x_i^{input},$$

where $x'_{i\cdot}$ indicates the original signal. See Diaz et al. (2012) for details about the computation of factor α.

3.2.6. Sample Normalization

A crucial aspect in the analysis of multiple ChIP-seq samples is the strategy for a genome-wide normalization of samples to bring them to a similar scale. Here, we describe two normalization approaches. We justify the use of these approaches and also point out their disadvantages. The disadvantages is the motivation to introduce a novel normalization strategy.

SDS-based Normalization Approach

An easy, intuitive approach for normalization is the global sequencing depth scaling (SDS) approach. Here, we simply increase the ChIP-seq signals that exhibit a genome-wide overall lower sequencing depth.
More formally, let $S^1 = \sum_{j=1}^{L} x_{G_1 j}$ and $S^2 = \sum_{j=1}^{L} x_{G_2 j}$ be the signal's total sum for $x_{G_1 \cdot}$ and $x_{G_2 \cdot}$. We scale up the genomic signal with less overall signal by the factor $f = \max(S^1/S^2, S^2/S^1)$. For example, if $S^1 < S^2$, we have

$$x_{G_1 \cdot}^{\text{norm}} = f \cdot x_{G_1 \cdot}$$

with $f = S^2/S^1$. We round all values to obtain count data again.

TMM-based Normalization Approach

Outliers in the data, that is, bins with unexpected high read counts, negatively influence the SDS approach. Due to technical issues in the protocol, high read counts usually occur in ChIP-seq data and the SDS approach consequently artificially increases the signal. A more robust way to normalize is to use the trimmed mean of the genome-wide logarithmic counts (TMM) (Robinson and Oshlack, 2010). Currently, most DPCs use TMM for the normalization of replicates.

More formally, for a given signal $x_{i \cdot}$ with $i \in G_k$, we first estimate the mean signal $\overline{x}_{G_k \cdot}$ of condition k. We add 1 to all count data to avoid zero counts. Then, the logarithmic ratio (M-values)

$$M_j = \log\left(\frac{\overline{x}_{G_k j}}{x_{ij}}\right),$$

and the logarithmic average (A-values)

$$A_j = 0.5 \cdot \log(\overline{x}_{G_k j} \cdot x_{ij}),$$

are estimated for all bins j. To reduce the number of outliers, we use trimmed values for M_j and A_j. The normalization factor f_i is the ratio of M- and A-values weighted by A-values

$$\log(f_i) = \frac{\sum_j A_j \cdot M_j}{\sum_j A_j}.$$

Housekeeping Gene Normalization Approach

TMM was initially devised for gene expression experiments which assumes that counts of most observations (genes or peaks) do not change. This is not necessarily the case for protein interactions, as two distinct cells can have distinct amounts of proteins or histone modifications bound to their DNA (Meyer and Liu, 2014). Particularly problematic in this normalization approaches is the effect of replicate specific signal-to-noise ratio (see Figure 2.5).

We propose a normalization approach that is based on the idea that particular control regions serve as reference points to bring the ChIP-seq signals to the same level independently from the biological condition and from the experiments. Among others, these reference values can be obtained by ChIP-PCR on selected genomic regions. For the case of activating histone modification, we use the promoter of housekeeping genes (HK). Karlić et al. (2010) show that histone modifications correlate well to gene expression and can therefore be used to predict the gene expression level. Housekeeping genes contribute to basic cell maintenance and are therefore expected to maintain constant gene expression level and consequently constant histone modifications (Eisenberg and Levanon, 2013). The overall working assumption for our normalization approach is that housekeeping genes do not change their expression or the abundance of histone marks in their promoter.

More formally, we define a set of control genomic regions, $R = \{r_1, \ldots, r_N\}$. The ChIP-seq signal of region r_n for sample i is

$$h_{in} = \sum_j x_{ij} \cdot \mathbf{1}(b_j \text{ overlaps } r_n).$$

First, for a given region n, we normalize the mean of each sample to the particular signal i

$$h'_{in} = \frac{\overline{h}_{\cdot n}}{h_{in}}.$$

The normalization factor for sample i is

$$f_i = \frac{\sum_n h'_{in}}{N},$$

where N is the number of HK genes. Finally, ChIP-seq count estimates for sample i are given by

$$x'_{i\cdot} = f_i \cdot x_{i\cdot}.$$

We use the promoter of housekeeping genes that are described by Eisenberg and Levanon (2013), namely C1orf43, CHMP2A, EMC7, GPI, PSMB2, PSMB4, RAB7A, REEP5, SNRPD3, VCP, VPS29, for the human genome. For the mouse genome, we use the corresponding genes but left out the human specific C1orf43. See Figure 3.2 for a schematic example of this normalization approach.

As example we perform the HK gene normalization for the signals shown in Figure 2.5. Similar to Figure 2.7, we show in Figure 3.1 the MA-plots without normalization (A), after applying the normalization by library size (B) and after the TMM normalization (C). Moreover, we depict in Figure 3.1D the MA-plot after normalizing the signals with the HK gene strategy. Here, we apply a normalization factor of 4.1 for S1 and of 1.1 for S2. In contrast to the normalization by library size and the TMM normalization, the usage of HK genes ensures not to take noise signal into account. We obtain a mean M-value of peak-associated bins (MMP) of 0.81 which is lower than after the library size (1.54) and TMM (1.79) normalization. Low M-values are expected as the peaks are based on two replicates of the same condition. Hence, this example demonstrates the advantage of the HK gene over the global and TMM normalization.

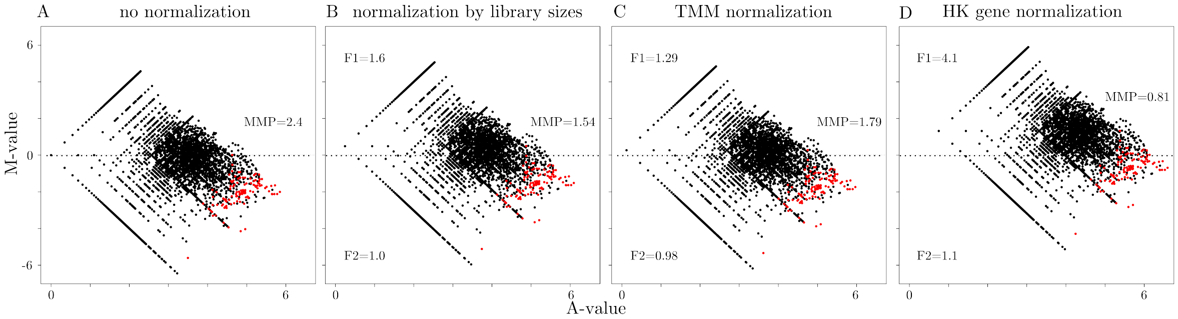


Figure 3.1.: (A)-(C) MA-plots (see Figure 2.7) for the region shown in Figure 2.5. (D) MA-plot for the HK gene normalization. We show the HK gene normalization factors $F1$ and $F2$ as well as the mean M-value of peak-associated bins (MMP).

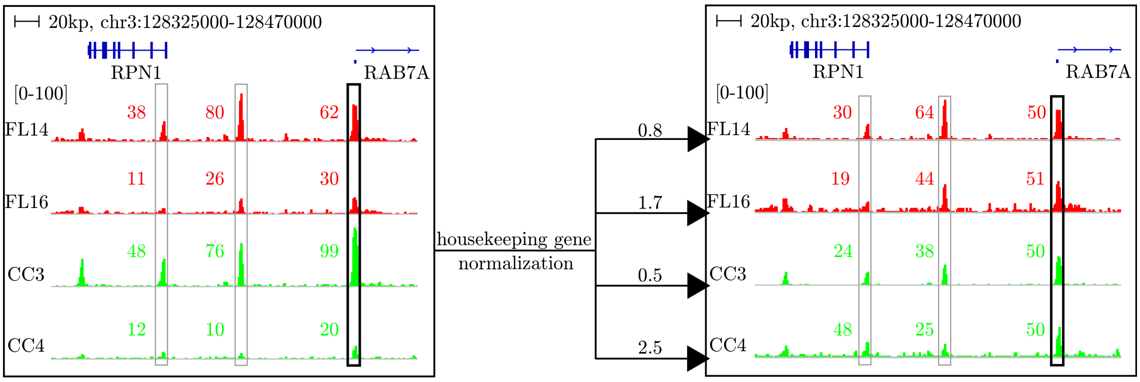


Figure 3.2.: HK gene normalization approach. The left panel shows four ChIP-seq signals assigned to two biological conditions (red and green). More details about the presented data can be found in Section 4.3. Boxes in signals contain peaks where its peak mass is given. The bold box gives the promoter of a HK gene used for normalization. In this carton, the normalization procedure gives 0.8 for ChIP-seq signal FL14, 1.7 for FL16, 0.5 for CC3 and 2.5 for CC4 as normalization factor. The right panel shows the normalized signal with updated mass values of each peak located in a box. The HK gene normalization approach brings all ChIP-seq signals to the same scale for any further downstream analysis steps.

3.3 Differential Peak Calling

We first give a brief introduction to HMMs. Second, we explain how we use HMMs to call candidate DPs. We describe the emission distribution used for the HMM, and explain how to initialize as well as how to train the HMM.

3.3.1. HMM Introduction

A hidden Markov model (HMM) is a stochastic model based on Markov chains. An HMM has a finite set $S = \{1,\dots,M\}$ of states and a probability density function assigned to each state. The HMM is in each time point at a particular state and emits a symbol with a probability given by a certain density function. The emitted symbol is also called observation. More formal, let $X = (X_1,\dots,X_L)$ be a random variable and let $x = (x_1,\dots,x_L)$ be a realisation of X that represents the observation sequence. Moreover, let $Q = (Q_1,\dots,Q_L)$ be an unknown variable. For observation $x = (x_1,\dots,x_j,\dots,x_L)$, we have an unknown sequence of states $q = (q_1,\dots,q_j,\dots,q_L)$ $(q_j \in S)$, where state q_j emits observation x_j. There are two major assumptions for HMMs:

(a) the probability to be in a state depends only of the previous state, that is,

$$\Pr(q_j \mid q_1,\dots,q_{j-1}) = \Pr(q_j \mid q_{j-1}), \quad \text{and} \tag{3.1}$$

(b) the probability of emitting observation x_j depends only on state q_j, that is,

$$\Pr(x_j \mid q_1,\dots,q_j) = \Pr(x_j \mid q_j).$$

The probability given by Equation 3.1 to reach a state is described by a transition matrix

$$A = (a_{kl}) \quad \text{with} \quad a_{kl} = \Pr(q_j = k \mid q_{j-1} = l),$$

for $1 \le k \le M$, $1 \le l \le M$, $a_{kl} > 0$ and $\sum_l a_{kl} = 1$. Furthermore, let $\pi = (\pi_1,\dots,\pi_M)$ be the initial state probabilities $\Pr(q_1 = k) = \pi_k$. In our case, the observation space X is discrete.

We consequently can reduce the probability density functions to probability mass functions $\mathbb{B}$. Hence, an HMM δ is parameterized by $\delta = (A, \mathbb{B}, \pi)$. Let b_s denote the probability mass function associated with state s, that is,

$$b_s(y) = \Pr(y = x_j \mid q_j = s, \Theta_s)$$

with $1 \leq s \leq M$ and $y \in X$. Parameter Θ_s describes the parameter of function b_s.

There are three fundamental computational problems associated with HMMs. The first problem is how to compute the likelihood of a sequence of observations x for a given HMM δ, that is, one has to evaluate

$$\Pr(x \mid \delta) = \sum_{q \in \mathbb{Q}} \Pr(x, q | \delta), \tag{3.2}$$

where $\mathbb{Q}$ gives all possible state sequences. Due to Equation 3.1, it is easy to see that Equation 3.2 is equivalent to

$$\Pr(x \mid \delta) = \sum_{q \in \mathbb{Q}} \pi_{q_1} b_{q_1}(x_1) a_{q_1 q_2} \dots a_{q_{L-1} q_L} b_{q_L}(x_L). \tag{3.3}$$

Evaluating Equation 3.3 needs $O(LM^L)$ operations. However, the forward-backward algorithm solves Equation 3.3 in $O(ML)$ by taking advantage of dynamic programming and two variables called forward and backward variables (Rabiner, 1989). The forward variable is defined as

$$\alpha_s(j) = \Pr((x_1, \dots, x_j), q_j = s \mid \delta).$$

Loosely spoken, for a given HMM δ the forward variable gives the probability to produce prefix $(x_1, \dots, x_j)$ of the observation x and end up in state s. The backward variable is defined as

$$\beta_s(j) = \begin{cases} \Pr((x_{j+1}, \dots, x_L), q_j = s), & \text{for } 1 \leq j \leq L, \\ 1, & \text{for } j = L. \end{cases}$$

For a given HMM δ, the backward variable gives the probability to obtain suffix $(x_{j+1}, \dots, x_L)$ of the observation x by starting from state s. Both variable are recursively defined and can be computed with dynamic programming. Combining the forward and backward variable leads to the forward-backward algorithm. The forward-backward algorithm can also be used to compute a further important measure, namely the posterior probability γ. The posterior probability $\gamma_s(j)$ is defined as the probability of being in state s at time j, that is,

$$\gamma_s(j) = \Pr(q_j = s \mid \delta, x). \tag{3.4}$$

It can be shown that the posterior probability $\gamma_s(j)$ is given by

$$\gamma_s(j) = \frac{\alpha_s(j)\beta_s(j)}{\sum_j^L \alpha_s(j)\beta_s(j)}.$$

The second problem is about finding a state sequence that maximizes its likelihood for a given HMM δ and observation x, that is, one has to evaluate

$$\hat{q} = \arg\max_q \Pr(x, q \mid \delta).$$

The Viterbi algorithm which works similar to the forward-backward algorithm can solve the problem in $O(M^2L)$ (Rabiner, 1989). The most likely state sequence is therefore also called Viterbi path.

The last problem is about the maximum likelihood estimation of an HMM δ for a given observation x, that is, one has to evaluate

$$\hat{\delta} = \arg\max_{\delta} \Pr(x \mid \delta). \tag{3.5}$$

Equation 3.5 cannot be solved analytically. A popular numeric maximization approach is the Baum-Welch algorithm, a specific instance of the EM-algorithm for HMMs. To apply the Baum-Welch algorithm for a discrete HMM with emission distribution b, the following equation has to be solved, either numerically or analytically,

$$\hat{\Theta} \in \arg\max_{\Theta} \sum_{s=0}^{M} \sum_{j=0}^{L} \gamma_s(j) \log b_s(y). \tag{3.6}$$

This chapter is based on Couvreur (1996). For more details and proofs, please see Rabiner (1989). Also, to get a deeper understanding of the Baum-Welch algorithm, we refer the reader to Bilmes (1998).

3.3.2. HMM for Differential Peak Calling

We model the differential peak calling problem with a three state HMM, which receives a $D \times L$ dimensional signal matrix $\mathbf{X}$ as input. The signals are the ChIP-Seq profiles after the application of all preprocessing steps described in Section 3.2. This first order HMM contains a state representing DPs gained in the first biological condition G_1 (`Gain 1`), a state for DPs gained in the second biological condition G_2 (`Gain 2`) and a background state (`Back`). We will call DPs to be `Gain 1` (`Gain 2`) for all competing methods, whenever they are detected to have higher signal in $x_{G_1\cdot}$ ($x_{G_12\cdot}$). Figure 3.3 shows the HMM topology, where all states have transitions to all other states and to themselves. We constrain the emission distribution to avoid label switching and to reduce the number of free parameters of the HMM.

The main idea to obtain candidate DPs is to first train an HMM with the Baum-Welch algorithm and then derive the most likely state sequence from the given data using the Viterbi algorithm. The state sequence is then associated with genomic regions exhibiting a DP in the first or second condition. This strategy depends crucially on the HMM's emission distribution that has to properly reflect the distribution of $\mathbf{X}$. In case with replicates, overdispersion typically occurs and has to be considered by the emission distribution. The application of the HMM to the signal matrix $\mathbf{X}$ is the second step in Algorithm 3.1.

3.3.3. Emission Distribution

For a given state s and observation $x_{\cdot j}$, the emission distribution b_s of the HMM is given by the product of probabilities for each biological condition G, that is,

$$b_s(y) = \Pr{}_s(y = x_{\cdot j} \mid q_j = s) = \textstyle\prod_{k \leq |G|} \Pr{}_{sk}(x_{G_k j}). \tag{3.7}$$

The probability of observing $x_{G_k j}$ in state s and condition k is given by the product of the

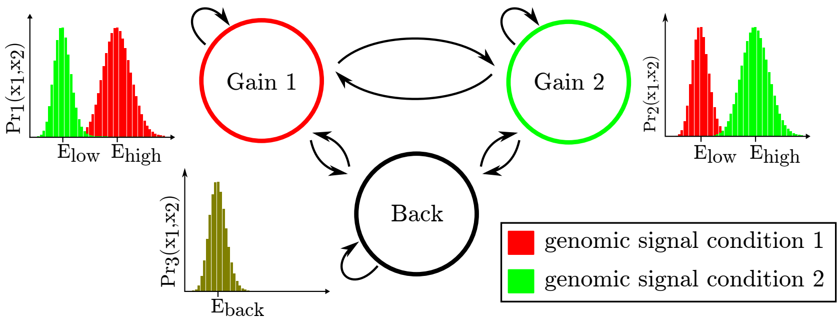


Figure 3.3.: HMM topology. The emission distributions are assigned to each state. To avoid label switching and to reduce number of free parameters, we constraint several parameters of the emission distributions. That is, the location parameter E_{low} and E_{high} of the emission distribution that are associated with state `Gain 1` *and state* `Gain 2` *are equal across the conditions. State* `Back` *exhibit location parameter E_{back}. In case with replicates, the HMM has a Negative Binomial distribution as emission. The location parameters μ_{low} and μ_{high} correspond to E_{low} and E_{high}. Furthermore, we set μ_{low} equal to E_{back}. In case without replicates, the HMM has either a Binomial or a mixture of Poisson distributions as emission. For the Binomial, we have $E_{low} = np_{low}$, $E_{high} = np_{high}$, and $E_{back} = np_{back}$. For the mixture of Poisson distributions, we use λ_{si1} respectively.*

observation's probabilities associated with condition k

$$\Pr{}_{sk}(x_{G_k j}) = \textstyle\prod_{i\in G_k} \Pr{}_{sk}(x_{ij}). \tag{3.8}$$

In case $D = 2$ without replicates, Equation 3.8 consists only of one factor, as G_1 and G_2 contain only one element. In the following, we will give the HMM emission distribution for three cases: Binomial, mixture of Poissons and Negative Binomial.

HMM without Replicates

In case without replicates two ChIP-seq profiles describing two biological conditions have to be taken into account. DPs are defined by changes among both profiles. The counts of both signals are modeled by a 2-dimensional emission distribution. We choose a Binomial distribution, as it models the number of successes in a sequence of independent Bernoulli experiments, that is, experiments with either a true or false outcome. Given a hypothetical Bernoulli experiment, the reads can either fall into the particular bin (true outcome) or into all other bins (false outcome). Hence, the number of reads in a genomic bin is modeled by a Binomial distribution.

More formally, we only have one ChIP-seq signal i per condition k. Equation 3.7 gives the emission distribution b_s for state s for a Binomial distribution by

$$b_s(y) = \Pr{}_s(y = x_{\cdot j} \mid q_j = s) = \textstyle\prod_{k\le|G|}\prod_{i\in G_k}\binom{n}{x_{\cdot j}} p_{sG_k}^{x_{ij}}(1-p_{sG_k})^{n-x_{ij}}. \tag{3.9}$$

with free parameters

$$\Theta = \{p_{sG_k}\}_{s=1,2,3,k=1,2} \cup \{n\}.$$

Parameter n is independent of state s and represents the number of reads of the largest library

$$n = \max\left(\sum_{j=1}^{L} x_{G_1 j}, \sum_{j=1}^{L} x_{G_2 j}\right).$$

Parameter p_{sk} is the probability of observing a read in state s and condition k.

For a large number of Bernoulli experiments, the Binomial distribution approximates the Poisson distribution. As we have a large number of reads in a ChIP-seq experiment, we therefore also evaluate the Poisson distribution as emission. We additionally extend our model by using a mixture of Poisson distributions. The rationale is that a mixture model is suitable to model outliers in the count data. In our case, due to various sources of bias in the ChIP-seq protocol, there are usually bins with unexpectedly large numbers of reads that we want to model. The Poisson distribution describes the probability of a given number of events occurring in a fixed interval. Similar to the Binomial distribution, we thereby model the number of reads falling into a genomic bin.

More formally, we have as emission distribution

$$b_s(y) = \Pr{}_s(y = x_{\cdot j} \mid q_j = s) = \textstyle\prod_{k \leq |G|} \prod_{i \in G_k} \sum_{l=1}^{N} c_{sl} \cdot b_{skl}(x_{ij}), \tag{3.10}$$

where N is the number of mixture components, where matrix $c \in \mathbb{R}^{M \times N}$ gives the mixing coefficient of the mixture for each state $s \in [1, 2, \ldots, M]$ (here $M = 3$) and component $l \in [1, 2, \ldots, N]$ with $c_{sl} \in [0, 1]$ and $\sum_{l=1}^{N} c_{sl} = 1$ for each state, and where

$$b_{skl}(x_{ij}) = \frac{\exp(-f(l) \cdot \lambda_{sk1}) \cdot (f(l) \cdot \lambda_{sk1})^{x_{ij}}}{x_{ij}!}$$

gives the Poisson component of the mixture model. We use the function $f(l) = l$ to ensure that the mean of each mixture component are multiple of each other. This is equivalent to the constraint

$$\lambda_{skl} = l \cdot \lambda_{sk1}. \tag{3.11}$$

This constraint was introduced to mitigate the problem that during mixture estimation some components can end up with little data support (or low mixing coefficients). This is usual when outliers (peaks with unusually large numbers of reads) are present in the data. In the case of the mixture of Poisson distributions, we have

$$\Theta = \{\lambda_{sk1}, c_{sl}\}_{s=1,2,3, k=1,2, l=\{1,\ldots,N\}}$$

as free parameters.

HMM with Replicates

In the case with replicates we model a DP by a D-dimensional emission distribution. There is additionally variance within the conditions which makes it in general harder to properly model the observations with the HMM emission distribution. In particular, given various ChIP-seq profiles of one condition, it is likely to observe overdispersion, that is, that the mean exceeds the variance. The Binomial and Poisson distribution cannot cope with overdispersion, as their variance linearly depends one the mean (Ismail and Jemain, 2007). We therefore use the Negative Binomial distribution as emission, since it is able to take overdispersion into account. Indeed, it can be shown that the Poisson distribution where the mean is separately drawn from a Gamma distribution results in a Negative Binomial distribution (Cameron and Trivedi, 2001). As there is no analytical solution for the Baum-Welch algorithm with a Negative Binomial distribution, we will show how to estimate the distribution.

More formally, Equation 3.7 gives the emission distribution b_s for state s, condition k and sample *i* by

$$
\begin{aligned}
b_s(y) &= \Pr{}_s(y = x_{\cdot j} \mid q_j = s) \\
&= \prod_{k \leq |G|} \prod_{i \in G_k} g(x_{ij} \mid \Theta_s = \{\mu_{sG_k}, a_{sG_k}\}) \qquad (3.12) \\
&= \prod_{k \leq |G|} \prod_{i \in G_k} \frac{\Gamma(x_{ij} + a_{sG_k}^{-1})}{\Gamma(x_{ij}+1)\cdot\Gamma(a_{sG_k}^{-1})} \cdot \left(\frac{a_{sG_k}^{-1}}{a_{sG_k}^{-1} + \mu_{sG_k}}\right)^{a_{sG_k}^{-1}} \cdot \left(\frac{\mu_{sG_k}}{a_{sG_k}^{-1} + \mu_{sG_k}}\right)^{a_{sG_k}^{-1}}, \qquad (3.13)
\end{aligned}
$$

with free parameters

$$\Theta = \{\mu_{sG_k}, a_{sG_k}\}_{s=1,2,3,k=1,2},$$

where a_{sG_k} is the dispersion parameter, μ_{sG_k} the location parameter and Γ the gamma function. The Negative Binomial distribution g has mean $\mathbb{E}(x_i) = \mu_{sG_k}$ and variance

$$Var(x_i) = \mu_{sG_k}(1 + a_{sG_k}\mu_{sG_k}). \qquad (3.14)$$

If $a_{sG_k} = 0$, the mean equals the variance and the distribution results in a Poison distribution. For $a_{sG_k} > 0$, variance increases with mean as usual when dealing with NGS data containing replicates (Anders and Huber, 2010).

3.3.4. HMM Training

The HMM is estimated with the Baum-Welch algorithm. Estimates of the initial state and transition probabilities follow usual methods (Rabiner, 1989). Training is performed until convergence. See for example Couvreur (1996) for a gentle introduction to the Baum-Welch algorithm. For a given emission distribution, we have to evaluate Equation 3.6 to obtain the estimates for the emission in the Baum-Welch algorithm. In the following we will explain how to compute the estimate for the Binomial, mixture of Poisson and the Negative Binomial distribution.

Binomial Distribution as Emission without Replicates

Equation 3.9 gives the HMM emission based on a Binomial distribution with free parameters

$$\Theta = \{p_{sG_k}\}_{s=1,2,3,k=1,2} \cup \{n\}.$$

To reduce the number of parameter estimates, we constrain the parameters from `Back` state ($s = 3$) to be equal $p_{\text{back}} = p_{31} = p_{32}$. We also constrain emissions for state `Gain 1` (s=1) and state `Gain 2` (s=2) by $p_{\text{low}} = p_{1G_1} = p_{2G_2}$ and $p_{\text{high}} = p_{1G_2} = p_{2G_1}$ to avoid label switching (Rabiner, 1989). This makes the distributions of enriched signals (non-enriched signals) from states `Gain 1` and `Gain 2` equal (Figure 3.3). In our case, we have $|G| = 2$ conditions and $M = 3$ HMM's states. We only have to solve 3 optimization problems, that is, determining p_{back}, p_{high} and p_{low}, to solve Equation 3.6. Here we show the estimation of p_{high}. The other parameter estimates follow in a similar manner. As we do not have replicates, we rewrite

Equation 3.6 as

$$
\begin{aligned}
&\underset{p_{\text{high}}\in\Theta}{\arg\max}\sum_{s=1}^{M}\sum_{j=1}^{L}\gamma_s(j)\log\left(\sum_{k\leq|G|}\sum_{i\in G_k}\binom{n}{x_{ij}}p_{1G_k}^{x_{ij}}(1-p_{1G_k})^{n-x_{ij}}\right)\\
=&\underset{p_{\text{high}}\in\Theta}{\arg\max}\sum_{s=1}^{M}\sum_{j=1}^{L}\gamma_1(j)\log\left(\binom{n}{x_{1j}}p_{sG_1}^{x_{1j}}(1-p_{sG_1})^{n-x_{ij}}\right)\\
&+\sum_{s=1}^{M}\sum_{j=1}^{L}\gamma_2(j)\log\left(\binom{n}{x_{2j}}p_{sG_2}^{x_{2j}}(1-p_{sG_2})^{n-x_{2j}}\right)\\
=&\underset{p_{\text{high}}\in\Theta}{\arg\max}\, f(p_{\text{high}}=p_{1G_1})+h(p_{\text{high}}=p_{2G_2})
\end{aligned}
$$

To compute the maximum of the function $f+h$, we first compute the derivative for p_{high}

$$
\begin{aligned}
\frac{f}{\delta p_{\text{high}}}+\frac{h}{\delta p_{\text{high}}}=&\sum_{j=1}^{L}\gamma_1(j)\frac{x_{1j}\binom{n}{x_{1j}}p_{1G_1}^{x_{1j}-1}(1-p)^{n-x_{1j}}-\binom{n}{x_{1j}}p_{1G_1}^{x_{1j}}(n-x_{1j})(1-p_{1G_1})^{n-x_{1j}-1}}{\binom{n}{x_{1j}}p^{x_{1j}}(1-p_{1G_1})^{n-x_{ij}}}\\
&+\sum_{j=1}^{L}\gamma_2(j)\frac{x_{2j}\binom{n}{x_{2j}}p_{1G_2}^{x_{2j}-1}(1-p)^{n-x_{2j}}-\binom{n}{x_{2j}}p_{1G_2}^{x_{2j}}(n-x_{2j})(1-p_{1G_2})^{n-x_{2j}-1}}{\binom{n}{x_{2j}}p^{x_{2j}}(1-p_{1G_2})^{n-x_{2j}}}\\
=&\sum_{j=1}^{L}\gamma_1(j)\frac{x_{1j}-np_{1G_1}}{p_{1G_1}(1-p_{1G_1})}+\sum_{j=1}^{L}\gamma_2(j)\frac{x_{2j}-np_{1G_2}}{p_{1G_2}(1-p_{1G_2})}.
\end{aligned}
$$

Then, we obtain the estimate of the maximum of the function $f+h$ for p_{high}:

$$
\hat{p}_{\text{high}}=\sum_{j=1}^{N}\frac{\gamma_1(j)x_{1j}+\gamma_2(j)x_{2j}}{n\cdot\gamma_1(j)+n\cdot\gamma_2(j)}.
$$

The other parameters can be computed accordingly. We obtain

$$
p_{\text{low}}=p_{1G_2}=p_{2G_1}=\sum_{j=1}^{N}\frac{\gamma_1(j)x_{2j}+\gamma_2(j)x_{1j}}{n\cdot\gamma_1(j)+n\cdot\gamma_2(j)},\quad\text{and}
$$

$$
p_{\text{back}}=p_{3G_1}=p_{3G_2}=\sum_{j=1}^{N}\frac{\gamma_3(j)x_{1j}+\gamma_3(j)x_{2j}}{2\cdot n\cdot\gamma_3(j)}.
$$

Mixture of Poisson Distributions as Emission without Replicates

The HMM emission based on the mixture of Poisson distributions is described by Equation 3.10. We use the Q function (see Section 4.2 in Bilmes (1998)) to obtain the equations for the EM-algorithm.

With regard to our constraint described in Equation 3.11, we consequently have to solve

$$
\max_{\Theta_{sk1}\in\mathbb{R}}\sum_{s=1}^{M}\sum_{l=1}^{N}\sum_{j=1}^{L}\log b_{sil}(x_{ij})\cdot p(O,q_j=s,m_{q_jj}=l\mid\lambda'),
$$

where $q=(q_1,\ldots,q_L)$ is a sequence of states and where $s_j\in\{1,\ldots,N\}$ is the state at time j. Furthermore m is a vector that indicates the mixture component for each state at each time. We obtain λ_{si1} for the first component as

$$\lambda_{si1} = \frac{\sum_{j=1}^{L} \sum_{l=1}^{N} x_{ij} \cdot r_{sl}(j)}{\sum_{j=1}^{L} \sum_{l=1}^{N} f(l) \cdot r_{sl}(j)},$$

where $r_{sl}(j)$ is the posterior probability of being in state s at time j with regard to the component l. The mixing coefficients c_{sl} and posterior probabilities $r_{sl}(j)$ follow standard parameterizations and are defined as

$$c_{sl} = \frac{\sum_{j=1}^{L} r_{sl}(j)}{\sum_{j=1}^{L} \gamma_s(j)}, \text{ and}$$

$$r_{sl}(j) = \gamma_s(j) \cdot \frac{c_{sl} \cdot b_{sl}(x_{ij})}{\sum_{l=1}^{N} c_{sl} \cdot b_{sl}(x_{ij})},$$

Furthermore, we compute λ_{si1} for the component l as

$$\lambda_{sil} = l \cdot \lambda_{si1},$$

where $\gamma_s(j)$ is the posterior probability to be at state s at time j.
We constrain the mixture distribution for each component l accordingly to the case of the Binomial distribution $\lambda_{11l} = \lambda_{22l}, \lambda_{12l} = \lambda_{21l}$ and $\lambda_{31l} = \lambda_{32l}$. All other parameters follow standard mixture model estimates.

Negative Binomial Distribution as Emission with Replicates

Equation 3.13 gives the HMM emission based on a Negative Binomial distribution with free parameters

$$\Theta_{sG_k} = \{a_{sG_k}, \mu_{sG_k}\},$$

for state s and condition k, where a_{sG_k} is the dispersion parameter and where μ_{sG_k} gives the location. For parameters Θ_{sG_k} of the Negative Binomial distribution, Equation 3.6 cannot be solved analytically. Instead we estimate μ_{sG_k} and a_{sG_k} based on a moment approach.

In our case, we have $|G| = 2$ conditions and $M = 3$ HMM's states. Given Equation 3.6, we therefore have to solve 6 optimization problems for each condition and state. We constrain location parameters of `Gain 1` (s=1) and state `Gain 2` (s=2) associated with enriched signals to be equal $\mu_{1G_1} = \mu_{2G_2} = \mu_{\text{high}}$. We also constrain location parameters of low values and background states to be equal $\mu_{1G_2} = \mu_{2G_1} = \mu_{3G_1} = \mu_{3G_2} = \mu_{\text{low}}$ (see Figure 3.3). This avoids label switching problems in the HMM (Rabiner, 1989). Consequently, we only have to solve 2 optimization problems, that is, determining μ_{high} and μ_{low}, to solve Equation 3.6. Here we show the estimation of $\mu_{\text{high}} = \mu_{11} = \mu_{22}$.

In Equation 3.6, we restrict our optimization space and obtain

$$\begin{aligned}
&\underset{\mu_{\text{high}} \in \Theta}{\text{arg max}} \sum_{s=1}^{M} \sum_{k \leq |G|} \sum_{i \in G_k} \sum_{j=1}^{L} \gamma_s(j) \log g(x_{ij} | \mu_{sG_k}) + \sum_{s=1}^{M} \sum_{k \leq |G|} \sum_{i \in G_k} \sum_{j=0}^{L} \gamma_s(j) \log g(x_{ij} | \mu_{sG_k}) \\
=&\underset{\mu_{\text{high}} \in \Theta}{\text{arg max}} \sum_{s=1}^{M} \sum_{i \in G_1} \sum_{j=1}^{L} \gamma_s(j) \log g(x_{ij} | \mu_{sG_1}) + \sum_{s=1}^{M} \sum_{i \in G_2} \sum_{j=0}^{L} \gamma_s(j) \log g(x_{ij} | \mu_{sG_2}) \\
=&\underset{\mu_{\text{high}} \in \Theta}{\text{arg max}}\, f(\mu_{\text{high}})
\end{aligned}$$

We define a function f depending on μ_{high}. As we want to optimize f, we take the derivative of f. The derivative of f is restricted to the case $s = 1$ and $s = 2$.

$$\begin{aligned}\frac{f}{\delta\mu_{\text{high}}} &= \frac{\sum_{i\in G_1}\sum_{j=0}^{L}\gamma_1(j)\log g(x_{ij}|\mu_{1G_1})}{\delta\mu_{\text{high}}} &&+ \frac{\sum_{i\in G_2}\sum_{j=0}^{L}\gamma_2(j)\log g(x_{ij}|\mu_{2G_2})}{\delta\mu_{\text{high}}}\\ &= \frac{f_1}{\delta\mu_{\text{high}}} &&+ \frac{f_2}{\delta\mu_{\text{high}}}\end{aligned} \tag{3.15}$$

Sums containing μ_{2G_1} and μ_{1G_2} are constants while deriving f with regard to μ_{high} and therefore are no longer considered. To simplify the notation, we introduce functions f_1 and f_2, which we have to derive separately to obtain the derivative of f. Accordingly to Ismail and Jemain (2007), we can rewrite Equation 3.12 as

$$\begin{aligned}g(x_{ij}|\Theta_{sG_k}) = &\left(\sum_{h=1}^{x_{ij}-1}\ln(1+a_{sG_k}h)\right) - x_{ij}\cdot\ln(a_{sG_k}) - \ln(x_{ij}!) + x_{ij}\cdot\ln(a_{sG_k}\cdot\mu_{sG_k})\\ &-(x_{ij}+a_{sG_k}^{-1})\cdot\ln(1+a_{sG_k}\cdot\mu_{sG_k})\end{aligned} \tag{3.16}$$

We plug in Equation 3.16 in function f_1 of Equation 3.15. The derivative of f_1 is given by

$$\frac{f_1}{\delta\mu_{\text{high}}} = \sum_{i\in G_2}\sum_{j=0}^{L}\gamma_1(j)\frac{x_{ij}-\mu_{\text{high}}}{\mu_{\text{high}}+a_1\mu_{\text{high}}^2}$$

The derivative estimation for f_2 works similar. We plug in $f_1/\delta\mu_{\text{high}}$ and $f_2/\delta\mu_{\text{high}}$ in Equation 3.15, set $f/\delta\mu_{\text{high}}$ to 0 and obtain the parameter $\hat{\mu}_{\text{high}}$ that optimize function f, that is,

$$\begin{aligned}\frac{f}{\delta\mu_{\text{high}}} &\overset{!}{=} 0 = \sum_{i\in G_1}\sum_{j=0}^{L}\gamma_1(j)\frac{x_{ij}-\mu_{\text{high}}}{\mu_{\text{high}}+a_1\mu_{\text{high}}^2} + \sum_{i\in G_2}\sum_{j=0}^{L}\gamma_2(j)\frac{x_{ij}-\mu_{\text{high}}}{\mu_{\text{high}}+a_1\mu_{\text{high}}^2}\\ \Rightarrow \qquad \hat{\mu}_{\text{high}} &= \frac{\sum_{i\in G_1}\sum_{j=0}^{L}\gamma_1(j)x_{ij}+\sum_{i\in G_2}\sum_{j=0}^{L}\gamma_2(j)x_{ij}}{|G_1|\sum_{j=0}^{L}\gamma_1(j)+|G_2|\sum_{j=0}^{L}\gamma_2(j)}\end{aligned}$$

Parameter $\hat{\mu}_{\text{low}}$ is computed in a similar manner.

To obtain reliable variance estimates on small sample sizes, we assume that the variance can be described by a smooth function based on the mean estimates similar as described by Anders and Huber (2010). We use a quadratic function

$$v_k(x) = c_{1G_k}\cdot x^2 + x + c_{2G_k}, \tag{3.17}$$

which is estimated for the ChIP-seq data on samples of condition k previous to the Baum-welch algorithm. The dispersion parameter a_{sk} is derived from Equation 3.14 and given by

$$a_{sG_k} = \frac{v_k(\mu_{sG_k})-\mu_{sG_k}}{\mu_{sG_k}^2}.$$

We apply the Viterbi algorithm to estimate a state sequence for the complete genomic signal. Finally, we merge consecutive bins associated with states `Gain 1` or `Gain 2` to obtain the candidate DPs.

3.3.5. Initial HMM Estimates

The Baum-Welch algorithm depends on the initial estimates to train the HMM. We use potential DPs as initial estimates and use two kinds of criteria to define them. First, a fold change criterion to quantify the intensity of a potential DP and second, a minimum signal criterion to avoid DP exhibiting too low counts. Based on these criteria, we annotate bins as initial DPs. We then use the initial DPs to obtain a posterior probability as well as to perform a single M-Step of the Baum-Welch algorithm to compute initial parameters. Given the large size of the genomic signals with initial DPs, we only use a random selection of regions to train the HMM. We select genomic regions formed by contiguous bins not filtered out in Section 3.2, which have at least a bin annotated with either `Gain 1` or `Gain 2` state. Depending on the absence of replicates, we follow two strategies to define initial DPs.

HMM without Replicates

A bin b_j will be assigned to state `Gain 1` if

$$x_{G_1 j}/x_{G_2 j} > t,$$

to state `Gain 2` if

$$x_{G_2 j}/x_{G_1 j} > t$$

and to `Back` state otherwise.

HMM with Replicates

For state `Gain 1`, we select bins if there is a difference t_1 in counts between two signals

$$\bar{x}_{G_1 j} - \bar{x}_{G_2 j} > t_1,$$

or if there is a high fold change t_2 and minimum signal support t_3

$$\bar{x}_{G_1 j}/\bar{x}_{G_2 j} > t_2 \quad \text{and} \quad \bar{x}_{G_1 j} + \bar{x}_{G_2 j} > t_3.$$

DPs associated with state `Gain 2` are defined accordingly. Otherwise, the bin is assigned the `Back` state.

3.4 Postprocessing Steps

We perform postprocessing steps to improve the DP estimation. First, we assign a p-value to each DP to evaluate how significant the difference between the two biological conditions is. Then, we remove DPs that are likely false positives caused by technical issues due to the ChIP-seq protocol (Pepke et al., 2009). This postprocessing pipeline is the third step in Algorithm 3.1.

3.4.1. *P*-Value Calculation

Statistical hypothesis testing is about the analysis of empirically collected data and in particular about the question whether the data provide enough evidence to reject a stated null hypothesis. We assume the null hypothesis to be true unless there is strong evidence in the data against it. If so, we reject the null hypothesis and assume the alternative hypothesis to be true. For a given null hypothesis, the p-value is a function describing the probability of obtaining a result equal or more extreme than the observed results. If the p-value is sufficient low, the observed result does not go in accordance with the null hypothesis which is then rejected. In our case, under the assumption that the count distribution of the first and second biological conditions describes the genomic background signal, the p-value gives the probability for a candidate DP. If the p-value is low, the null hypothesis is rejected and the potential DP is considered to be a true candidate DP.

More formally, we follow the idea of Anders and Huber (2010) to assign a p-value to each DP. Let

$$y_1 = \sum_{j=u}^{v} x_{G_1 j} \quad \text{and} \quad y_2 = \sum_{j=u}^{v} x_{G_2 j}$$

be the read counts of a DP spanning from bin u to v with two biological conditions G_1 and G_2. For a DP gaining a peak in condition G_1, the p-value is the sum of probabilities of the tuple (a,b) with $a > y_1$ and $a+b = y_1+y_2$. In other words, given a DP with counts (y_1, y_2), we add up all probabilities of tuples with more extreme values. More extreme values a defined as $a > y_1$ for a fixed margin sum $a+b = y_1+y_2$.
More formally,

$$\Pr(a > y_1 | y_2) = \sum_{\substack{a+b=y_1+y_2 \\ a>y_1}} \Pr(a,b), \tag{3.18}$$

where $a, b \in \mathbb{N}$. We compute the probability $\Pr(a,b)$ as

$$\Pr(a,b) = \frac{\Pr(a \mid s=3, \Theta_{31}) \cdot \Pr(b \mid s=3, \Theta_{32})}{\sum_{c+d=a+b} \Pr(c \mid s=3, \Theta_{31}) \cdot \Pr(d \mid s=3, \Theta_{32})}, \tag{3.19}$$

where $c, d \in \mathbb{N}$ with $c+d = a+b$. We use the distribution of the HMM's `Back` state $(s=3)$, that is, the genomic background signal, to compute the p-value. The p-value for DPs gaining a peak in condition G_2 can be defined accordingly.

For large y_1, y_2 values, the computation of the sums in Equation 3.18 and Equation 3.19 are computationally expensive which makes improvements in the formalization for a faster p-value calculation necessary. We combine both equations and obtain

$$\Pr(a > y_1 | y_2) = \frac{\sum_{\substack{a+b=y_1+y_2 \\ a>y_1}} \Pr(a \mid s=3, \Theta_{31}) \cdot \Pr(b \mid s=3, \Theta_{32})}{\sum_{c+d=y_1+y_2} \Pr(c \mid s=3, \Theta_{31}) \cdot \Pr(d \mid s=3, \Theta_{32})}. \tag{3.20}$$

The sum of the nominator is a subset of the sum of the denominator. Consequently, we only need to evaluate the sum of the denominator and take into account the appropriate values for the nominator.

We model the probability of the counts with the HMM's emission distribution that is assigned to the `Back` state. The distribution has parameters $\Theta_{31} = \Theta_{32}$. In the following we assume that a Binomial distribution B is assigned to the `Back` state of the HMM. Given that $b = y_1 + y_2 - a$ we can then rewrite the main term in the nominator (and denominator) as a combination of Binomial distributions:

Definition 3.1 (Combined Binomial (CB) Distribution). *Let B a Binomial distribution, the combined Binomial (CB) distribution is defined as*

$$f(x) = B(x \mid n,p) \cdot B(y_1 + y_2 - x \mid n,p), \quad \textit{with}$$

$$B(x \mid n,p) = \binom{n}{x} p^x (1-p)^{n-x},$$

with fixed $y_1, y_2 \in \mathbb{N}$ and where $n \in \mathbb{N}$ $p \in [0,1]$ are the parameters of the Binomial distribution.

To further improve the computation of Equation 3.20, we show that DB distributions are axially symmetrical and have a global maximum at $q = (y_1 + y_2)/2$ given that $p_{31} = p_{32}$. To investigate these characteristics, we first give a statement about the binomial coefficient.

Lemma 3.2. *Let $\binom{n}{k}$ be the binomial coefficient with $n, k \in \mathbb{B}$ and $k \leq n$. It is*

$$\binom{n}{k} = \binom{n}{k-1} \frac{n-k+1}{k}.$$

Proof.

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{n!}{(k-1)!(n-k+1)!} \cdot \frac{n-k+1}{k} = \binom{n}{k-1} \frac{n-k+1}{k}$$

□

We show that the DB distribution is symmetrical.

Lemma 3.3. *Let f be a CB distribution, f is symmetrical to the y-axis parallel going through the point $q = (y_1 + y_2)/2$.*

Proof. Let $d \in \mathbb{N}$, it is

$$\begin{aligned} f(q+d) &= B(\frac{y_1+y_2}{2} + d \mid n, p) \cdot B(y_1 + y_2 - \frac{y_1+y_2}{2} - d \mid n, p) \\ &= B(y_1 + y_2 - (\frac{y_1+y_2}{2} - d) \mid n, p) \cdot B(\frac{y_1+y_2}{2} - d \mid n, p) \\ &= f(\frac{y_1+y_2}{2} - d) \\ &= f(q-d). \end{aligned}$$

□

Next, we estimate the maximum of a CB distribution. We will show that the CB distribution monotonically decrease starting from point $q = (y_1 + y_2)/2$. Due to its symmetrical property, we conclude the maximum at point q.

Lemma 3.4. *Let f be a CB distribution, f monotonically decreases starting from point $q = (y_1 + y_2)/2$, $q \in \mathbb{N}$, for all $x \geq q$.*

Proof. We will use complete induction with the induction hypothesis

$$\frac{f(q+k)}{f(q+k+1)} > 1, \tag{3.21}$$

with $k \in \mathbb{N}$ and $k \geq 0$.
First, we start with the base case $k = 0$.

$$\begin{aligned}\frac{f(q)}{f(q+1)} &= \frac{\binom{n}{q}p^{q}(1-p)^{n-q}\binom{n}{q}p^{q}(1-p)^{n-q}}{\binom{n}{q+1}p^{q+1}(1-p)^{n-q-1}\binom{n}{q-1}p^{q-1}(1-p)^{n-q+1}}\\ &= \frac{(q+1)!(n-q-1)!(q-1)!(n-q+1)!)}{q!(n-q)!q!(n-q)!}\\ &= \frac{q+1}{q}\cdot\frac{n-q+1}{n-q} > 1.\end{aligned}$$

We assume that the induction hypothesis in Equation 3.21 holds for k. In the induction step, we will proof that the induction hypothesis holds for the case $k+1$ as well. Here, we use Lemma 3.2 to bring the binomial coefficient in the right form.

$$\begin{aligned}\frac{f(q+k+1)}{f(q+k+2)} &= \frac{\binom{n}{q+k+1}p^{q+k+1}(1-p)^{n-q-k-1}\binom{n}{y_1+y_2-q-k-1}p^{y_1+y_2-q-k-1}(1-p)^{n-y_1-y_2+q+k+1}}{\binom{n}{q+k+2}p^{q+k+2}(1-p)^{n-q-k-2}\binom{n}{y_1+y_2-q-k-2}p^{y_1+y_2-q-k-2}(1-p)^{n-y_1-y_2+q+k+2}}\\ &\overset{3.2}{=} \frac{\binom{n}{q+k}\frac{n-q-k}{q+k+1}p^{q+k}p(1-p)^{n-q-k}\binom{n}{y_1+y_2-q-k}\frac{y_1+y_2-q-k}{n-y_1-y_2+q+k+1}p^{y_1+y_2-q-k}}{\binom{n}{q+k+1}\frac{n-q-k-2}{q+k+2}p^{q+k+1}p(1-p)(1-p)^{n-q-k-2}}\\ &\quad\cdot\frac{p^{-1}(1-p)^{n-y_1-y_2+q+k}(1-p)}{\binom{n}{y_1+y_2-q-k-1}\frac{y_1+y_2-q-k-1}{n-y_1-y_2+q+k+2}p^{y_1+y_2-q-k-1}p^{-1}(1-p)^{n-y_1-y_2+q+k+1}(1-p)}\\ &= \underbrace{\frac{f(q+k)}{f(q+k+1)}}_{>1,\text{ induction}}\cdot\underbrace{\frac{n-q-k}{n-q-k-2}}_{>1}\cdot\underbrace{\frac{y_1+y_2-q-k}{y_1+y_2-q-k-1}}_{>1}\cdot\underbrace{\frac{q+k+2}{q+k+1}}_{>1}\cdot\underbrace{\frac{n-y_1-y_2+q+k+2}{n-y_1-y_2+q+k+1}}_{>1}\end{aligned}$$

□

We can now state that a CB distribution has its maximum at $q = (y_1+y_2)/2, q \in \mathbb{N}$.

Lemma 3.5. *Let f be a CB distribution. Function f has its maximum at $q = (y_1+y_2)/2$, $q \in \mathbb{N}$.*

Proof. Lemma 3.4 state that function f monotonically decreases from point $q = (y_1+y_2)/2$ for all $x > q$, $q \in \mathbb{N}$. As function f is symmetrical to a y-axis parallel going through point q (Lemma 3.3), it is clear that f has a maximum at $q, q \in \mathbb{N}$. If $q \notin \mathbb{N}$, f has two maxima at $q_1 = \lfloor (y_1+y_2)/2 \rfloor$ and $q_2 = \lfloor (y_1+y_2)/2 \rfloor + 1$. For this case, it is clear that the main ideas of the lemmata presented here still hold. □

As CD distributions are symmetrical (Lemma 3.3), we only have to evaluate half of the sum's values of the numerator (denominator) of Equation 3.20. Furthermore, as DB distributions decrease monotonically (Lemma 3.4 and Lemma 3.5) departing from q, we can approximate the p-value calculation by making $f(e) = f(a)$ for all $e > a$ given that $f(a) - f(a+1) < \varepsilon$. These steps allow a speed up of 100 times on the p-value calculations on our experiments. The lemmata presented here do not hold for the mixture of Poisson distribution or the Negative Binomial distribution.

3.4.2. Filtering of ChIP-seq experimental artifacts

We perform several postprocessing steps to remove spurious DPs. The rationale is that the ChIP-seq protocol induces biases that lead to peaks in the ChIP-seq signal that are not caused by biological events (Pepke et al., 2009).

First, we ignore all DPs with a size smaller than the estimated fragment size $\hat{f}$. We also merge concordant DPs, which have a distance less than the estimated fragment size $\hat{f}$. The second step is only suggested for histone modification data, which is usually localized in broader genomic regions. P-values are re-estimated after merging and corrected for controlling the False Discovery Rate (Benjamini and Hochberg, 1995).

For the case with replicates, we use the mean of estimated fragment sizes. Additionally, false positive DPs may be caused by a high strand lag (Pepke et al., 2009). For each DP, we therefore count the forward and reverse reads, normalize the ratio by computing the z-scores and filter out all DPs that exhibit a high/low z-score. By default, we choose a z-scores threshold of 2 which corresponds to a two fold standard deviation from the normal distribution. Also, DPs falling into blacklisted genomic regions (see Section 2.2.5) are filtered out.

3.5 Implementation

We implemented our HMM-based strategy to detect DP in ChIP-seq signals associated with two biological conditions as Python command line tools. ODIN (One-stage DIffereNtial peak caller) comprise the HMM with a Binomial and mixture of Poisson distributions for the case of a ChIP-seq analysis without replicates. THOR can be seen as an extension of ODIN and provides an HMM with a Negative Binomial distribution to take replicates into account. Both tools perform the pre- and postprocessing steps described in Section 3.2 and Section 3.4.

THOR and ODIN are part of the Regulatory Genomics Toolbox (RGT) which provides functions for Python to handle genomic signals. In particular RGT gives an infrastructure to analyze ChIP-seq signals. THOR and ODIN are available under the terms of the *GNU General Public Licence v3 (GPL v3)*. At their current version (ODIN 0.4 and THOR 0.1), they exhibit 3253 and 2821 lines of code with 136 and 112 functions, respectively. ODIN was released in October 2014 and THOR was released in July 2015. The HMM both tools are using is implemented in the HMMlearn package (see `https://github.com/hmmlearn/hmmlearn`), which is based on the machine learning package Scikit-learn (Pedregosa et al., 2011).

ODIN and THOR take as input BAM files describing the aligned reads, a fasta file describing the genome to consider, and a tab separated file describing the chromosome size of the genome. They output bigwig files to describe the normalized ChIP-seq signal for each replicate and condition. Furthermore they give BED and narrowPeak files describing the identified DPs. Please see `https://genome.ucsc.edu/FAQ/FAQformat.html` for more details about the file formats. Besides HMMlearn, further important non-standard python packages that are used are HTSeq to process FASTA files, scipy to process BAM files and scipy and numpy to cope with the ChIP-seq signal.

We tested ODIN and THOR with Python 2.7, Numpy 1.4.0, Scipy 0.7, Scikit-learn 0.14, Pysam 0.7.5, HTSeq 0.6.5 and HMMlearn 0.0.1. We use a local Linux Ubuntu 14.04.4 LTS x86 64-bit machine running with 8 Intel(R) Core(TM) i7-4770 CPU at 3.40GHz and 16 GB RAM. Furthermore, we run both tools on an HPC cluster mainly based on Intel Xeon-based 8- to 128-way SMP 64-bit nodes with Scientific Linux release 6.6 (Carbon).
For more information how to use ODIN, please see

`www.regulatory-genomics.org/ODIN`,

and for THOR please see

`www.regulatory-genomics.org/THOR`.

All websites mentioned in this Section were accessed on 14th November, 2015.

3.6 Summary

In this chapter, we introduced our algorithms ODIN and THOR to identify DPs in ChIP-seq signals without and with replicates. Algorithm 3.1 gives an schematic overview of both methods. Our algorithms apply new concepts to solve the differential peak calling problem:

- we introduce the preprocessing steps that are necessary to make the ChIP-seq signals applicable to our algorithms. Importantly, we introduced a novel normalization strategy for the ChIP-seq profiles which is based on the use of control regions (see Section 3.2.6).

- ODIN and THOR apply an HMM to segment the signal by detecting peaks with variable size through the use of posterior decoding algorithms. In the case without replicates, we use a Binomial or a mixture of Poisson distributions as emission (see Section 3.3.3). The rationale is that the Binomial distribution models the number of successes in a sequence of independent Bernoulli experiments. We can see the ChIP-seq profile construction as a sequence of Bernoulli experiments, where each reads falls or does not fall into a particular genomic bin. Furthermore, the Binomial distribution approaches the Poisson distribution for a large number of Bernoulli experiments. In the case with replicates, we apply a Negative Binomial distribution as emission (see Section 3.3.3), as it is equivalent to the Poisson distribution, where the mean is separately drawn from a Gamma distribution. Furthermore, the Negative Binomial distribution accounts for overdispersion, which typically occurs when dealing with NGS count data.

- from a methodological aspect, the use of an HMM to segment the signal is the favourable method of choice. Window-based approaches (PePr, DiffReps, csaw, see Table 2.1) depend on heuristic methods and do not take advantage of decoding algorithms to call DPs. Moreover, using pre-defined peaks (DiffBind, MACS2, DBChIP, DESeq, MAnorm, see Table 2.1) highly depends on the strategy how to compute the initial peak set.

- we explained our postprocessing strategy to get rid of technical artifacts (see Section 3.4). Moreover, we compute a p-value for each DPs (see Section 3.4.1). The calculation is based on the emission distribution of the used HMMs.

CHAPTER 4

Experimental Methods

In the previous chapter, we introduced our algorithms ODIN and THOR to call DPs in ChIP-seq signals without and with replicates. Here, we describe the experimental setup for our evaluation studies. Evaluation and comparison of DPCs are challenging problems, as there are neither direct metrics to rate DP estimates nor datasets which could serve as gold standards in the evaluation procedure. We apply two alternative strategies to evaluate DPs. In this chapter, we first describe the simulation algorithm for ChIP-seq data to produce artificial gold standards. The algorithm is highly parametrized to enable producing various gold standards for the evaluation. Second, we propose indirect metrics by associating DPs with gene expression changes in the same cellular conditions. Moreover, we describe the biological data sets used for the evaluation studies. Next, we describe the experiments performed with and without replicates to evaluate THOR, ODIN and their competing methods. Also, we describe two use cases for THOR. The first use case describes the ability of THOR to call DPs that support regulatory single nucleotide polymorphisms (rSNPs). The second use case is about the association of DPs to genes. We check whether these genes go in accordance with prior biological knowledge. Finally, we briefly explain the statistical test we apply to quantify different methods solving the differential peak calling problem. Figure 4.1 gives an overview of all experiments.

4.1 Evaluation with Simulated Data

With the simulation of ChIP-seq data it is possible to produce a gold standard which can be used to evaluate DPCs. This allows us to evaluate methods for data with distinct characteristics such as the number of replicates and the number of reads per sample. For a given DPC, we check whether DPs in the simulated data are (in-)correctly called or not called. The simulation of single ChIP-seq datasets has already been addressed by Zhang et al. (2008) and Humburg (2011) (see Section 2.4.2), but none of these approaches can be used directly for the differential peak calling problem.

4.1.1. Simulation Method

Algorithm 4.1 describes how we simulate ChIP-seq reads that contain replicates. Figure 4.4 pictures the simulation procedure. In the following, we describe each step in more detail.

1. Step **Creating Protein Domains** We define n protein domains $(D_i)_{i=1\ldots n}$ for a genome g (see Figure 4.4, Step 1). Protein domains are regions in the genome that contain proteins. Depending on the number of proteins, these domains model histone modifications (large number of proteins) or TFs (low number of proteins) within the genome. Genomic regions with repetitive or unassembled parts are ignored for the domain placement. For each protein domain D_i, we sample the actual number q_i of proteins $(P_{i,j})_{j=1\ldots q_i}$ that are contained. The protein number q_i follows a

Negative Binomial distribution $q_i \sim NB_{m_1,p_1}$. We determine the positions $r_{i,1}$ of the first protein $P_{i,1}$ by uniformly selecting a position within the genome: $r_{i,1} \sim U[g]$. We then place further proteins $r_{i,j}$ with a particular space between each other, that is, $r_{i,j} = r_{i,1} + \sum_{k=1}^{j-1} b_k$ $(j \in \{2 \dots q_i\})$. The spacing variable b_k follows a mixture of normal distributions $b_k \sim \sum_l c_l \cdot N_{\mu_l,\sigma_l^2}$.

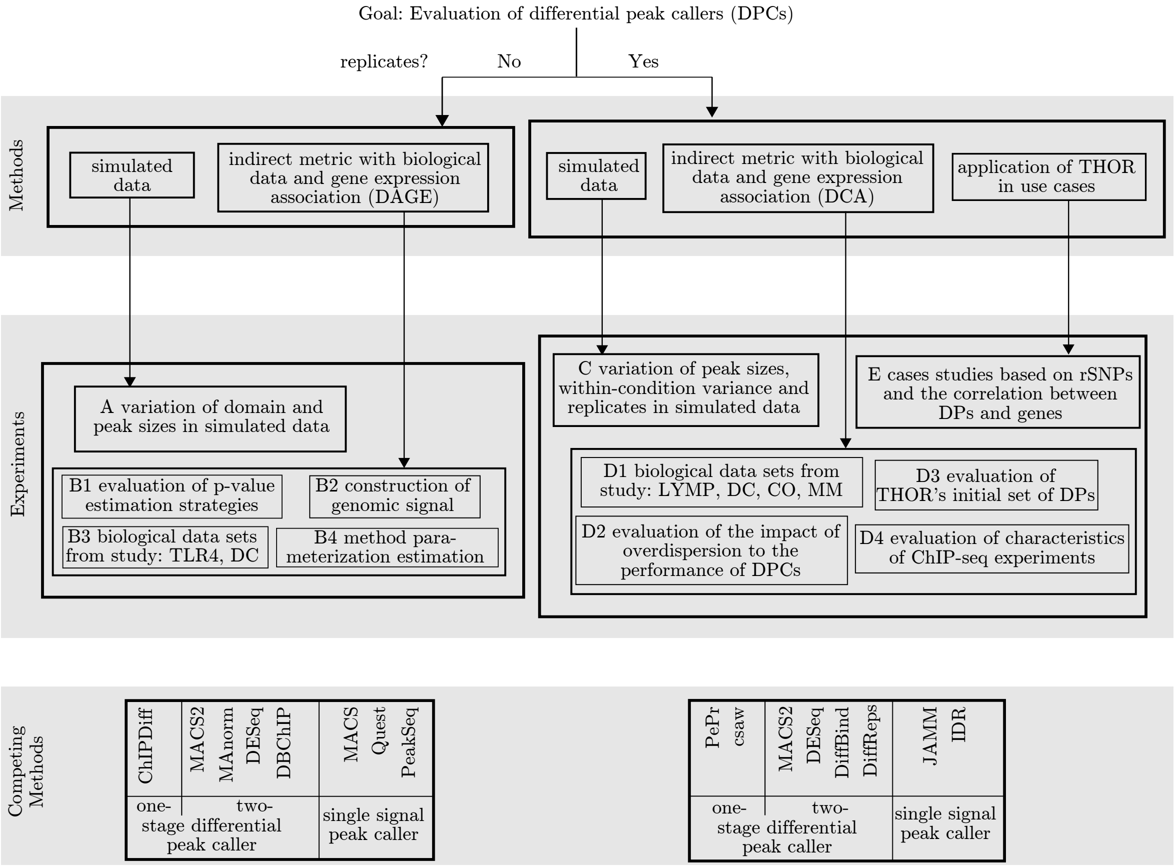

Goal: Evaluation of differential peak callers (DPCs)
replicates?
No
Yes
Methods
simulated data
indirect metric with biological data and gene expression association (DAGE)
simulated data
indirect metric with biological data and gene expression association (DCA)
application of THOR in use cases
Experiments
A variation of domain and peak sizes in simulated data
B1 evaluation of p-value estimation strategies
B2 construction of genomic signal
B3 biological data sets from study: TLR4, DC
B4 method parameterization estimation
C variation of peak sizes, within-condition variance and replicates in simulated data
E cases studies based on rSNPs and the correlation between DPs and genes
D1 biological data sets from study: LYMP, DC, CO, MM
D3 evaluation of THOR's initial set of DPs
D2 evaluation of the impact of overdispersion to the performance of DPCs
D4 evaluation of characteristics of ChIP-seq experiments
Competing Methods
ChIPDiff
MACS2
MAnorm
DESeq
DBChIP
MACS
Quest
PeakSeq
one-
two-
stage differential peak caller
single signal peak caller
PePr
csaw
MACS2
DESeq
DiffBind
DiffReps
JAMM
IDR
one-
two-
stage differential peak caller
single signal peak caller

Figure 4.1.: Overview of Experiments. If no replicates are available, we use the DAGE score for the studies TLR4 and DC to evaluate ODIN and its competitors ChIPDiff, MACS2, MAnorm, DESeq, and DBChIP (B3). We also evaluate the p-value estimation strategies (B1), the construction of the genomic signal ODIN uses for the analysis (B2) and the parametrization of the applied DPCs (B4). Moreover, we take advantages of simulation, where we vary the protein domain sizes as well as the peak sizes (A). For the case with replicates, we apply the DCA metric for the studies LYMP, DC, CO and MM to quantify the DP predictions of THOR and its competitors PePr, csaw, MACS2, DESeq, DiffBind and DiffReps (D1). Moreover, we evaluate the impact of overdispersion to THOR (D2), the characteristics of the data sets (D4) and the initial DPs THOR is using (D3). We also use simulated data for the evaluation with different peak sizes, within-condition variance and numbers of replicates (C). Furthermore, we apply THOR to two use cases to evaluate its ability to evaluate prior biological knowledge (E).

2. Step **Sampling Fragments** We sample the fragments $\{F_{i,j,l}\}$ that are bound to the protein $P_{i,j}$ (see Figure 4.4, Step 2). The length $s_{i,j,l}$ of each fragment $F_{i,j,l}$ follows a normal distribution $s_{i,j,l} \sim N_{\mu,\sigma^2}$. Fragments are assigned randomly to each DNA strand and always cover the entire length $o_{i,j}$ of the protein $P_{i,j}$ to which they are associated. However, since fragments are usually larger than the corresponding proteins, they are randomly moved up- or downstream. That is, for a given fragment's midpoint $m_{i,j,l}$,

$$m_{i,j,l} = r_{i,j} + t \quad with \quad t \sim U[-(s_{i,j,l} - o_{i,j}), (s_{i,j,l} - o_{i,j})].$$

For MA-plots of biological data, we typically observe a non-linear decrease of M-values for higher A-values. We model this non linearity by using function f which is described by a Laplace distribution:

$$f_{b,\mu}(d_{i,j}) = \frac{1}{2b}\exp\left(-\frac{|d_{i,j}-\mu|}{b}\right), \tag{4.1}$$

with $b = 0.5$, $\mu = 0.2$ and where $d_{i,j}$ gives the ratio of the fragments assigned to one of the biological conditions. The number l of fragments we sample for a protein is given by

$$l = f_{0.5,0.2}(d_{i,j}) \cdot p,$$

where p follows a Negative Binomial distribution $p \sim NB_{m_2,p_2}$. Figure 4.2 shows an example of an MA plot of simulated data. The factor $f_{0.5,0.2}$ causes the typical non-linear relationship between M and A values, that is, the M-values decrease stronger than linear with increasing A-values.

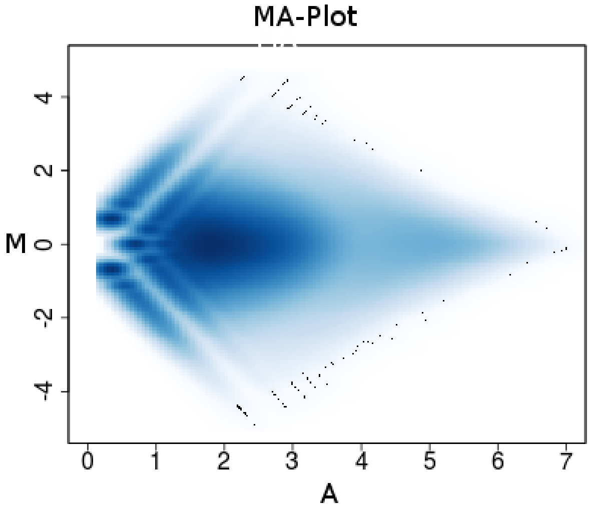


Figure 4.2.: MA-plot example of simulated ChIP-seq data. We use mean $m_1 = 8$ and variance $p_1 = 14$ for the negative binomial distribution describing the protein domains. The number of fragments assigned to each protein follows a Negative Binomial distribution with mean $m_2 = 150$ and variance $p_2 = 10000$. We have 2 replicates for each condition with $\alpha_0 = 5$ for a moderate variance between the replicates. The Laplace function (Equation 4.1) leads to the typical shape of the MA-plot.

3. Step **Assigning Fragments** For a given protein $P_{i,j}$, factor $d_{i,j}$ describes the ratio of the protein associated fragments that are assigned to the biological conditions. In our model, the ratio $d_{i,j}$ follows a beta distribution $B(0.5, 0.5)$. The beta distribution $B(0.5, 0.5)$ is symmetrical to 0.5 and tends to assume the extreme values 0 and 1. We thereby increase the probability that fragments are mostly assigned to one condition which could potentially result in a DP.

   For each protein domain $P_{i,j}$ and each biological condition, we randomly choose a replicate and assign fragments to it (see Figure 4.4, Step 3). For n replicates in a condition and for a constant vector $\overline{\alpha} = (\alpha_0, \ldots, \alpha_0)$ of length n, where α_0 describes the variance to distribute fragments among the replicates, the probability distribution

Algorithm 4.1 ChIP-seq read simulator

```
Input: reference genome $g$
Output: ChIP-seq read $\langle r_{kG_i} \rangle_{k \in \mathbb{N}, i \in 1,2}$ for condition $G_i$ with replicates
  1. select genomic regions in $g$, include protein domains $D_i$, and sample proteins $P_{i,j}$ in domain $D_i$
  2. sample and place fragments $F_{i,j,l}$ per protein $P_{i,j}$
  3. assign a proportion $d_{i,j}$ of fragments of a protein $P_{i,j}$ to a biological condition $G_i$, and assign each fragment to a ChIP-seq replicate of a biological condition $G_i$
  4. add noise to the data
  5. define DPs for each protein $P_{i,j}$ and output reads $\langle r_{kG_i} \rangle_{k \in \mathbb{N}}$ for condition $G_i$
```

to assign fragments to replicates is given by a Dirichlet distribution of order n, that is,

$$f(\overline{x}, \overline{\alpha}) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \quad \text{with}$$

$$B(\overline{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}.$$

For each fragment, we follow the sampled probabilities to assign it to a replicate. The lower α_0, the higher is the variance within the replicates.

4. Step **Adding Noise** We follow Zhang et al. (2008) to add noise to each replicate (see Figure 4.4, Step 4). We divide the genome into bins and assign a random weight to each bin. We assume that the majority of noise fragments in a ChIP-seq experiment appear in single locations, but some of them build dense clusters. We therefore use a right skewed gamma distribution to model the weight of a bin.

Accordingly to the weights, we randomly sample t bins with replacement. For each sampled bin, we add a noise fragment with a uniformly chosen position to the bin. The number t of chosen bins for replicate r is defined as

$$t = \min\left(\frac{\text{\#fragments}}{\text{FRiP}}, \frac{b \cdot \text{genome's length}}{\text{read's length}}\right).$$

FRiP is the fraction of reads in peaks. To have the number t invariant towards genome's length, we multiply the ratio of genome's and read's length by b. The variable b gives the average background coverage.

5. Step **Deriving Reads from Fragments and Defining Differential Peaks** Reads are obtained by getting the initial u base pairs of fragments in the forward strand (or the last u base pairs of the reverse strand). We define a DP gaining signal in condition G_i, if the number of fragment in condition G_i is higher than a given threshold e and at least v fragments are present, that is,

$$\frac{| \{F_{i,j,l}\}|_{G_i} |}{| \{F_{i,j,l}\} |} > e \quad \text{and} \quad \left|\{F_{i,j,l}\}\right|_{G_i} \geq v,$$

where $|\{F_{i,j,l}\}|_{G_i}$ gives the fragments of condition G_i (see Figure 4.4, Step 5). The position of the DP is defined by the protein position $r_{i,j}$.

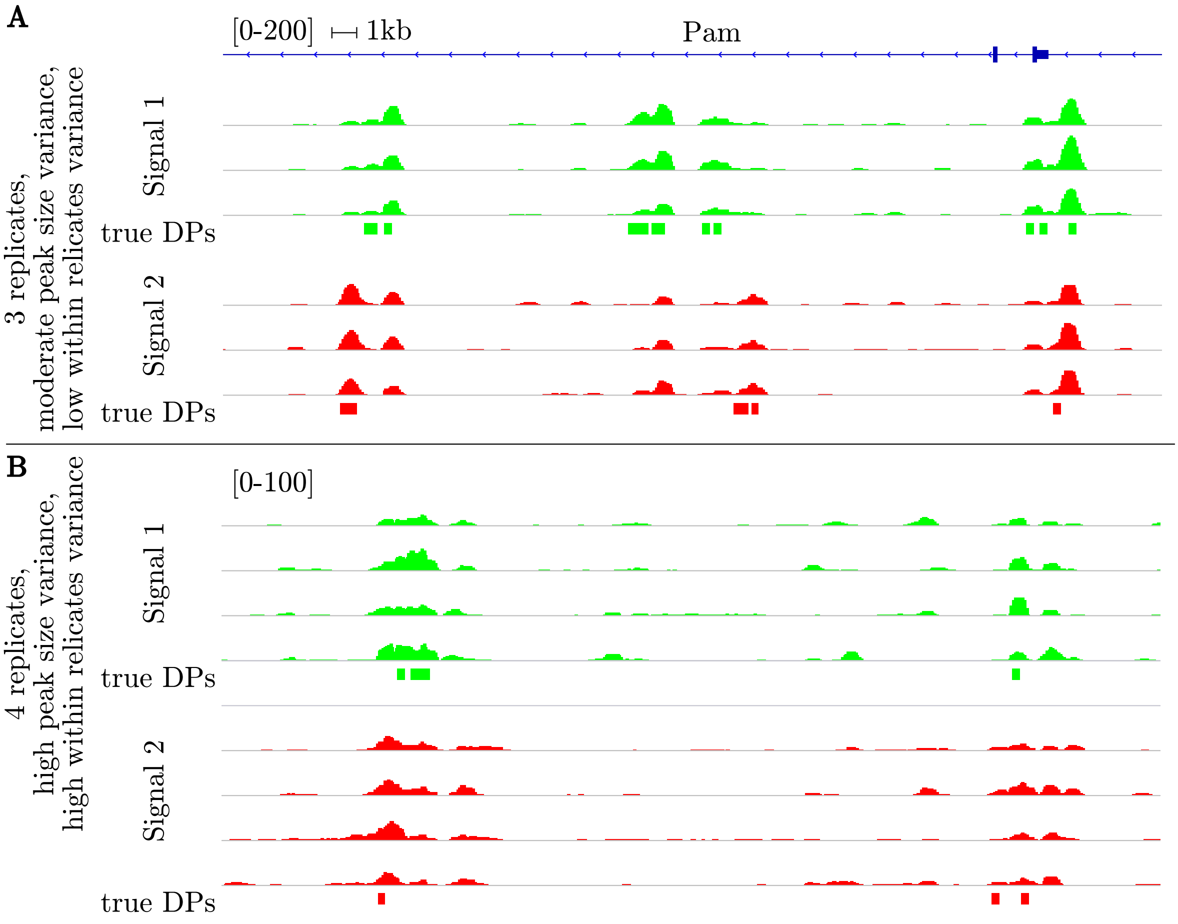


*Figure 4.3.: Example for simulated data. **(A)** A differential peak calling problem with 3 replicates in each condition, moderate peak size variance and low within-condition variance. **(B)** A hard differential peak calling problem with 4 replicates, high peak size variance and high within-condition variance.*

Figure 4.3 gives two examples for simulated data. We use the simulation algorithm with different parameters to obtain an easy and a hard differential peak calling problem with replicates.

To simulate ChIP-seq reads without replicates, we use a simpler version of Algorithm 4.1. First, we use a fixed spacing b between the proteins within a domain (see Step 1). Second, for the number l of fragments to sample per protein (see Step 2), we use a random variable following a Negative Binomial distribution. In the case with replicates, we use the Laplace function to model the non-linearity of the MA-plot (see Step 2). Here, we do not consider the non-linear property. Next, we use a constant ratio to assign fragments to one of the biological conditions (see Step 3). Finally, we do not add background noise to the ChIP-seq data, that is, we do not perform Step 4. The rationale of this simulation version is a historical one, as we first developed the simulation algorithm without replicates and then extended the approach to account for replicates.

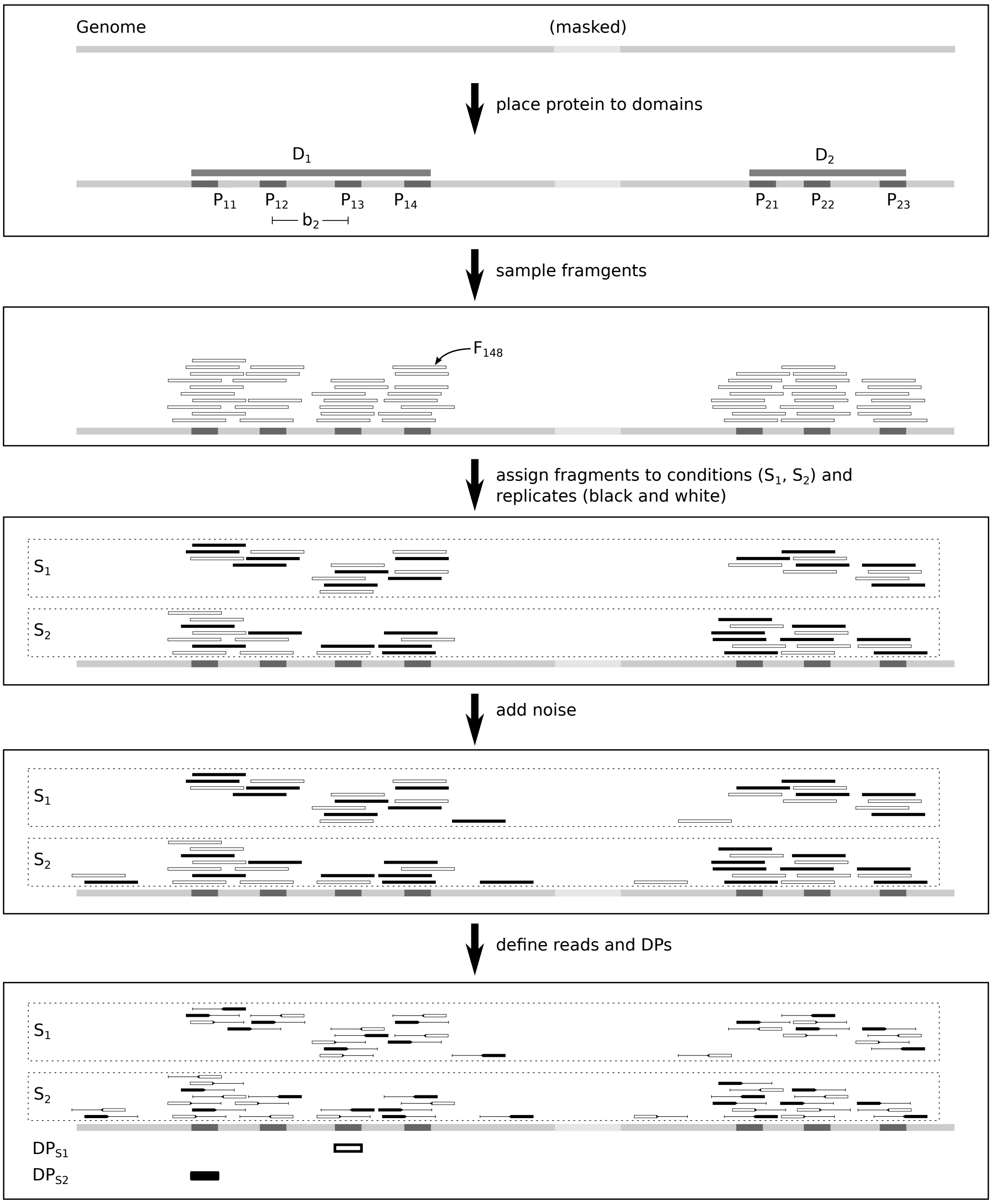


Figure 4.4.: *Workflow to simulate ChIP-seq data. First, unassembled and repeated regions are marked and ignored in the further progress. We then uniformly place domains of proteins in the genome. Here, domain D_1 contains proteins P_{11}, P_{12}, P_{13} and P_{14}, and Domain D_2 contains proteins P_{21}, P_{22} and P_{23}. The spacing between two proteins of a domain, for example b_2 between protein P_{12} and P_{13}, is sampled from a mixture normal distribution. Next, fragments are assigned to a protein, e.g. fragment F_{148} is associated with protein P_{14}. In the next step, fragments are assigned to both biological conditions (S_1, S_2) as well as replicates (black, white). We add noise to the data and define reads as the beginning or ending part of the fragments.*

4.1.2. Evaluation

We describe how we use simulated data to evaluate DPCs. For given simulated data and DP predictions, Table 4.1 gives an overview of the possible classification of DPs. For a gentle introduction to the evaluation of a classifier, please see Fawcett (2004).

	simulated DP	no simulated DP
called DP	true positives	false positives
not called DP	false negatives	true negatives

Table 4.1.: Possible classification. For a given genomic region, a DPC calls or does not call a DP (rows), while the simulated data actually contain or do not contain a DP (columns). This results in four possible classifications performed by the DPC, namely true/false positive/negative DPs.

If the simulated region is a DP and is classified as positive (negative), the region is called a true positive (false negative). If the simulated region is no DP and is classified as positive (negative), the region is called a false positive (true negative). For a DPC, the numbers in the major diagonal (true positives and true negatives) give the correct decisions. The numbers outside the major diagonal (false negatives and false positives) give the incorrect decisions. Moreover, we define the true positive rate (TPR) as the ratio between positive called DPs and the total number of positives, that is,

$$\text{TPR} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}.$$

The false positive rate (FPR) is given by

$$\text{FPR} = \frac{\text{false positives}}{\text{false positives} + \text{true negatives}}.$$

DPCs assign a p-value to the called DPs, where a lower p-value indicates a higher probability that the DP is a true positive. DPCs typically use a p-value threshold to define final DPs. Stricter thresholds lead to a stricter classification and in particular stricter FPR. The receiver operating characteristic (ROC) curve describes the relationship between FPR and TPR for distinct p-value thresholds. The ROC curve therefore allows a visual representation of a classifier performance under distinct FPRs. The area under the curve (AUC), that is, the integral, of a ROC curve gives a single score for a DPC. The higher the AUC ROC, the better the DP predictions of the considered method. We use ROC and AUC ROC to evaluate DPCs with simulated data.

In our case, true or false positives and negatives are given as genomic regions. To classify these regions, we define an interval based algebra. A genomic region $r = (r_s, r_e)$ is described by its starting position r_s and ending position r_e. We omit the chromosome information as we restrict our analysis to one chromosome. The intersection of two genomic regions $r_1 = (r_{1s}, r_{1e})$ and $r_2 = (r_{2s}, r_{2e})$ is defined as

$$r_1 \cap r_2 = \begin{cases} (\max(r_{1s}, r_{2s}), \min(r_{1e}, r_{2e})) & \text{if } r_1 \text{ and } r_2 \text{ overlap,} \\ \emptyset & \text{else.} \end{cases}$$

The subtraction of two genomic regions is defined as

$$
r_1 - r_2 = \begin{cases} (r_{1s}, r_{2s}) & \text{if } r_1 \text{ and } r_2 \text{ overlap, } r_{1s} < r_{2s},\ r_{1e} < r_{2e}, \\ (r_{2e}, r_{1e}) & \text{if } r_1 \text{ and } r_2 \text{ overlap, } r_{1s} > r_{2s},\ r_{1e} > r_{2e}, \\ \{(r_{1s}, r_{2s}), (r_{2e}, r_{1e})\} & \text{if } r_1 \text{ and } r_2 \text{ overlap, } r_{1s} < r_{2s},\ r_{1e} > r_{2e}, \\ \emptyset & \text{if } r_1 \text{ and } r_2 \text{ overlap, } r_{1s} > r_{2s},\ r_{1e} < r_{2e}, \\ (r_{1s}, r_{1e}) & \text{else.} \end{cases}
$$

For two sets of genomic regions the subtraction and intersection operation is performed element-wise, that is, for each element of the first set the operation is performed for each element of the second set. The size of a genomic region set is defined as the sum of all genomic regions' length.

With the interval based algebra on genomic regions, we are able to quantify the classification outcome of DP predictions. For a given simulation instance, we define g as a genomic region spanning the entire genome which we simulate. Moreover, we have a set of genomic regions T describing true DPs in the simulated data and a set P_A of genomic regions describing DPs predicted by algorithm A. We obtain true positive DPs by computing $T \cap P_A$, false positive DPs by computing $P_A - T$, false negative DPs by computing $T - P_A$ and true negative DPs by computing $g - T - P_A$. Figure 4.5 gives an example of the operations on genomic regions and the resulting classification. We use the algebra based classification and obtain

$$
\text{TPR} = \frac{|T \cap P_A|}{|T|} \quad \text{and} \quad \text{FPR} = \frac{|P_A - T|}{|\text{g} - T|}.
$$

In the case without replicates, we use for historical reasons a simpler approach to evaluate DP predictions. DPCs are evaluated by sorting the called DPs by smallest p-value and calculating the proportion of true positives among the top r called DPs.

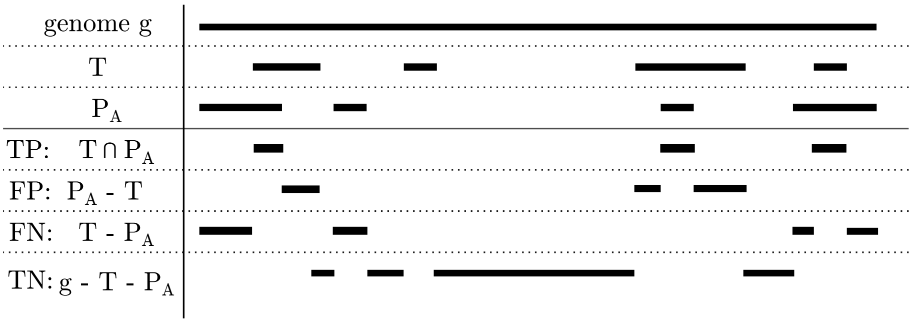


Figure 4.5.: Algebra on genomic regions. For a given genome g, true DPs T in the simulated data and predicted DPs P_A, we perform operations on genomic regions to compute all possible classification outcomes: true positive (TP), false positive (FP), false negative (FN) and true negative (TN) DPs. For the differential peak calling problem, we typically obtain a large number of true negative DPs, as the majority of the genome does not contain ChIP-seq signal.

4.1.3. Implementation

We implemented both strategies to simulate ChIP-seq data with and without replicates as Python command line tools. The tools are public available at `http://costalab.org/wp/ODIN` and `http://costalab.org/wp/THOR` under the terms of the *GNU General Public Licence v3 (GPL v3)*. The required input is the reference genome in fasta format and the genomic regions in BED format describing repeated regions within the genome. In the case with replicates, the user additionally has to define the number of replicates for each condition. The tools are highly parametrized, that is, all variables introduced here, such as the number of protein within a domain and the number of reads, can be customized for each run which allows a flexible tool usage. We use the same computational landscape of THOR and ODIN for the simulators. All websites mentioned in this section were accessed on 17th November, 2015.

4.2 Evaluation with Biological Data

We motivate and discuss two approaches to measure the quality of DPs called by DPCs on real datasets. These approaches are based on the correlation of DPs and associated gene expression data.

4.2.1. Indirect Metric

We associate changes in protein-DNA interactions with changes in gene expression whenever gene expression is measured in the same cellular conditions. The idea is based on the fact that the level of histone modifications correlates with the expression of the surrounding genes (Karlić et al., 2010). Several groups have already used histone modifications to predict gene expression (Cheng et al., 2011; Maze et al., 2014). We will use gene expression data to evaluate DPs.

Our measure is independent of the number of called peaks and can be applied either to gene expression data from sequencing or microarray data. For sequencing data, we first extend the DPs to have a length at least of 1000bps. Next, we count the reads of the gene expression data falling into the DPs. The use of minimum windows around DPs is based on the fact that we want to capture the expression of known genes or uncharacterized, long non-coding RNA (lncRNA) in the close proximity of the DPs. For microarray data, DPs are assigned to genes if (1) they are located in the gene or close to the promoter of a gene (1000bp upstream) or (2) if the DPs are located 50 Kbps away from the TSS without a TSS of another gene in between. The average expression value of genes assigned to a peak is used. Peaks not assigned to genes are ignored. We use the gene annotations from Cunningham et al. (2015).

DAGE

For the case without replicates, we sort DPs called by algorithm A by increasing p-value and take the top k ranked DPs that are associated with the HMM state `Gain 1` (or `Gain 2`). Let s_{Ai1} and s_{Ai2} be the gene expression values associated with the ith DP called by algorithm A in the first and second condition. We define the function

$$e(k) = \frac{\sum_{i \leq k} \log\left(\frac{s_{Ai1}}{s_{Ai2}}\right)}{k}, \tag{4.2}$$

for $k \in \mathbb{N}$. Function e gives the average logarithmic ratio of condition-specific expression values for the top k DPs. The higher the value is, the higher is the association between DP and changes in expression.

We compute the integral of function e and obtain the single statistic DAGE (D̲ifferential

Average Gene Expression). We use Equation 4.2 and define

$$\mathrm{DAGE} = \left(\sum_{k \in K} |e(k)| \right) \cdot h, \tag{4.3}$$

for $K = [h, 2 \cdot h, \ldots, H]$ where h is the step size and H the maximum number of DPs used.

DCA

If replicates are available, we perform a differential expression analysis with DESeq (Anders and Huber, 2010) for RNA-seq and limma (Ritchie et al., 2015) for microarray data. We compute p-values of the differential expression analysis and use them to indicate expression changes. This approach improves the DAGE statistic, which is based on a simple fold change.

We rank the DPs by increasing p-values. For a given DP with rank i, let p_{Ai1} be the corresponding p-value of the DPC A and let p_{Ai2} be the corresponding p-value of the gene expression analysis. We compute the Spearmann rank-correlation (Spearman, 1904) between p-value lists (p_{Ai1}) and (p_{Ai2}) for the top k ranked DPs, that is,

$$f(k) = \mathrm{cor}\big((p_{Ai1}), (p_{Ai2})\big)_{\mathrm{Spearman}}, \tag{4.4}$$

for all $i < k,\ k \in \mathbb{N}$. We obtain DCA (Differential Correlation Analysis) curves which indicate the association of gene expression and DPs for a distinct number of called peaks. Furthermore, we obtain a single score for algorithm A by estimating the normalized area under the DCA curve, that is, we use Equation 4.4 and obtain

$$\mathrm{DCA} = \frac{\left(\sum_{k \in K} \max\big(0, f(k)\big) \right) \cdot h}{H}, \tag{4.5}$$

for $K = [h, 2 \cdot h, \ldots, H]$, where h is the step size and H the maximum number of DPs used. We ignore DPs with a negative correlation between gene expression and count data, as they represent spurious solutions. The DCA score detects the positive correlation between gene expression changes and DPs.

DAGE and DCA evaluation are based on the use of cumulative values computed by the functions e and f. Evaluation of $e(k)$ takes into account s_{Ai1} and s_{Ai2} with $i \leq k$. Evaluation of $e(k+1)$ also considers s_{Ai1} and s_{Ai2} with $i \leq k$. Additionally, the succeeding values $s_{A(k+1)1}$ and $s_{A(k+1)2}$ are evaluated. This pattern makes the values s_{Ai1} and s_{Ai2} gain higher impact on distinct evaluations of function e than all succeeding values s_{Aj1} and s_{Aj2} with $j > i$. The corresponding statement is valid for function f. This characteristic of DAGE and DCA is also shared by AUC ROC (see Section 4.1.2), which is typically used by the machine learning community (Fawcett, 2004).

4.3 Biological Datasets

We list the biological datasets that we use to evaluate DPCs. Depending on the availability of replicates, we use the DAGE or DCA statistic for the evaluation. Table 4.2 gives an overview of the differential peak calling experiments without replicates and Table 4.3 of the datasets with replicates. We use BWA (Li and Durbin, 2010) version $0.6.1 - \mathrm{r}104$ with default parameters for read mapping to the mouse (mm9) or human genome (hg19).

- **Dendritic Cell (DC) Differentiation** This in-house study measures regulatory changes during the development of antigen-presenting dendritic cells (DC) which develop from hematopoietic stem cells in bone marrow. Our collaborators have established an in-vitro protocol to differentiate multipotent progenitors (MPP) from adult mouse bone marrow to common DC progenitors (CDP) (Felker et al., 2010). CDP cells are further differentiated to either classical DC (cDC) or plasmacytoid DC (pDC). For these four cell types, we have performed a DP analysis comparing the lineage commitment steps (MPP to CDP, CDP to cDC, CDP to pDC) and DC subset specification (cDC and pDC).

  ChIP-seq experiments without replicates were performed for the histone modification H3K4me1 and the TF PU.1 by Lin et al. (2015). The data is available at the Gene Expression omnibus (GEO) database (Edgar et al., 2002) with accession number GSE57563. It has been shown that the TF PU.1 is associated with active gene regulation (Lin et al., 2015). Hence, similar to histone modifications, we use PU.1 to evaluate DPs with the DAGE score. A single input-DNA profile which serves as control for all cell types is available. We also have gene expression data from microarrays for all four cell types from Felker et al. (2010) (GEO accession GSE22432). Altogether, we obtain 8 experiments which are listed in Table 4.2.

  Furthermore, ChIP-seq experiments with two technical replicates were performed for the histone modification H3K27ac (GEO accession GSE73143). Input-DNA for each cell type is available. This study represents a scenario with potentially very low variability within the biological conditions and leads to 4 experiments in Table 4.3.

- **TLF4 Pathway Analysis (TLR4)** Kaikkonen et al. (2013) investigate the response of macrophages after activation of the TLR4 signaling pathway in mice. They provide ChIP-seq experiments without replicates for the TF PU.1 at time points 0h, 1h, 6h, 12h and 24h and for the histone modification H3K4me2 at time points 0h, 1h, 6h and 24h (time point 12h was not available). We perform differential peak calling by comparing the time point 0h with all other time points, which leads to 7 experiments (Table 4.2). The study provides an input-DNA signal of untreated cells, which is used as control. Moreover, we use the genomic run-on sequencing (GRO-seq) experiments, which measure the quantity of nascent transcripts, at time points 0h, 1h, 6h, 12h and 24h for evaluation. These data were obtained from GEO accession number GSE48759.

- **Epigenomics Effects of Cocaine (CO)** Feng et al. (2014) analyze epigenetic changes after cocaine intake on mouse nucleus accumbens. The study measures histone modifications of three biological replicates after treatment with a cocaine or saline solution. We use data from histone modifications H3K4me1 and H3K36me3, which leads to two DP calling experiments. The authors provide RNA-seq data matching the samples, but no input-DNA (GEO accession number GSE42811 and GSE24850). This study represents a scenario with biological replicates that exhibit a similar genomic background. Therefore, we expect a low variance within the biological conditions.

- **Monocyte and Macrophages (MM)** This study provides samples of monocytes activated to macrophages in up to 8 human samples (Saeed et al., 2014). We consider the histone modifications H3K4me1, H3K27ac and H3Kme3. For histone modification H3K4me1 there are 6 monocytes and 10 macrophages, for H3K27ac there are 5 monocytes and 8 macrophages and for H3K4me3 there are 6 monocytes and 10 macrophages samples. We perform DP estimations between monocytes and macrophages for all histone modifications. Condition-specific RNA-seq data (36 macrophages and 25 mono-

Experiment	Protein	Cond. 1	Cond. 2
TLR4-PU.1-0h-1h	PU.1	0h	1h
TLR4-PU.1-0h-6h	PU.1	0h	6h
TLR4-PU.1-0h-12h	PU.1	0h	12h
TLR4-PU.1-0h-24h	PU.1	0h	24h
TLR4-H3K4me2-0h-1h	H3K4me2	0h	1h
TLR4-H3K4me2-0h-6h	H3K4me2	0h	6h
TLR4-H3K4me2-0h-24h	H3K4me2	0h	24h
DC-PU.1-MPP-CDP	PU.1	MPP	CDP
DC-PU.1-CDP-cDC	PU.1	CDP	cDC
DC-PU.1-CDP-pDC	PU.1	CDP	pDC
DC-PU.1-cDC-pDC	PU.1	cDC	pDC
DC-H3K4me1-MPP-CDP	H3K4me1	MPP	CDP
DC-H3K4me1-CDP-cDC	H3K4me1	CDP	cDC
DC-H3K4me1-CDP-pDC	H3K4me1	CDP	pDC
DC-H3K4me1-cDC-pDC	H3K4me1	cDC	pDC

Table 4.2.: Overview of DP experiments without replicates. We give the experiment name, protein type as well as the cellular conditions for each of the evaluated differential peak problems.

cytes samples) are used for evaluation. The study does not provide input-DNA data for the ChIP-seq experiments. The data are available with restricted access at the European Genome-phenome Archive (EGA) (Lappalainen et al., 2015), accession number EGAD00001001011. This study represents a scenario with human biological replicates with a moderate within-group variability.

- **B cell lymphoma (LYMP)** Koues et al. (2015) performed a comprehensive analysis of regulatory genomic features in lymphomas. We use ChIP-seq data of the histone modification H3K27ac on follicular lymphoma cells (FLs), as well as distinct populations of B cells from healthy donors: proliferative centroblasts (CC) and peripheral blood B cells (PBBA). We only consider samples with a matching input-DNA and gene expression (measured with microarrays): CC samples 1-5, FL samples 1, 2, 5, 8, 10, 11, 14, 16 and PBBA sample 1-3 (GEO accession number GSE62246). We evaluate DPs in the cases FL vs. CC, FL vs. PBBA and CC vs. PBBA. This dataset contains human biological replicates and disease samples and is expected to have a high within-group variability.

4.4 Experiments without Replicates

Here, we explain how the above mentioned evaluation procedures are used for our experiments without replicates. We describe the experiments we perform to evaluate ODIN and other methods that do not account for replicates. ODIN is run with a Binomial and a mixture of Poisson distribution as emission. Moreover, we evaluate ChIPDiff, MACS2, MAnorm, DBChIP and DESeq. MAnorm, DBChIP and DESeq are based on initial candidate peaks called by SPCs (see Table 2.1). We also evaluate the use of the SPCs PeakSeq, Quest and MACS. First, we describe how we use the simulated data. Second, we resort to the biological data for the evaluation experiments.

Experiment	Histone	Cond. 1	Cond. 2	#rep
DC-H3K27ac-MPP-CDP	H3K27ac	MPP	CDP	2, 2
DC-H3K27ac-CDP-cDC	H3K27ac	CDP	cDC	2, 2
DC-H3K27ac-CDP-pDC	H3K27ac	CDP	pDC	2, 2
DC-H3K27ac-cDC-pDC	H3K27ac	cDC	pDC	2, 2
CO-H3K36me3	H3K36me3	saline	cocaine	3, 3
CO-H3K4me1	H3K4me1	saline	cocaine	3, 3
MM-H3K27ac	H3K27ac	monoc.	macrop.s	5, 8
MM-H3K4me1	H3K4me1	monoc.	macrop.	5, 8
MM-H3K4me3	H3K4me3	monoc.	macrop.	6, 10
LYMP-FL-CC	H3K27ac	FL	CC	8, 5
LYMP-FL-PBBA	H3K27ac	FL	PBBA	8, 3
LYMP-CC-PBBA	H3K27ac	CC	PBBA	5, 3

Table 4.3.: Overview of DP experiments with replicates. For each experiment, we describe the experiment name, histone modification type, cellular conditions and number of replicates.

4.4.1. Evaluation of Methods with Simulation Data

We describe the simulation experiments without replicates. First, we explain the experimental setup and second, we give details about the parametrization of the simulator and all used DPCs.

Experimental Setup

We investigate the effect of protein domain sizes as well as the number of reads in the libraries. We therefore vary the parameters

- m_1 to obtain larger protein domains,
- p_1 to obtain more variable sized protein domains,
- m_2 to obtain peaks with higher number of reads, and
- p_2 to obtain peaks with higher variance in their size.

See Section 4.1.1 for a detailed description of the parameters. We evaluate the parameter settings

$$(m_1,\ p_1) \in \{(1,4),(4,6),(8,14)\} \quad \text{and}$$

$$(m_2,\ p_2) \in \{(20,200),(20,2000),(100,200)\}.$$

We combine the two-stage DPC DBChIP and MAnorm with the SPC MACS, as MACS provides good performance (Chen et al., 2012; Wilbanks and Facciotti, 2010) and does not require input-DNA for the execution. Moreover, we evaluate the usage of DESeq in this scenario. Hence, we combine DESeq with MACS and refer to this algorithm as DESeq-MACS. We run ODIN with the Binomial and mixture of Poisson emission distribution. For the mixture of Poisson we consider 1, 2, 3, and 4 components.

Parametrization of Methods

The following values are set as constants in our simulated ChIP-seq experiments. We generate 10,000 protein domains per dataset. The spacing b between proteins is defined as 200 bp, which reflects the average spacing between nucleosomes (Mammana et al., 2013). Furthermore, ChIP fragments typically have a length of 200 bp (Furey, 2012). We therefore model the fragment's size with mean $\mu = 200$ and standard deviation $\sigma = 20$. The standard deviation follows estimates taken from paired-end sequencing data reported in (Marschall et al., 2012). The minimum number of reads v to support a DP is 25 and the ratio e for definition of a DP is defined as $e = 0.6$. We use a ChIP-seq read size of 26 bp. We choose chromosome 1 of the mouse genome (mm9) as reference genome and align the simulated reads with BWA (Li and Durbin, 2010) version $0.6.1 - \mathrm{r}104$ with default parameters. For each parametrization choice, we generate 50 simulated datasets. We run all methods with default parameters.

4.4.2. Evaluation of Methods with Biological Data

We give details about the experiments performed with biological datasets without replicates. First, we introduce the experimental setup and second, we describe the parametrization of all considered DPCs.

Experimental Setup

We use all datasets which are listed in Table 4.2, that is, all 15 ChIP-seq experiments from the TLR4 study (Kaikkonen et al., 2013) and the DC study (Lin et al., 2015) without replicates. The two-stage DPCs DBChIP, DESeq and MAnorm require peaks of each ChIP-seq signal as input. In contrast to the simulated data which does not provide input-DNA, we here are able to separately evaluate the SPCs PeakSeq, Quest and MACS to compute the candidate peaks. The SPCs were selected based on their good performance (Chen et al., 2012; Wilbanks and Facciotti, 2010). We also define a two-stage DPC which merges all candidate peaks and uses them as input for DESeq.

DBChIP uses predefined short windows of 250bps around the peak summits as candidates for DPs. As proposed by the authors, we apply DBChIP to TF-based ChIP-seq data which typically exhibit well defined, sharp peaks. In contrast to the DPC that combines DESeq with candidate peaks, DBChIP finds DPs with variable size which is common for histone ChIP-seq data. Hence, we distinguish between experiments with TFs and histone modifications. As ChIPDiff does not provide *p*-values or any criteria to sort DPs, we can only obtain points for the DAGE curve.

In our experiments, ODIN (with a single component in the mixture model) requires averagely 12GB of memory. The calculations last on average 4 hours on a 3.4GHz machine. Computational time increases linearly with the number of components in the case of mixture of Poisson distributions.

Parametrization of Methods

For ODIN, we use the mappability files that are provided by Landt et al. (2012)[1] to compute the genomic signal (Section 3.2.1). We only consider regions with a mappability value of 1. We compute the fragment size (Section 3.2.2) with $\hat{f} = \arg\max_{f \in G} c(f)$ for the range $G = [0, 5, \ldots, 600]$. Moreover, we use a step size s of 50 and window size w of 100 to compute the signal profile (Section 3.2.3). This choice was based on visual inspection of peaks: smaller windows did not affect peaks and larger windows induced too large peaks. We only use input-DNA signal of chromosome 1 to build the GC-content histogram (Section 3.2.4).

[1] `http://hgdownload-test.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeMapability/`, last access: 25th November 2015,

ChIPDiff is also run with default parameters ($FC = 3$, minRegionDist = 1000 and minP = 0.95). It finds less than 20 peaks for half of the experiments from the TLR4 study. For these experiments, we change parameters ($FC = 1.5$, minRegionDist = 200 and minP = 0.7) to obtain at least 100 DPs. MACS2 is run with parameter $C = 0.5$ for the TLR4 study and $C = 1.5$ for the DC study. MAnorm, DBChIP and DESeq and all SPCs are run with default parameters.

4.4.3. Evaluation of *P*-value Estimation Strategies

We evaluate distinct estimations strategies to compute *p*-values of DPs. For this, we call DPs with ODIN for all biological datasets (see Table 4.2) and re-compute the *p*-values with DESeq and edgeR. As we do not consider replicates, we have to choose the following parameters for DEseq's function *estimateDispersions*: the method *blind*, the sharingMode *fit-only* and the fitType *local*. For edgeR, we follow the user guide and use a dispersion factor of 0.04 in the *exactTest* function. We use edgeR version 3.6.8 and DEseq version 1.16.0. We compare the DAGE scores based on the *p*-value estimates of ODIN, edgeR and DESeq.

4.5 Experiments with Replicates

We describe the experiments to evaluate THOR and competing DPCs that take replicates into account. First, we resort to simulated data and then we describe how we use biological data for the evaluation experiments.

We call DPs with THOR and all methods described in Table 2.1 that account for replicates, that is, MACS2, DiffBind and DiffReps, PePr and csaw. Moreover, we combine DESeq with the SPC JAMM, which can handle replicates. We also combine DESeq with IDR, which uses peaks called by MACS2 on single ChIP-seq profiles to estimate common peaks within a condition (see Section 2.4.3). We refer to these approaches as DESeq-JAMM and DESeq-IDR respectively.

As described in Section 3.3.3, ODIN uses a Binomial or, for large n, where n is the number of reads in the ChIP-seq libraries, an equivalent Poisson distribution. We evaluate THOR with a Poisson distribution by fixing $a_{sG_k} = 0$ (see Section 3.3.3), which can be seen as a version of ODIN that supports replicates. We refer to this approach as Poisson-THOR.

4.5.1. Evaluation of Methods with Simulation Data

We describe the simulation experiments with replicates. First, we explain the experimental setup and second, we give details about the parametrization of the simulator as well as all DPCs.

Experimental Setup

We are interested in how methods perform when the number of reads of each protein in a domain, the number of replicates and the variance within replicates changes. We therefore simulate the following parameter settings:

- $(m_2, p_2) \in \{(100, 200), (100, 400)\}$ to obtain peaks with moderate and high variance in their sizes. Thereby, we model distinct types of histone modifications which have either uniform or varying peak sizes;
- $(r_1, r_2) \in \{(2, 2), (4, 4)\}$ to evaluate experiments with 2 and 4 replicates in each condition, and

- $\alpha_0 \in \{5, 10, 60\}$ to obtain data with low (60), moderate (15) and high (5) variance within a biological condition. This parameter controls the consistency between replicates: higher variance imposes lower consistency and more difficult differential peak calling problems.

See Section 4.1.1 for a detailed description of the parameters. Experiments with 2 replicates are obtained by discarding 2 ChIP-seq experiments from each biological condition of the experiments with 4 replicates. We were not able to run csaw on the simulated data, even when trying out distinct parameters as used in the real data. Furthermore, PePr requires input-DNA which is not provided by our simulation model.

Parametrization of Methods

We model the space between the proteins b_k within the same domain. Since we are interested in modelling histones, we estimate mixture model parameters by using histone position data in yeast (Weiner et al., 2010). For this, we randomly take $10,000$ consecutive histone positions and fit a mixture normal distribution to their distance. We ignore positions which are 500bp away from each other, as we assume that these positions belong to two different histone domains. Bayesian information criterion shows that 2 components fit best for the mixture model ($-1.5 \cdot 10^2$). To model histone characteristics, we define the minimum distance between proteins in a domain as the sum of the usual estimate of histone size (147bps) and the average linker size (55bps) (Szerlong and Hansen, 2010).

Similar to the case without replicates, we generate $n = 10,000$ protein domains per dataset. ChIP fragments typically have a length of 200 bp (Furey, 2012). We therefore model the fragment's size with mean $\mu = 200$ and standard deviation $\sigma = 20$. The standard deviation follows estimates taken from paired-end sequencing data reported by Marschall et al. (2012). The minimum number of reads to support a DP v is 25 and the ratio e for definition of a DP is defined as $e = 0.6$. We use a read size u of 26. Reads are sampled from chromosome 1 of mm9 and aligned with BWA with default parameters. We use a FRiP of 0.05 (Landt et al., 2012) for the estimation of the noise in the simulated signal. Our empirical studies have shown that the average background coverage b should be around 0.25 in ChIP-seq experiments. We use $m_1 = 8$ and $p_1 = 14$ for the Negative Binomial distribution NB_{m_1,p_1} describing the number of proteins in a protein domain. We repeat each experiment 25 times. We run all methods DPCs with default parameters.

4.5.2. Evaluation of Methods with Biological Data

We describe the experiments performed with biological datasets that contain replicates. First, we explain the experimental setup and second, we detail the parametrization of all considered DPCs.

Experimental Setup

We use all 12 dataset which are listed in Table 4.3, that is, 4 experiments from the DC (Lin et al., 2015), 2 experiments from the CO (Feng et al., 2014), 3 experiments from the MM (Saeed et al., 2014) and 3 experiments from the LYMP study (Koues et al., 2015).

We run THOR with two normalization strategies, the housekeeping gene approach as well as TMM (see Section 3.2.6), and refer to them as THOR-HK and THOR-TMM. The evaluation of the normalization strategies was not necessary in the case of the simulated data, as we ensure that both conditions have the same overall number of reads.

On the dataset with largest number of ChIP-seq samples (MM-H3K4me3), THOR required 4 hours and 16 GBs of memory on a 3.4GHz machine.

Parametrization of Methods

As described in Section 3.3.5, we use certain criteria to define initial DPs used to train the HMM of THOR. We use $t_1 = \langle x \rangle^{.95}$ as minimum difference between signals, where $\langle x \rangle^{.95}$ is the value in the 95% percentile of $\mathbf{X}$; $t_2 = 1.6$ as fold change criteria, and $t_3 = t_1/2$. If these parameters yield a training set smaller than $t^{\min} = 100$, we decrease t_2 by 15 and t_1 by 0.1, and repeat the training set construction procedure. To estimate the mean-variance function for each biological condition k, we randomly choose 20.000 bins, estimate mean and variance for each bin and fit the quadratic model described in Equation 3.17 using a non-linear least squares approach (Levenberg, 1944). We use regions 500bps upstream of housekeeping genes (C1orf43, CHMP2A, EMC7, GPI, PSMB2, PSMB4, RAB7A, REEP5, SNRPD3, VCP, VPS29) described by Eisenberg and Levanon Eisenberg and Levanon (2013) as control regions for the human genome.

We use the following parametrization for the competing methods. For csaw, as suggested by the authors, we use a window size of 150bp and a step size of 25bp. All other parameters are set as default. For Pepr, we follow the instructions on their webpage (see `https://ones.ccmb.med.umich.edu/wiki/PePr/`, last access on 12th November 2014) including a procedure to remove artefacts in ChIP-seq data. To obtain a number of DPs comparable to other tools, we increase the p-value threshold parameter to 0.01. Initially, MACS2 called too few DPs, such that we had to decrease both the minimum length for DPs by using $l = 50$ and the fold-change cutoff by using $C = 1.5$ in the algorithm *bdgdiff*. Moreover, we increased the p-value threshold to 0.2 to increase the number of peaks for the algorithm *callpeak*. MACS2's *callpeak* algorithm serves as internal SPC by identifying peaks in single ChIP-seq profiles. These peak estimates represent the base for all downstream steps to call DPs. For DiffBind, as recommended by the authors, we choose parameter *minOverlap* to be 3 in the count function to only consider peaks supported in up to three replicates across all conditions. Moreover, we increase the threshold for significant DPs (*th=0.1*). We run DiffReps with default parameters, that is, we use a window size of 1000bp and a step size of 100bp, but increase significance threshold for called DP (by using the option *–pval 0.1*). For DESeq-IDR, we follow the framework of ENCODE (see `https://sites.google.com/site/anshulkundaje/projects/idr`, last access on 21th November 2014) to estimate common peaks with IDR. We use an IDR threshold of 0.01 for the replicates, an IDR threshold of 0.02 for the self-consistency replicates, and an IDR threshold of 0.0025 for the pooled pseudo replicates. We then apply DESeq with default parameters to check for DPs. Moreover, we use JAMM in combination with DESeq where we use default parameters for both methods.

4.6 Use Cases of THOR

We use THOR in two studies that investigate biologically motivated questions. We resort to the data described in Section 4.3. The aim is to present biological results that are expected to arise, as we thereby ensure that THOR is performing proper DP predictions. The prior knowledge about the results stems from the study-specific biological background.

4.6.1. Identifying rSNPs

We evaluate the ability of THOR with the housekeeping normalization approach to detect DPs that support regulatory single nucleotide polymorphisms (rSNPs). Regulatory single nucleotide polymorphisms are point mutations of the DNA which may modify the specific transcription factor binding site (TFBS) such that the TF binding affinity is influenced. These rSNPs thereby influence the gene expression as well as the chromatin state (Hawkins et al., 2010; Guo et al., 2014). Figure 4.6 gives an example for a rSNP.

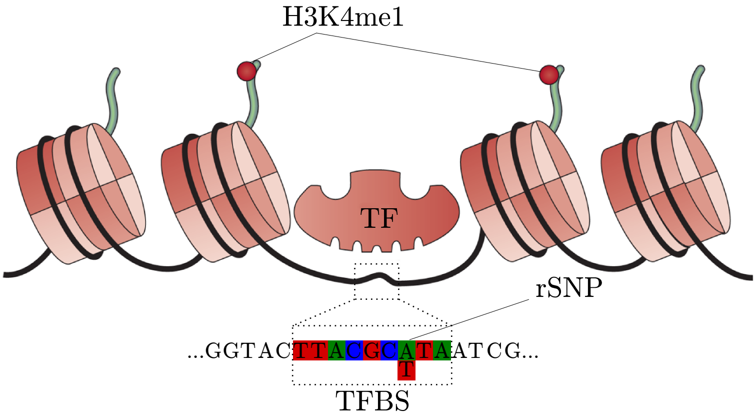


Figure 4.6.: A regulatory SNP, here a substitution of a single nucleotide A to T, effects the TBFS such that the TF cannot bind at that genomic position. Hence, the gene expression is influenced and the chromatin structure is modified. The figure is based on Hawkins et al. (2010).

We evaluate the presence of disease-associated rSNPs in DPs by considering samples of tumour B-cell from patients follicular lymphoma (FL) and centroblasts B-cell from healthy donors (CC) (dataset LYMP-FL-CC from the study of Koues et al. (2015), see Table 4.3).

We call SNPs in FL samples using GATK's *UnifiedGenotyper* (McKenna et al., 2010). Concerning filtering steps performed by GATK, we use a threshold value of 20 for read depth across samples (DP), the Variant Confidence/Quality by Depth (QD) and the RMS Mapping Quality (MQ). We filter SNPs that lie on chromosomes chrY and chrM. Moreover, we exclude SNPs falling into blacklisted regions (ENCODE Project Consortium, 2012) and restrict our analysis to loci with at least 4 reads in more than 2 CC and 3 FLs samples. This yields 4390 candidate SNPs. We further filter SNPs, if the frequency of alternative alleles is higher in FL than in CC samples (p-value < 0.05; Fisher Exact Test). This procedure results in 243 candidate SNPs, of which 143 overlapped with DPs called by THOR (FL vs. CC). Moreover, we evaluate all transcription factor binding sites with JASPAR (Mathelier et al., 2014) and UNIPROBE (Robasky and Bulyk, 2011) motifs and a FDR of 10^{-4} using the motif analysis tool from `www.regulatory-genomics.org` (last access: 3rd December 2015). Altogether, 117 rSNPs were associated with DPs gained in FL (vs CC) and 20 rSNPs were associated with peaks gained in CC (vs. FL). See Sup. Table A.40 for a selection of the rSNPs. The overlap with DPs called by MACS2 only covers 41 candidate rSNPs, which is the highest overlap among all competing methods of THOR.

4.6.2. Analysing the Development of Dendritic Cells

We evaluate the confirmability of DPs called by THOR by analysing the datasets DC-H3K27-ac-CDP-cDC and DC-H3K27ac-CDP-pDC (see Table 4.3) from the study of Lin et al. (2015) in more detail. We apply THOR with the housekeeping gene normalization approach and restrict our analysis to the case where peaks are gained respectively in the cDC and pDC condition. Similar to the DCA and DAGE approach (see Section 4.2.1), we assign DPs to genes if (1) they are located in the gene or close to the promoter of a gene (1000bp upstream) or (2) if the peaks are located 50 Kbps away from the TSS without a TSS of another gene in between. We sort the gene list by the associated *p*-values of the DPs. If multiple DPs are assigned to a gene, we take the DP with the smallest *p*-value. The rationale for this experiment is that under the assumption that THOR calls reasonable DPs, genes that are known to be associated with the differentiation of CDP to cDC or pDC cells should be ranked in the top of the list, that is, close to DPs with lowest p-values.

4.7 Statistical Analysis

We apply the Friedman-Nemenyi test (Demšar, 2006) to the DAGE and DCA values as well as to the measures provided by the ChIP-seq simulation to evaluate DPCs. The Friedman-Nemenyi test consists of two parts. First, the non-parametric Friedman test (Friedman, 1937) detects differences in observations estimated by various methods across multiple datasets. The test computes for each method a rank for the observations of all datasets. Under the assumption of uniformly distributed ranks across the methods, that is, that the methods give similar observations, the test checks for significant differences within the ranks. Depending on the number of observations and methods, the test statistic approaches a χ^2 distribution. Second, the Nemenyi test (Nemenyi, 1962) is applied to identify the method that causes the significant difference in the rank statistic. In our case, the Friedman-Nemenyi test indicates whether one of the DPCs is assigned to significant higher values than others across multiple datasets.

4.8 Summary

In this chapter, we described the experimental methods for our evaluation studies of DPCs. First, we introduced a simulation algorithm for ChIP-seq profiles with DPs to obtain artificial gold standards. Second, we described two indirect metrics to rate DPs. Both metrics are based on the idea that histone modifications correlate with gene expression. If replicates are available, we apply the DAGE metric for 15; and otherwise, the DCA metric for 12 differential peak calling problems. We also detailed the experimental setup of our studies. In the case with replicates, we simulate ChIP-seq data with variable protein domains sizes as well as variable peak sizes for our study. With regard to the DAGE metric, we investigate distinct parameter settings for ODIN. Moreover, we perform a DAGE-based evaluation study with ODIN and its competitors. In the case without replicates, we simulate ChIP-seq data with variable peaks sizes, different number of replicates and distinct variances between the conditions. We also perform an evaluation study with regard to the DCA metric based on THOR and its competitors. We furthermore evaluate THOR's ability to support prior biological findings. In particular, we use THOR to call DPs between distinct differentiation steps of dendritic cells. We assign called DPs to genes and check whether these genes go in accordance with prior knowledge of dentritic cells. Figure 4.1 gives an overview of all experiments.

CHAPTER 5

Results

In the previous chapter we described the experimental methods for our evaluation studies and detailed the performed experiments comprising ODIN, THOR and their competitors. In this chapter, we first describe the results of the experiments performed without replicates, that is, we evaluate in particular ODIN. Second, we explain the results of experiments that contain replicates. Here, we investigate the performance of THOR. Figure 4.1 gives an overview of all performed experiments.

5.1 Experiments with ODIN

We describe the results of experiments with ODIN. First, we describe the findings of the simulated ChIP-seq data without replicates (see Figure 4.1A). We produce simulated, artificial gold standards of ChIP-seq profiles to extensively evaluate DPCs with regard to distinct data characteristics. We then give the results for the biological data without replicates evaluated with the DAGE statistic (see Figure 4.1B1-B4). The DAGE metric is based on the idea of associating changes in protein-DNA bindings with changes in gene expression of the same cellular condition.

5.1.1. Experiments with Simulated Data

As described in Section 4.4.1 we investigate the effect of variable size in the protein domains as well as a variable number of reads in the ChIP-seq experiments. Figure 5.1 shows the results for simulated data without replicates. As expected, methods perform best for experiments with more reads and lower number of proteins per domain (bottom left). The reason is that lower number of proteins per domain shape isolated peaks in the signal. In combination with a large number of reads per protein, these peaks additionally are well-defined. As a consequence, this scenario presents an easy differential peak calling problem. The performance of MAnorm for top ranked DPs is quite competitive with ODIN variants for data with few proteins per domains (left column), but its performance deteriorates whenever more proteins are present in the domains. MACS2 has similar performance as ODIN variants when large number of reads are present (bottom), but ODIN clearly outperforms MACS2 when peak sizes have a high variance (middle row). We calculate the Area Under the Curve (AUC) of ROC curves for each experiment to perform the Friedman-Nemenyi test. We then evaluate the overall performance of methods for all conditions (see Table 5.1 and Sup. Table A.1 for the Friedman ranking). Results indicates that ODIN with Binomial or single Poisson distribution has a significantly higher AUC scores than MACS2, MAnorm, DESeq-MACS and DBChIP (p-value < 0.1). Other tools do not significantly outperform any other tool.

	ODIN-Binomial	ODIN-Poisson-1	ODIN-Poisson-4	ODIN-Poisson-3	ODIN-Poisson-2	MACS2	MAnorm	DESeq-MACS	DBChIP
ODIN-Binomial									
ODIN-Poisson-1									
ODIN-Poisson-4									
ODIN-Poisson-3									
ODIN-Poisson-2	+								
MACS2	*	+							
MAnorm	*	+							
DESeq-MACS	*	*	*	+					
DBChIP	*	*	*	*					

Table 5.1.: Results for the simulated datasets without replicates. The table is based on the Friedman-Nemenyi hypothesis test and the AUC scores. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.

5.1.2. Genomic Signal Construction

In the following, we resort to biological data. For this experiment we restrict our analysis to chromosome 1, which we then discard for the further analysis with biological data sets.

We evaluate the impact of the preprocessing steps performed to create and improve the ChIP-seq signal (see Section 3.2 and Figure 4.1B2). In particular, we analyze all 8 combinations of using: (1) the GC-content model, (2) filtering reads aligned to poor mappability regions and (3) the subtraction of input-DNA. The Friedman-Nemenyi test on DAGE statistics ($h = 50$, $H = 500$) indicates a slight advantage of using input-DNA subtraction and GC-content model compared to using none of the steps for TF data (p-value < 0.1). No significant difference is detected on histone data. However, the Friedman score ranks are similar in both scenarios reinforcing the advantage of the input-DNA subtraction and GC-content model, which will be further used in all experiments (see Sup. Table A.2 – Sup. Table A.5).

5.1.3. Method Parametrization

We evaluate the use of parameter constraints and the choice of the HMM's emission distribution as presented in Section 3.3.4. We compute the DAGE statistic ($h = 50$, $H = 500$) with or without parameter constraints.

First, the constrained model has statistically significant higher DAGE values (p-value $<$ 0.006, one-tailed Wilcoxon test) for experiments with TFs, while no significant differences are obtained for histone modification experiments. This reinforces the advantage of parameter constraints, which are used in further experiments.

Furthermore, we evaluate the use of distinct emission distributions for ODIN: Binomial and mixture of Poisson with 1 to 4 components. As shown in Sup. Table A.6 – Sup. Table A.9, no significant difference was found. We therefore use the Poisson mixture with the number of components that offers the highest ranking (four components for histones experiments

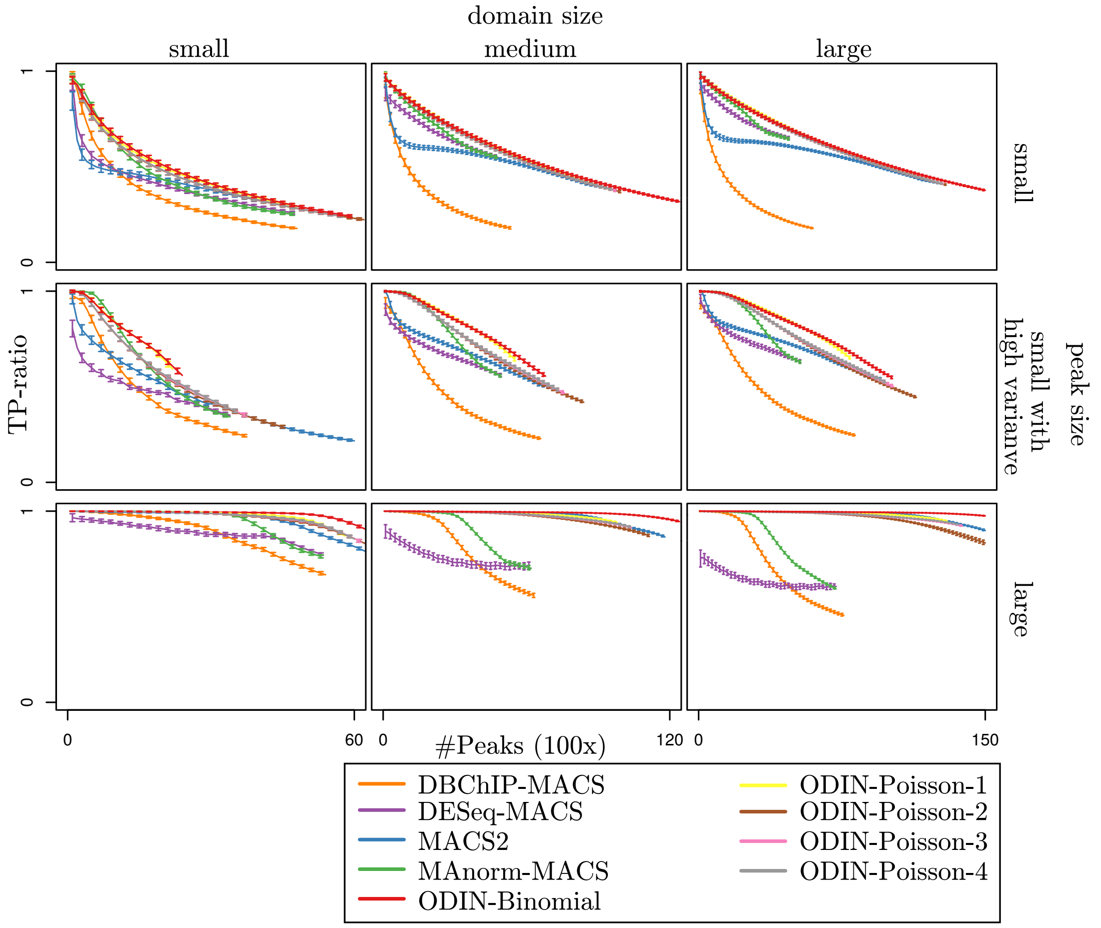


Figure 5.1.: Average ratio of true positive DPs for all compared methods over nine distinct parameters of the simulated data. The number of reads per peak increases from top to bottom (small: $m_2 = 20,\ p_2 = 200$, small with high variance: $m_2 = 20,\ p_2 = 2000$, large: $m_2 = 100,\ p_2 = 200$). The number and variance of proteins within a domain increases from left to right (small: $m_1 = 1,\ p_1 = 4$, medium: $m_2 = 5,\ p_2 = 6$, large: $m_2 = 8,\ p_2 = 14$).

and one component for TF experiments) as well as the Binomial distribution in the following experiments.

Finally, we inspect the impact of SPCs (MACS, QUEST and PeakSeq) on two-stage DPCs DPChIP, DESeq and MAnorm. As shown in Sup. Table A.10 – Sup. Table A.13, no significant difference between the used SPCs was found. We therefore use the best ranked combination for the follow-up experiments: MAnorm-macs, DESeq-quest and DBChIP-quest.

5.1.4. Evaluation *P*-value Estimation Strategies

We evaluate the *p*-value estimation methods of ODIN, DESeq (Anders and Huber, 2010) and EdgeR (Robinson et al., 2010) (see Figure 4.1B1). We use DPs predicted by ODIN with a Binomial distribution for all 15 datasets. ODIN's *p*-value estimation leads to a significant higher DAGE score than the *p*-value estimation of DESeq and EdgeR for TF experiments and a significant higher DAGE score than EdgeR for histone experiments (see Table 5.2; Sup. Table A.14 and Sup. Table A.15 give the Friedman rankings for the experiments).

	TF			histone mod.		
	ODIN	ODIN-DESeq	ODIN-edgeR	ODIN	ODIN-DESeq	ODIN-edgeR
ODIN						
ODIN-DESeq	*					
ODIN-edgeR	*			*	*	

Table 5.2.: P-value estimation evaluation based on histone modification and TF experiments. We estimate p-values with our strategy (ODIN, see Section 3.4.1), the strategy of DESeq (ODIN-DESeq) and the strategy of edgeR (ODIN-edgeR) for histone modification and TF experiments. The asterisk and the cross, respectively, mean that the method in the column outperformed the method with regard to the DAGE scores in the row with significance levels of 0.05 and 0.1.

5.1.5. Comparative Evaluation on Biological Data

We describe the DAGE results for all considered methods (see Table 2.1) based on all biological datasets without replicates (see Table 4.2). See B3 in Figure 4.1 for this specific experiment. In Figure 5.2 we display DAGE curves for DBChIP, MACS2, our DESeq approach, MAnorm and ODIN for four selected experiments on real data. As ChIPDiff does not provide information to sort the DPs, its results are only represented as points, where the x-axis location corresponds to the number of called DPs. In most scenarios, curves approximate to zero for higher ranks, which indicates that higher ranked DPs are associated with higher expression changes. In some scenarios, such as TLR4-H3K4me2-0h-6h, the curve associated with `Gain 2` DPs (Figure 5.2H) are further from 0 than `Gain 1` DPs (Figure 5.2G). This is an indication that there are more changes in ChIP-seq peaks and gene expression in signal 2 (6h) than in signal 1 (0h). This is in accordance with the main message of the TLR4 study, which shows that induction of TLR4 promotes new enhancers marked by H3K4me2 (Kaikkonen et al., 2013). While there are some experiment-specific variations, both ODIN variants outperform other methods on DC-PU.1-MPP-CDP (Figure 5.2A, B), DC-H3K4me1-MPP-CDP (Figure 5.2C, D). Furthermore, the performance of ODIN with Binomial distribution is similar to other methods on TRL4-H3K4me1-0h-6h (Figure 5.2G, H) and TRL4-PU.1-0h-6h (Figure 5.2E, F). All DAGE curves can be found at Sup. Figure A.1 – Sup. Figure A.4.

Moreover, we evaluate the performance of all methods for all 15 real data experiments listed in Table 4.2 by computing the DAGE scores for `Gain 1` and `Gain 2` peaks. The Friedman-Nemenyi test indicates that both ODIN variants have significantly higher DAGE scores than DBChIP and MACS2 on TF experiments (p-value < 0.1, see Table 5.3, Sup. Table A.16 gives the Friedman ranking) and significantly higher DAGE scores than DESeq on histone data (p-value < 0.05, see Table 5.4, Sup. Table A.17 gives the Friedman ranking). Moreover, ODIN with Binomial distribution has a significant higher DAGE score than MACS2 and MAnorm on histone and TF data.

We also performed an evaluation of ChIPDiff by comparing the DAGE values of all methods with H equal to the number of peaks called by ChIPDiff. ODIN with Binomial distribution has significantly higher DAGE scores than ChIPDiff on TF experiments (p-value < 0.1, see Sup. Table A.18 and Sup. Table A.19), while no statistical difference was detected on histone data (see Sup. Table A.20 and Sup. Table A.21). In all cases, both ODIN with Bino-

	ODIN-poisson-1	ODIN-binomial	MAnorm-macs	MACS2	DBChIP-quest
ODIN-poisson-1					
ODIN-binomial					
MAnorm-macs	∗				
MACS2	∗	+			
DBChIP-quest	∗	∗			

Table 5.3.: DAGE results based on TF experiments. Friedman-Nemenyi hypothesis test results for the AUC metric. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.

	ODIN-binomial	ODIN-poisson-4	MAnorm-macs	MACS2	DEseq-quest
ODIN-binomial					
ODIN-poisson-4					
MAnorm-macs	∗				
MACS2	∗				
DEseq-quest	∗	∗			

Table 5.4.: DAGE results based on histone experiments. Friedman-Nemenyi hypothesis test results for the AUC metric. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.

mial and mixture of Poisson distribution ranked best by the Friedman score compared to all competing methods.

We perform a visual inspection of the DPs from experiment TLR4-H3K4me2-0h-24h around gene Irf1. In Figure 5.3 we show the DP estimates for the same genomic region already shown in Figure 2.6. MAnorm and our approach DESeq-quest can successfully detect changes in large peak areas. ChIPDiff detects most DPs, but have a tendency to call large regions. ODIN and MACS2 are able to detect detailed changes within the large domains. MACS2 and ChIPDiff are not able to recover a DP upstream of Irf1 (marked as DP2) in H3K4me2 0h. The loss of this histone mark after TLR4 treatment is supported by gain of PU.1 on the very sample location for PU.1 ChIP-seq profiles.

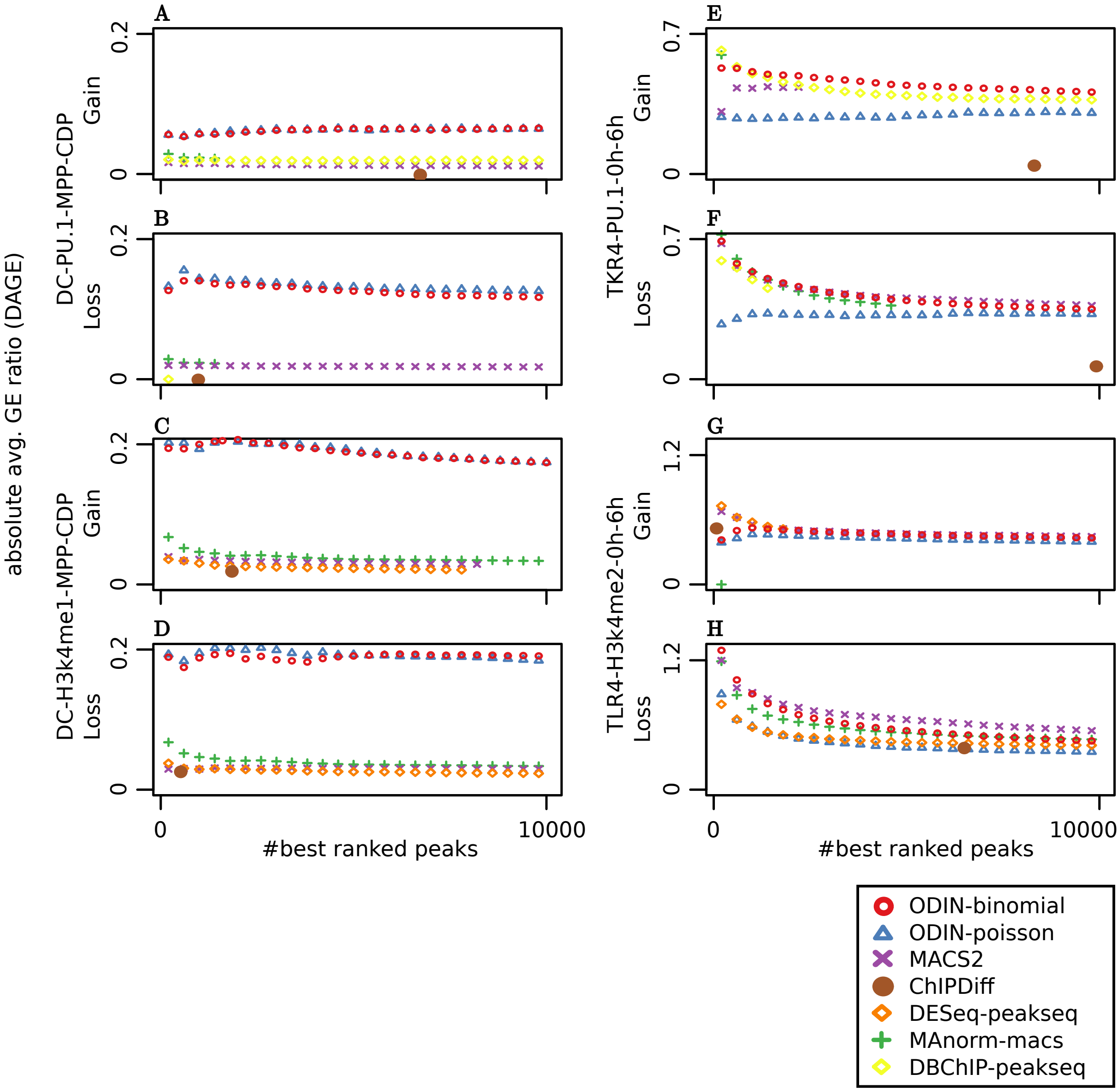


Figure 5.2.: Here we depict the DAGE curves for selected experiments from TLR4 and DC studies. Lines in the first and third row represent DP gained in the first signal (`Gain 1`*), while lines in the second and fourth row in the second signal (*`Gain 2`*).*

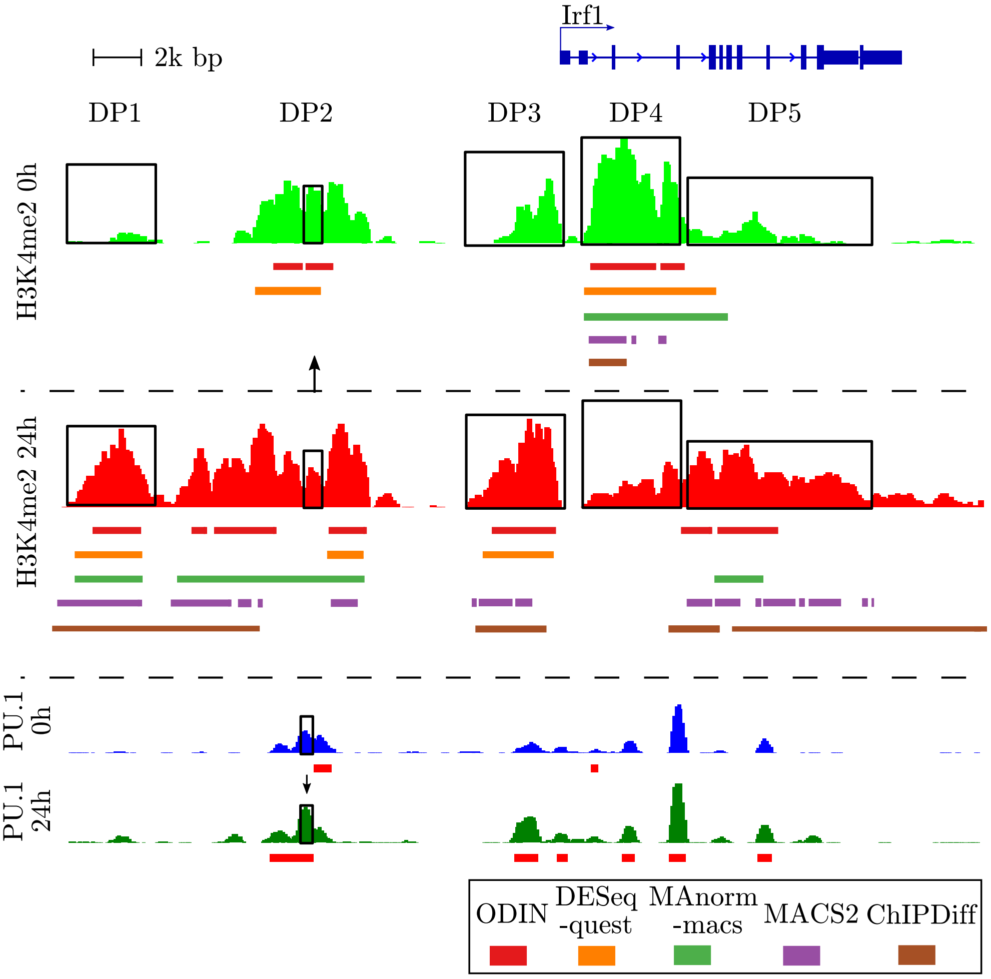


Figure 5.3.: DPs detected on experiment TLR4-H3K4me2-0h-24h around the Irf1 gene. Bars below the ChIP-seq signal indicate the regions called as DPs by distinct methods.

5.2 Experiments with THOR

We describe the results of experiments with THOR and its competing methods. First, we describe the findings of the simulated ChIP-seq data with replicates (see Figure 4.1C). Simulated gold standards of ChIP-seq profiles help to evaluate DPCs with regard to distinct data characteristics. Next, we detail the results for the biological data with replicates which we evaluate with the DCA statistic (see Figure 4.1D1-D4). Similar to the DAGE metric, the DCA metric is based on the idea of associating changes in protein-DNA bindings with changes in gene expression of the same cellular condition. Finally, we describe two use cases for THOR (see Figure 4.1E). The first use case describes the ability of THOR to call DPs that support rSNPs. The second use case is about the association of DPs to genes.

5.2.1. Experiments with Simulated Data

Figure 5.4 shows the distributions of AUC values for all methods and experimental combinations. The first simulation parameter is the number of replicates (red vs. green lines). We observe that most methods have lower AUC values in experiments with 2 replicates (red line) than with 4 replicates (green line) (p-value < 0.05; one-sided Wilcoxon test). Exceptions are Poisson-THOR and IDR. IDR returns very few peaks on cases with 4 replicates (green line), even when using an lenient threshold the SPC used as input for IDR. Poisson-THOR's poor performance on 4 replicates stems from its simple distribution, which does not cope with overdispersion.

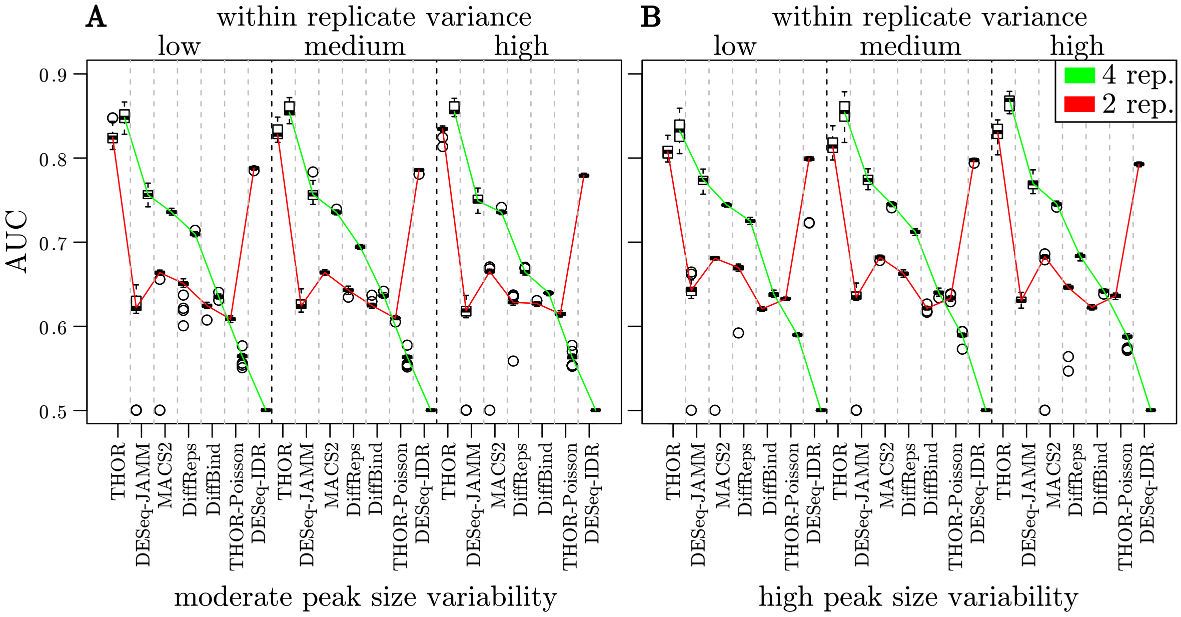


Figure 5.4.: Results for simulated data with replicates. We show the AUC distribution for 25 repetitions of each scenario. Simulated data were based on moderate (A) and high (B) condition peak size variability and 2 (red lines) and 4 (green lines) replicates. Each boxplot is divided by the level of within-condition variance (low, medium and high). Methods (x-axis) are ordered by decreasing median AUC values (y-axis) for the cases with 4 replicates.

The second simulation parameter is the variance of the peak sizes, where we evaluate scenarios with moderate (Figure 5.4A) and high (Figure 5.4B) variance. Two methods have higher AUC values in scenarios with moderate peak variance. THOR in case of low and medium within-condition variance and 2 replicates (p-value $< 10^{-4}$, one-sided Wilcoxon test), as well as DiffBind in the case with low, medium and high within-condition variance and 2 replicates (p-value $< 4.7 \cdot 10^{-8}$, one-sided Wilcoxon test). All other methods show no significant changes in AUC values.

The third characteristic is the level of variance within the replicates. DESeq-JAMM, DiffReps and DESeq-IDR show decrease in AUC values for increasing variance. Interestingly, the performance of THOR, MACS2 and DiffBind shows increase in AUC values with increasing variance for respectively 5, 3 and 6 of the eight cases (p-value < 0.05; one-sided Wilcoxon test).

Finally, we apply the Friedman-Neymeni test for all data together to evaluate the statistical significance in AUC value differences for distinct methods. THOR has significantly higher AUC scores than all competing methods. MACS2 has significant higher AUC values than DiffReps, DESeq-IDR, DiffBind and Poisson-THOR; and DESeq-JAMM and DiffReps have significantly higher AUC values than DiffBind and Poisson-THOR (p-value < 0.05, see Table 5.5 and Sup. Table A.22 for the Friedman ranking). Evaluating specific conditions, THOR has significantly higher AUC values than DiffReps, DiffBind and Poisson-THOR for all 12 cases (p-value < 0.05, Sup. Table A.23 – Sup. Table A.34). In the case with 2 replicates, THOR additionally has significantly higher AUC values than DESeq-JAMM (p-value < 0.05, Sup. Table A.23 – Sup. Table A.28) and in the case with 4 replicates significantly higher AUC values than DESeq-IDR (p-value < 0.05, Sup. Table A.29 – Sup. Table A.34). THOR is ranked top in all of the 12 cases.

	THOR	MACS2	DESeq-JAMM	DiffReps	DESeq-IDR	DiffBind	Poisson-THOR
THOR							
MACS2	$*$						
DESeq-JAMM	$*$	$+$					
DiffReps	$*$	$*$					
DESeq-IDR	$*$	$*$					
DiffBind	$*$	$*$	$*$	$*$			
Poisson-THOR	$*$	$*$	$*$	$*$	$*$	$*$	

Table 5.5.: Friedman-Nemenyi test results based on the AUC statistic of simulated data for all scenarios. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.

5.2.2. Initial DP Estimation

In the following, we resort to biological data. Here, we evaluate the impact of the initial DP estimates to train THOR's HMM (see Figure 4.1D3). Similar for the experiments with ODIN, we restrict in this experiment our analysis to chromosome 1. We then discard chromosome 1 for the further analysis with biological data sets.

According to our parametrization (see Section 4.5.2), the initial DPs we use to train THOR's

HMM are based on a minimum difference between signals t_1 and a fold change t_2. We set the minimum signal support $t_3 = t_1/2$. We evaluated different parameter settings for $t_2 \in \{1.3, 1.6\}$ and $t_1 \in \left\{\langle x \rangle^{.95}, \langle x \rangle^{.99}\right\}$ by predicting DPs for chromosome 1 for all 12 experiments with replicates. The Friedman-Nemenyi test on DCA statistics for $h = 100$ and $H = 1000$ shows no statistically significant differences, which indicates that THOR is robust against distinct initial parameter setups (see Sup. Table A.35 and Sup. Table A.36). We used the parametrization with the highest ranking ($t_1 = 1.6$, $t_2 = \langle x \rangle^{.95}$) for all further experiments.

5.2.3. Quality Analysis on Biological Data Sets

To better understand the characteristics of ChIP-seq experiments evaluated in our study (see Table 4.3), we first perform a quality check. For this we use the FRiP (fractions of reads in peaks) score from the ENCODE consortia (Landt et al., 2012) which gives an estimate of the signal-to-noise ratio of ChIP-seq experiments (see Section 2.2.6). We also propose the use of the quadratic coefficient, that is, the variable c_{1G_k} in Equation 3.17 which describes the mean-variance function, for a given biological condition as an indicator for "overdispersion". Figure 5.5A and B give two examples of the mean-variance function of selected experiments. Overdispersion positively correlates with the number of replicates in the condition (R=0.74, adjusted p-value=0.0001; Spearman Correlation). Moreover, higher overdispersion is associated with lower FRiP scores (R=-0.78; adjusted p-value=$2.9 \cdot 10^{-5}$). As depicted in Figure 5.5C, average FRiP vs. overdispersion space separates the experiments by their expected complexity. The dendritic cell (DC) differentiation experiments, which were obtained by in vitro differentiation of cells with technical replicates, have highest FRiP values and lowest overdispersion values. The follicular lymphoma experiments (LYMP), which arise from patients with distinct genetic background and with potential tissue heterogeneity, have both highest overdispersion scores and lowest average FRiP. This indicates that the experiments evaluated here cover a large spectrum of peak size variance within biological conditions.

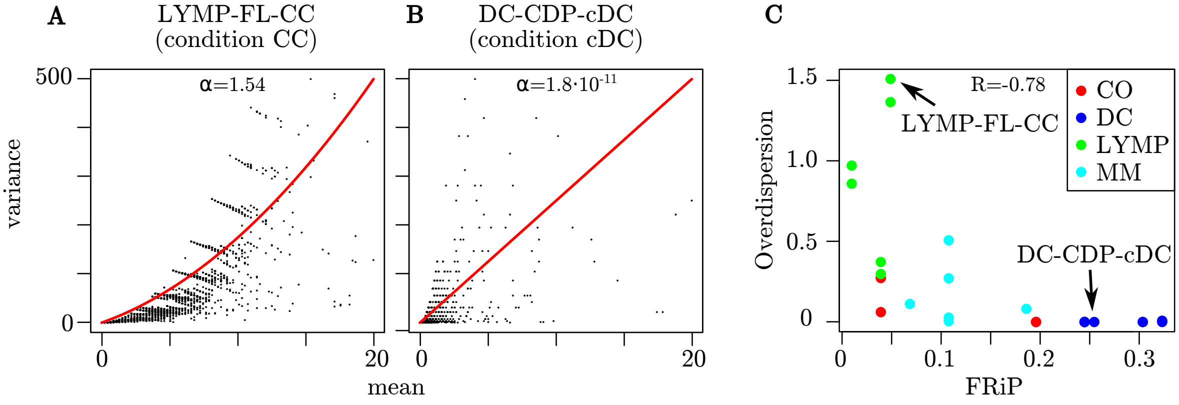


*Figure 5.5.: **(A)** and **(B)** Two examples for the mean-variance function described by Equation 3.17. A high value (A) of c_{1G_k} gives the function a quadratic and a low value (B) a linear shape. **(C)** Association between average FRiP and overdispersion scores. FRiP and overdispersion scores for the 24 biological conditions analyzed: cocaine intake (CO), monocyte differentiation (MM), lymphoid cancer (LYMPH) and dendritic cell differentiation (DC). Higher FRiP indicates higher signal-to-noise ratio and better ChIP-seq experiments. Higher overdispersion scores indicates higher within-condition variability. Arrows indicate the experiments described in (A) and (B).*

5.2.4. Comparative Analysis of Biological Data

We evaluate THOR and six competing methods (csaw, MACS2, DiffReps, PePr, DiffBind and DeSeq-IDR) on 12 differential peak calling problems using data from the cocaine intake on mice (CO), dendritic cell differentiation (DC), B cell follicular lymphoma (LYMP) and monocyte differentiation (MM) study (see Section 4.3 and Table 4.3)[1]. We also evaluate the application of THOR with either the TMM (THOR-TMM) or the housekeeping genes (THOR-HK) normalization approach (see Section 3.2.6). The performance of the methods was evaluated with the DCA (Differential Correlation Analysis) statistic. See D1 in Figure 4.1 for this specific experiment. Figure 5.6 shows selected examples for DCA curves. Sup. Figure A.5 – Sup. Figure A.8 give all 12 DCA curves.

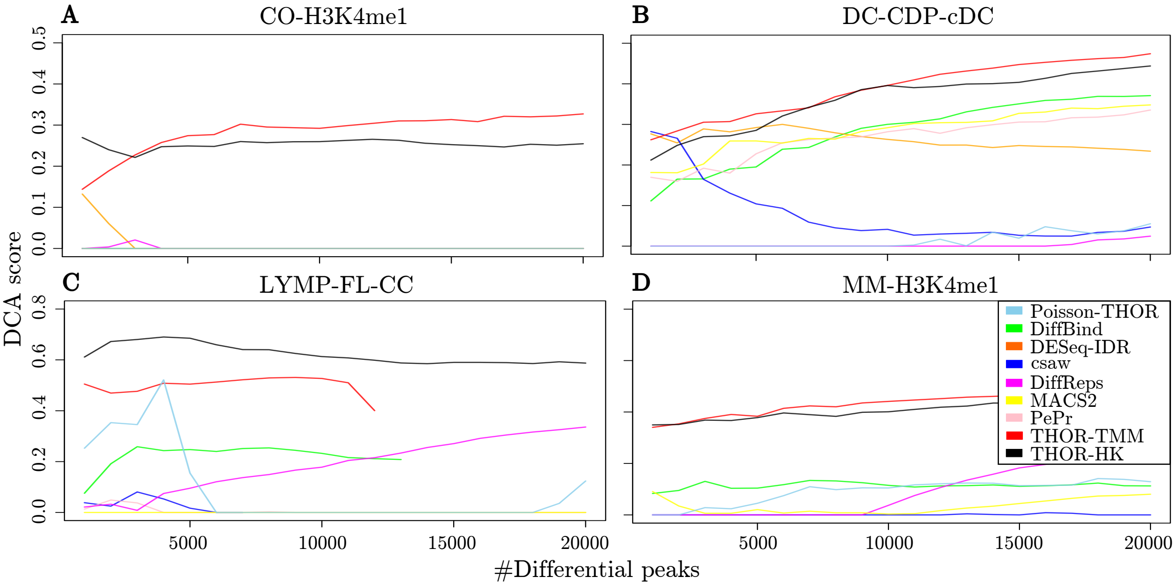


Figure 5.6.: DCA curves for four selected differential peak calling problems. We show DCA curves for (A) CO-H3K4me1, (B) DC-H3K27ac-CDP-cDC, (C) MM-H3K4me1 and (D) LYMP-FL-CC. Higher DCA values indicate higher association between differential peaks and differential expression.

We use the Friedman-Nemenyi test to check for significant differences in the area under the DCA curves (see Section 4.7). THOR with both normalization strategies is the best ranked method and has significantly higher DCA values than DESeqIDR, csaw and Diffbind (adjusted p-value<0.05, Table 5.6 and Sup. Table A.37 for the Friedman ranking). MACS2 also has significantly higher DCA values than csaw (adjusted p-value<0.05).
As PePr requires input-DNA data and therefore cannot be executed for MM and CO, we repeat the Friedman-Nemenyi test on DCA values from DC and LYMP only. In this case, THOR-HK has significantly higher DCA score than csaw (p-value<0.05) as well as Poisson-THOR (p-value<0.1, Sup. Table A.38 – Sup. Table A.39) and THOR-TMM outperforms csaw (p-value<0.05). There is no significant differences for all other competing methods.

As an example, we show in Figure 5.7 DPs for H3K4me3 histone modification on the monocyte to macrophage differentiation (MM) experiments of genes discussed in the original study from Saeed et al. (2014). DiffBind and DESeq-IDR do not call any DPs in this region. THOR calls a combination of gain (green) and decrease (red) in H3K4me3 levels in IRAK3's and PDK2's promoters. Csaw peaks misses regions with large histone changes in

[1] We were not able to execute JAMM on these data sets, which was therefore left out of this analysis.

	THOR-HK	THOR-TMM	macs2	DiffReps	DiffBind	DESeqIDR	Poisson-THOR	csaw
THOR-HK								
THOR-TMM								
macs2								
DiffReps								
DiffBind	$*$	$+$						
DESeqIDR	$*$	$*$						
Poisson-THOR	$*$	$*$						
csaw	$*$	$*$	$*$					

Table 5.6.: Friedman-Nemenyi hypothesis test results for the DCA score ($h = 500, H = 10000$). The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.

both cases. MACS2 only detects a small lost peak in IRAK3 promoter, while DiffReps detects rather large gain peaks in both promoters. The average peak size of all analyzed data supports the fact that DiffReps tends to call larger DPs (1893bps) and MACS2 smaller DPs (296bps) than all other tools (1133bps) (see Figure 5.9).

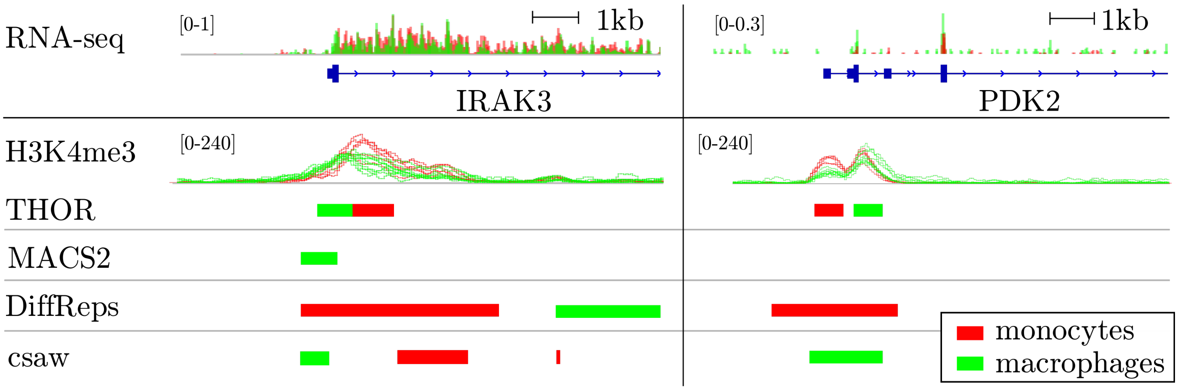


Figure 5.7.: Example of DPs estimates in biological data. We depict an overlay of all H3K4me3 and RNA-seq signals for monocytes (red) and macrophages (green) around the promoter of IRAK3 and PDK2 for THOR and competing methods. We show the 10,000 most significant DPs of each method.

5.2.5. Overdispersion Impact on Differential Peak Calling Performance

We analyze the effect of overdispersion to the DPCs (see Figure 4.1D2). As previously described, follicular lymphoma (LYMP) experiments exhibit highest overdispersion values, while the dendritic cell differentiation study has the lowest. Interestingly, the DCA scores supports the notion that THOR has relatively better performance than competing tools in data with high within-condition variance such as LYMP-FL-CC (Figure 5.6C), while it per-

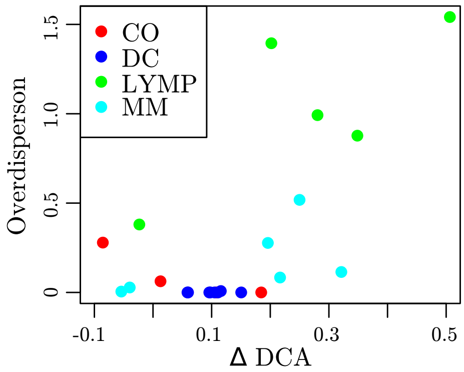


Figure 5.8.: Example of DCA-overdispersion association. Association between the difference of THOR DCA scores with the best competing method and the overdispersion score of 12 differential peak calling problems divided by gain and lost peaks.

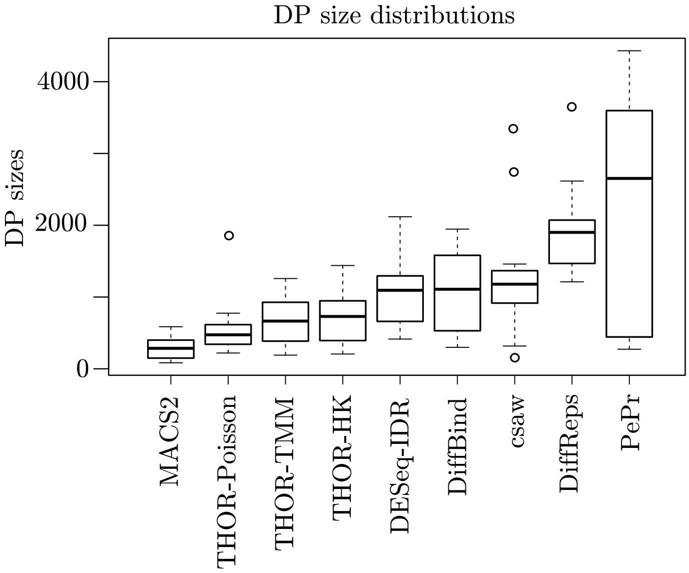


Figure 5.9.: DP size distribution for each tool. The boxplot of each tool gives the DP size distribution obtained from predictions on all biological data.

forms comparatively well with other competing methods such as in DC-CDP-cDC (Figure 5.6B). To investigate this more systematically, we measured the difference between the THOR-HK vs. the best DCA for competing methods. As indicated in Figure 5.8, there is a moderate association between Δ DCA and the overdispersion score (R=0.36; adjusted *p*-value<0.1). The Δ DCA is large for experiments with large number of replicates and genetic variance between samples from MM and LYMP.

Another important question is the performance of the two normalization approaches supported by THOR. While the Friedman rank indicates THOR-HK has overall better ranking, the difference in ranks between the TMM and HK approaches is not statistically significant. Considering the difference in DCA scores for THOR-HK and THOR-TMM (Figure 5.10), we observe that both methods perform similarly in most data sets. However, THOR-HK clearly outperforms THOR-TMM in four conditions from LYMPH (LYMP-CC-PBBA gain, LYMP-CC-PBBA loss, LYMP-FL-CC loss and LYMP-FL-PBBA loss). These conditions have small FRiP (< 0.05) and large overdispersion estimates (> 0.03). These results indicate that in cases with high variance or low signal-to-noise ratio it is advisable to perform a housekeeping gene normalization strategy.

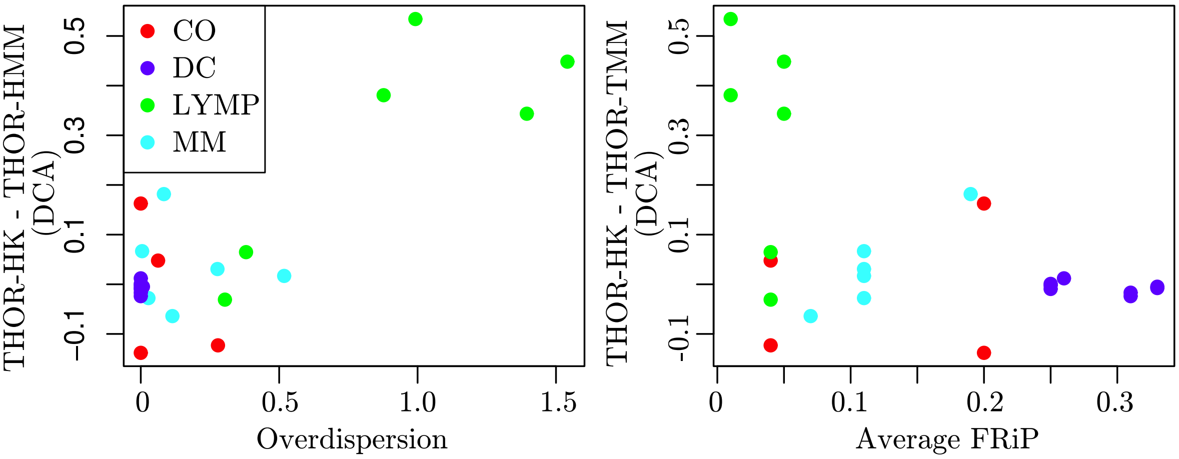


Figure 5.10.: Association between the difference in DCA values for THOR normalization approaches (THOR-HK - THOR-TMM) vs. the average FRiP or overdispersion scores.

5.2.6. Use Cases

We describe two use cases for THOR (see Figure 4.1E). First, we apply THOR to the LYMP-FL-CC data set to evaluate the ability of THOR to call DPs that support rSNPs. Second, we evaluate whether DPs called by THOR for the data sets DC-H3K27ac-CDP-cDC and DC-H3K27ac-CDP-pDC are associated with dendritic cell genes.

Identifying rSNPs

In this experiment we analyze THOR's ability to call DPs that support rSNPs. For the data set LYMP-FL-CC (see Table 4.3), we first call DPs with THOR and second use GATK to identify rSNPs. Next, we filter SNPs that do not lie within DPs (see Sup. Table A.40 for an selection). In Section 4.6.1, we describe in detail how we obtain the 137 candidate rSNPs.

We apply GREAT (McLean et al., 2010) to perform an enrichment analysis of the 203 genes that are neighbouring the candidate rSNPs. GREAT assigns biological meaning to non-coding genomic regions such as rSNPs. For that, GREAT assigns the rSNPs to genes in the vicinity and details the biological function of the genes. All resulting 27 enriched gene sets are associated with lymphoid cells, such as *abnormal lymphocyte morphology* (adjusted p-value=$6.5 \cdot 10^{-4}$; 39 annotated genes) and *abnormal B cell morphology* (p-value=0.0011; 23 annotated genes). Concerning the 28 rSNPs detected in the original study (Koues et al., 2015), there is no overlap of our candidate rSNPs and the GREAT analysis indicates no enriched terms. However, note that Koues et al. (2015) employ a distinct strategy to detect rSNPs, which was based on comparing the reads of single FL patients vs. all CC cells.

Next, we select six genomic regions with seven rSNPs which are close to genes with "abnormal lymphocyte morphology" as indicated by GREAT, lie within a DP with low p-value ($< 10^{-14}$) and the rSNP disrupted (or enhanced) transcription factor binding sites (TFBS). One interesting cluster of rSNPs is present in the locus of the G-receptor gene family RGS. The second ranked (by lowest DP p-value) rSNP is located in the promoter of RGS2 (Figure 5.11A) and two rSNPs (DP ranks 6 and 7) lie in an enhancer region between RGS1 and RGS13 (Figure 5.11B). There are lower levels of H3K27ac around all rSNPs and in the promoters of these genes indicating a decrease of gene activity. We also find that rSNPs change binding affinity of TFBS of B-cell factors Bcl6 (disruption) and Ikaros (enhancement) as well as the repressive chromatin remodelling factor YY1 (enhancement). Genes in these loci (RGS1, RGS2 and RGS13) have been previously reported to be associated with B cell motility (Han et al., 2005) and to regulation of germinal center B cells (Shi et al., 2002).

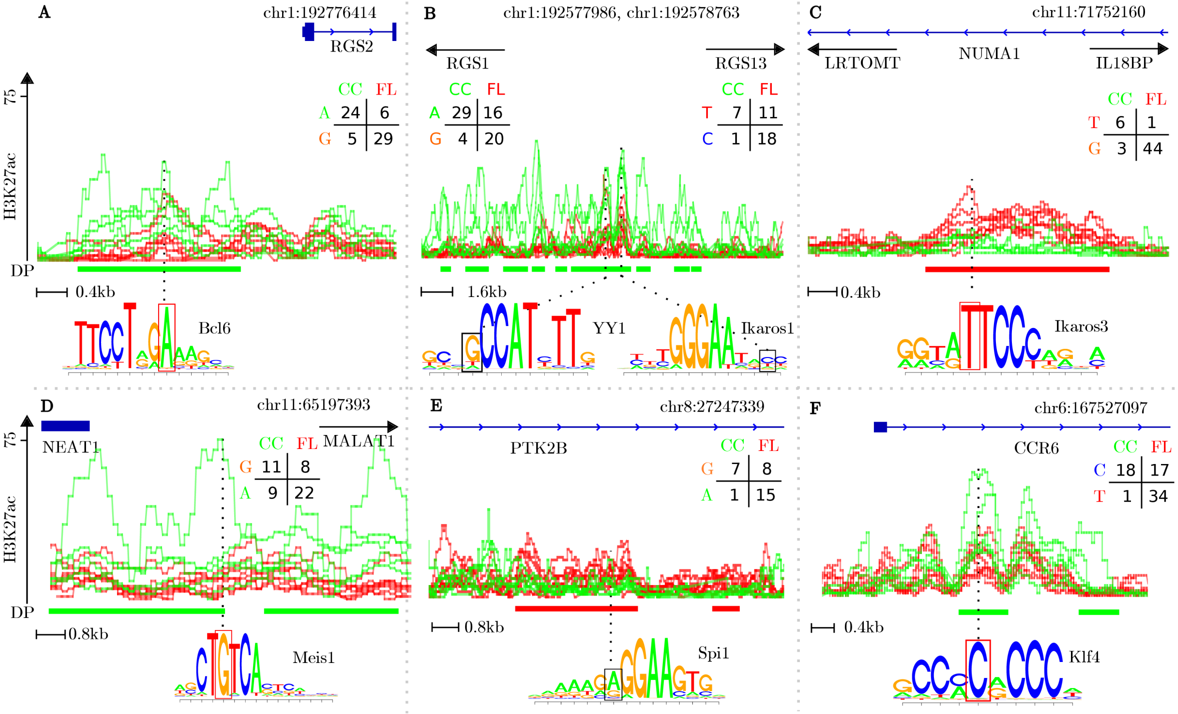


Figure 5.11.: *Selection of regulatory SNPs. We depict FL (red signal) and CC (green signal) located in DPs called by THOR (red/green bars under ChIP-seq signals). For each rSNP, we indicate close genes and a table with the frequency of the common (top) and alternative (bottom) alleles. We also show examples of TFBS motifs being disrupted by the rSNPs. Red (black) boxes indicate the motif position that is disrupted (enhanced).*

The rSNPs in DP rank 5 is in the vicinity of IL-18BP (Figure 5.11C). Both rSNP region and IL-18BP have increased H3K27ac levels in FL patients. Interestingly, the rSNP disrupt a Ikaros binding site. IL-18BP is known to antagonize the IL-18 receptor and Interferon responses in immune cells (Yoshimoto et al., 1998). Another interesting rSNP locus (rank 10) is close to NEAT1 (and MALAT1) (Figure 5.11D). This genomic region displays high losses of H3K27ac on FL condition. Among others, we find disruption of motifs of the Meis1 factor. While MALAT1 and NEAT1 have not been associated with B cell lymphomas, these long non-coding genes have prominent functions in RNA splicing and cancer (Gutschner et al., 2013). Moreover, a rSNP (rank 13, Figure 5.11E) is in the vicinity of the kinase PTK2B, which has increased H327ac levels in FL patients. The rSNP enhanced the binding of the hematopoietic master regulated Sfp1. This kinase has been shown to be relevant for marginal zone B cells in mice (Guinamard et al., 2000).

Finally, rSNP on rank 19 (Figure 5.11F) lies in a intergenic region of gene CCR6 and disrupts the binding of a Klf4 factor. CC chemokines are known regulators of both B cell as well as cancer cell migration and were also among the genes found in the original study (Koues et al., 2015). Altogether, these results indicate the power of THOR by detecting DPs supporting rSNPs.

Dendritic Cell Development Analysis

We evaluate if DPs are close to genes which are relevant for dendritic cell differentiation. Here, we only evaluate genes gained in cDC (compared to CDP) and pDC (compared to CDP) using H3K27ac (see data set DC-H3K27ac-CDP-cDC/pDC in Table 4.3).

First, we perform an analysis with GREAT to evaluate the biological functions associated with genes. For cDC peaks, we obtain *MHC class II protein complex* (adjusted p-value=$1.8 \cdot 10^{-106}$; 8 annotated genes) and *antigen processing and presentation of exogenous peptide antigen via MHC class II* (adjusted p-value=$1.1 \cdot 10^{-76}$; 6 annotated genes). The major histocompatibility complex (MHC) class II is a molecule family that occurs in antigen presenting cells such as dendritic cells (Ting and Trowsdale, 2002). For pDC peaks, the GREAT analysis yields *interferon receptor activity* (adjusted p-value=$5.5 \cdot 10^{-41}$; one annotated gene) and *type I interferon production* (adjusted p-value=$6.1 \cdot 10^{-39}$; one annotated gene). Interferons are a protein family in the response against viruses and cancer cells, which are known functions of dendritic cells (De Andrea et al., 2002).

Second, we evaluate DPs by assigning them to genes in their vicinity. Sup. Table A.41 gives the top 50 ranked genes that are close to DPs gained in cDC cells. ID2 (rank 4) is an important factor associated with cDC differentiation, and is known to have low expression in the precursors cell CDP (Jackson et al., 2011). Also, receptor genes ADAM19 (rank 13) and KIT (rank 47) are known markers of cDC cells (Miller et al., 2012). Six of the top eight ranked genes, that is, H2-AA, H2-AB1, H2EB-1, H2-EA-PS, H2-EB2 (see Figure 5.12C) and CD74 (see Figure 5.12D), as well as several further other listed genes (H2-DMB1 (rank 16), H2-DMB2 (rank 17), CIITA (rank 28) and H2-OB (rank 37)) code proteins that are part of the MHC class II family. Moreover, gene IRF8 (rank 14) regulates distinct stages of the DC differentiation (Jackson et al., 2011).

Sup. Table A.42 lists the top 50 genes close to pDC gain peaks. The top ranked gene SIGLECH (see Figure 5.12A) is a receptor widely used as pDC identification marker (Zhang et al., 2006). Moreover, gene IRF8 (rank 2, see Figure 5.12B), gene TCF4 (rank 41) and gene RUNX2 (rank 49) play key roles in the development of pDC cells (Jaiswal et al., 2013; Cisse et al., 2008; Sawai et al., 2013). Gene PACSIN1 (rank 36) regulates the interferon response specifically in pDC cells. Gene IFNAR1 (rank 9) and IL10RB (rank 10) code interferon receptors. Altogether, THOR identifies several genes that are associated with either cDC or pDC cells. This indicates that THOR's results go in accordance with prior biological knowledge.

5.3 Discussion

Figure 4.1 gives an overview of all experiments we have performed for the evaluation of ODIN, THOR and their competitors. We evaluate ODIN with simulated and biological data sets. ODIN significantly outperforms all competing methods on the simulated data, where we vary the size of the protein domains as well as the size of peaks within the domain. With regard to the DAGE metric, the performance of MACS2, DBChIP and DESeq is worse than ODIN's independently from the protein of interest (TF or histone modifications). This emphasizes ODIN's flexibility which is caused by the use of the HMM for the signal segmentation step. MACS2 predicts simulated DPs well for the easy scenario with high peak sizes, but shows poor performance for peaks with a high variance in their size. MAnorm in combination with the SPC MACS also provides good results for the simulated as well as biological data which can be explained by its sophisticated normalization strategy. ChIPDiff also uses an HMM to identify DPs. However, it lacks the application of pre- and postprocessing steps resulting in a poor overall performance.

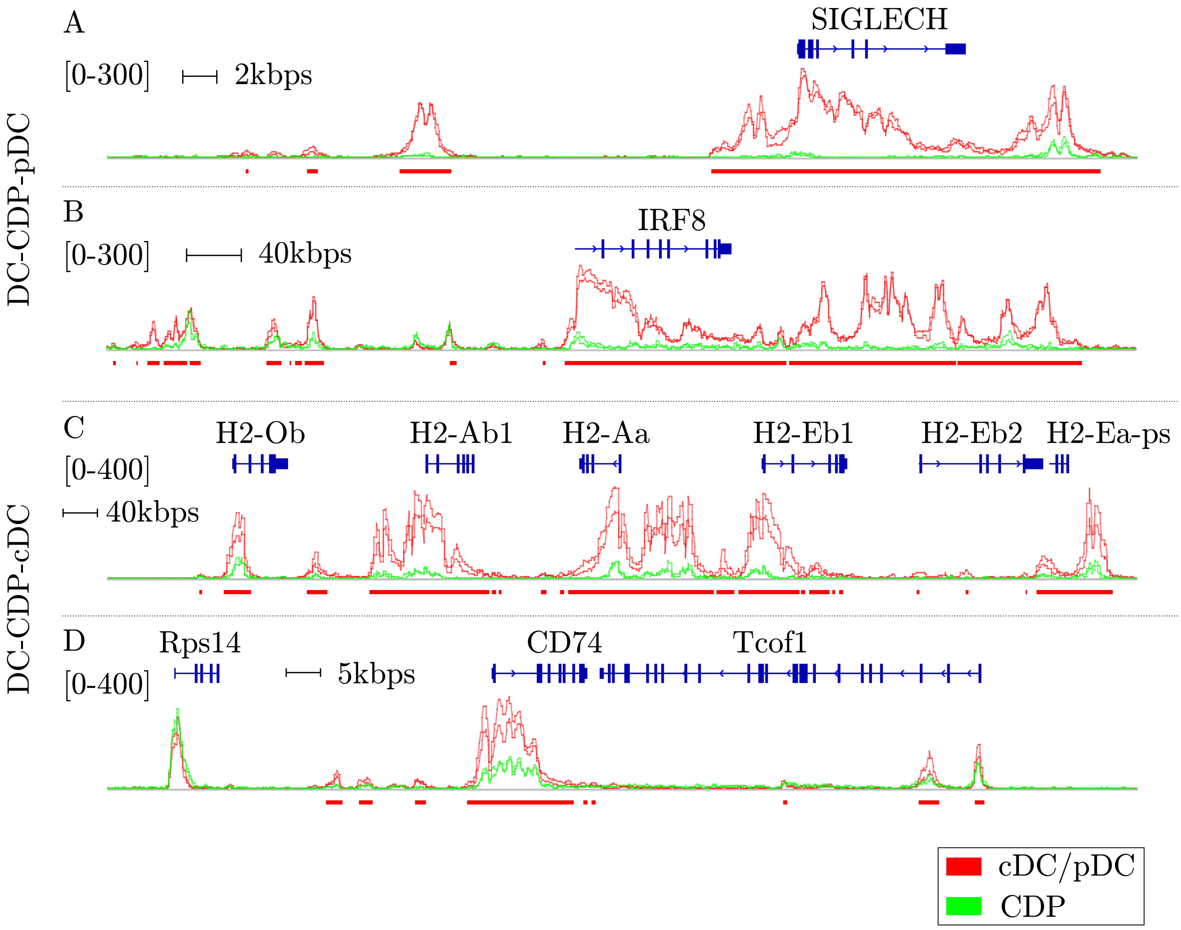


Figure 5.12.: Genes with most significant DPs. We give examples of genes that are associated with the most significant DPs for histone modification H3K27ac with two replicates. For the data set DC-H3K4ac-CDP-pDC (top) we show genes SIGLECH (A) and IRF8 (B), and for data set DC-H3K4ac-CDP-cDC (bottom) we picture gene CD74 (C) and genes that code of the MHC class II family (D).

THOR outperforms competing methods for most simulated and biological data sets. Concerning the biological data, the difference in performance between THOR and its closest competing method is relatively higher for data with high overdispersion and low quality. Moreover, THOR with the housekeeping gene normalization approach is the top ranked method for the biological data. In particular, THOR performs best for experimental conditions from the follicular lymphoma study. This study has overall lowest quality statistics (FRiP) and highest within-condition variance scores (overdispersion). Indeed, THOR's framework includes the estimation of overdispersion quality measures, which can be used to guide the choice of normalization strategy.

One competing method with an overall good performance is MACS2 (unpublished), which was ranked third on simulated data and second on biological data. Although there is no current description of MACS2, it is based on the framework of the widely used SPC MACS (Zhang et al., 2008). The performance of other tools varied across distinct experiments. While DESeq-IDR performed well on simulated data cases with low within-condition variance and low number of replicates, it failed to call peaks on data with large variance. This is expected as IDR was conceived for a conservative peak detection on technical replicates. JAMM (with DESeq) had good performance on simulated data and is the only framework performing integrative analysis of single signal peak calling problems with replicates. Some methods, such as PePr and DiffReps, had a tendency to call peaks larger than other tools and the observed histone changes. This explains the average performance of these methods in our evaluation. Poisson-THOR, which can be seen as a version of ODIN supporting replicates with a distribution not coping with overdispersion, has poor results in most evaluated scenarios. This reinforces the importance of support to overdispersion on the presence of replicates.

We also demonstrate THOR's power to call DPs by applying it to two use cases. First, we show that THOR calls DPs that support rSNPs. We call SNPs for the leukemia study (LYMP-FL-CC in Table 4.3) and check whether they fall into DPs called by THOR. Among all methods, THOR provides DPs with the highest amount of covered SNPs. We show that SNPs potentially influence TFBS which are associated with leukemia.

In the second study, we assign DPs called by THOR to genes for cDC and pDC peaks compared to CDP for H3K4me1. As expected, the resulting genes are associated with dendritic cells and their development. The identified genes go in accordance with prior knowledge of dendritic cells and in particular of their differentiation. Both use cases give reasonable findings which emphasize THOR's usefulness for the DP identification. In combination with our evaluation studies, which are based on simulated, artificial gold standards as well as the DCA metric, the use case studies emphasize that THOR is a powerful method for the ChIP-seq analysis.

CHAPTER 6

Conclusion

This thesis contributes to the computational analysis of ChIP-seq data. The main goal of this work was to develop algorithms that call differential peaks (DPs) in ChIP-seq data. We propose the one-stage DPCs ODIN and THOR for the case without and with replicates in ChIP-seq experiments. Both methods are based on a hidden Markov Model (HMM) to identify DPs, as HMMs segment the ChIP-seq signal by detecting peaks with variable size through the use of posterior decoding algorithms. Competing methods, such as MACS2, MAnorm, DBChIP, DiffBind and DiffReps, use pre-defined candidate peaks to identify DPs. This approach highly depends on the initial peak calling step and fails to detect subtle changes within complex histone modification peak landscapes. Other methods, such as PePr and csaw, use a window-based approach to segment the signal. However, the performance of this strategy highly depends on the choice of the window size. Moreover, heuristic methods have to be applied to merge windows in close vicinity to each other.

In the case without replicates, ODIN uses an HMM with a Binomial or Poisson distribution and in the case with replicates, THOR uses a Negative Binomial distribution as emission to handle overdispersion. Modelling overdispersion, which ChIP-seq data typically exhibits, is crucial for an accurate DP calling process. As there is no analytical solution for the Baum-Welch algorithm with a Negative Binomial distribution, we estimate the corresponding parameters based on an empirically evaluated mean-variance function. With regard to the HMM emission, we also propose a p-value estimation strategy to identify significant DPs.

A crucial aspect for the differential peak calling problem is the normalization, as ChIP-seq profiles typically exhibit different sequencing depths as well as different signal-to-noise ratios. The widely used TMM normalization approach, which was originally developed for gene expression analysis, is based on the assumption that the number of reads in most genomic regions does not change across the conditions. This is not necessarily the case for protein interactions, as two distinct cells can have distinct amounts of proteins or histone modifications bound to their DNA. Particularly problematic with TMM is the effect of replicate-specific background noise. Hence, we propose the use of control regions for normalization. We propose housekeeping (HK) genes as control regions for the analysis of active histone marks. Our experiments show that THOR's DP predictions lead to highest DCA scores when we apply the HK gene normalization (see Table 5.6). For other proteins of interest, ChIP-PCR measurements can provide regions which can serve as control regions.

Several pre- and postprocessing steps are necessary to call DPs in ChIP-seq data: read filtering, fragment size estimation, GC-content normalization, control DNA normalization, sample normalization and artefact filtering. See Table 2.1 for an overview of the features implemented in the competing methods. ODIN and THOR perform all required steps which makes them the most complete methods for solving the differential peak calling problem.

Evaluation of DPCs is still an open problem in the research community. To our best knowledge, we perform the most comprehensive evaluation study available. First, we use biological data sets for the evaluation. In the case without replicates, we propose the DAGE and, in the case with replicates, the DCA metric. Both metrics are based on the idea of associating changes in protein-DNA bindings with changes in gene expression of the same cellular condition. For the case without replicates, we use 15 data sets to quantify 5 competing methods (MACS2, ChIPDiff, DESeq, MAnorm and DBChIP); and for the case with replicates, we use 12 data sets to evaluate 7 competing methods (DiffReps, DiffBind, MACS2, csaw, DESeq-IDR, DESeq-JAMM and PePr). To obtain a comprehensive picture of the performance, we use data sets from different kinds of proteins (TFs and activating histone marks), which results in different peak sizes in the signal.

Second, we propose an algorithm to simulate ChIP-seq reads of two biological conditions with potential replicates that contain DPs. The simulation of ChIP-seq profiles can produce artificial and customized gold standards which can be used to extensively evaluate DPCs with regard to distinct data characteristics. Our algorithm is highly parametrized, such that a broad range of gold standards can be produced. Furthermore, our algorithm produces raw ChIP-seq reads, which have to be mapped to a reference genome. Thereby, bias induced by the mapping process is taken into account. In the case without replicates, we vary the protein domain size and the peak sizes. In the case with replicates, we account for various peak sizes, within-condition variance and number of replicates. These experimental setups cover a wide range of characteristics of differential peak calling problems.

Altogether, our evaluation studies show that ODIN and THOR are the best performing methods. The performance is justified from their methodological aspects, as both tools use an HMM to intrinsically analyses windows of varying size during the detection of DPs.

6.1 Future Work

A natural extension of ODIN and THOR is to consider more than 2 biological conditions which can yield more intricate epigenetic findings in biological and medical research. For example, we could evaluate time series data as it is provided by Kaikkonen et al. (2013). Also, we could analyze developmental series as described in the dendritic cells study.

ODIN and THOR use an HMM to segment the ChIP-seq signal. We restrict our analysis to the case of small or medium sized peaks. However, HMMs are flexible, such that our methods could also call peaks in large protein domains. A systematic evaluation of this idea could potentially offer a wide range of further applications for ODIN and THOR.

Our normalization approach is based on housekeeping genes as control regions in the case of active histone marks. The rationale is that control regions on the one hand should have a similar signal across the conditions and on the other hand exhibit a low noise. We could also evaluate the use of consensus peaks across the conditions as control regions. Consensus peaks could be called by IDR or JAMM. However, this approach includes the additional peak calling step in the normalization procedure. Peak calling is error prone and its performance depends on the method's parametrization as well as the shape of the peaks in the ChIP-seq profiles.

The idea of DCA and DAGE is that gene expression correlates to certain histone modifications of the same cellular conditions. However, this approach can oversimplify the problem. Maze et al. (2014) state that histone modification levels can exhibit opposite changes with regard to the gene expression among a gene. Moreover, the interplay between certain proteins and histone modifications may lead to changes in the histone modification level. For example, while the histone modification H3K4me3 is usually enriched at active gene promoters and therefore associated with transcription, its level can also decrease at certain genes when RNA polymerase II interacts with the gene. Hence, the working assumption for our metrics is that in the majority of cases certain histone modification are consistently associated with gene expression changes in the same biological condition. Furthermore, it is possible to generalize this idea of DCA and DAGE. Instead of only considering gene expression data, DPs that are based on active histone modifications can be associated with other active histone modification that are known to be featured in the same biological conditions.

Furthermore, calling DPs in signals describing distinct biological conditions is not restricted to the ChIP-seq technique. A novel sequencing application is SHAPE-seq (Lucks et al., 2011; Loughrey et al., 2014) which investigates the RNA structure. For SHAPE-seq, solving the differential peak calling problem is potentially as important as for ChIP-seq, since comparing the RNA structure of two biological conditions leads to a deeper understanding of the underlying biological mechanisms in cells. However, one challenge is that the SHAPE-seq protocol produces artefacts that differ to the ones from ChIP-seq. Applying SHAPE-seq makes it necessary to consider other technology-specific artefacts.

APPENDIX A

Appendix

A.1 Results with ODIN

A.1.1. Simulation

AUC	
ODIN-binomial	1.3333
ODIN-poisson-1	2.0
ODIN-poisson-4	3.4444
ODIN-poisson-3	4.2222
ODIN-poisson-2	5.3333
MACS2	5.8889
MAnorm	6.0
DESeq-macs	8.2222
DBChIP	8.5556

Table A.1.: Friedman ranking for the results based on the simulated data. For each metric, the methods are displayed in decreasing order with their respective Friedman ranking.

A.1.2. Parameter Selection

AUC	
input-DNA, nomapReg, GC	3.5
input-DNA, mapReg, GC	3.7813
noinput-DNA, nomapReg, GC	3.9375
input-DNA, nomapReg, noGC	4.4375
noinput-DNA, mapReg, GC	4.4688
input-DNA, mapReg, noGC	4.7813
noinput-DNA, mapReg, noGC	5.1563
noinput-DNA, nomapReg, noGC	5.9375

Table A.2.: The Friedman ranking for the construction of the genomic signal based on TF experiments. For the DAGE statistic we use $h = 50$, $H = 500$. We restrict our analysis to DPs in chromosome 1. We are interested in all 8 combinations of using: (1) the GC-content model (GC, noGC), (2) filtering reads aligned to poor mappability regions (mapReg, nomapReg) and (3) the subtraction of input-DNA (input-DNA, noinput-DNA).

	input-DNA, nomapReg, GC	input-DNA, mapReg, GC	noinput-DNA, nomapReg, GC	input-DNA, nomapReg, noGC	noinput-DNA, mapReg, GC	input-DNA, mapReg, noGC	noinput-DNA, mapReg, noGC	noinput-DNA, nomapReg, noGC
input-DNA, nomapReg, GC								
input-DNA, mapReg, GC								
noinput-DNA, nomapReg, GC								
input-DNA, nomapReg, noGC								
noinput-DNA, mapReg, GC								
input-DNA, mapReg, noGC								
noinput-DNA, mapReg, noGC								
noinput-DNA, nomapReg, noGC	+							

Table A.3.: Friedman-Nemenyi hypothesis test results for the construction of the genomic signal based on TF experiments. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.

AUC	
input-DNA, nomapReg, GC	3.9286
input-DNA, mapReg, GC	4.0
noinput-DNA, mapReg, GC	4.0
noinput-DNA, nomapReg, GC	4.2143
input-DNA, nomapReg, noGC	4.7143
input-DNA, mapReg, noGC	4.7857
noinput-DNA, nomapReg, noGC	4.9286
noinput-DNA, mapReg, noGC	5.4286

Table A.4.: The Friedman ranking for the construction of the genomic signal based on histone modification experiments. For the DAGE statistic we use $h = 50$, $H = 500$. We restrict our analysis to DPs in chromosome 1. We are interested in all 8 combinations of using: (1) the GC-content model (GC, noGC), (2) filtering reads aligned to poor mappability regions (mapReg, nomapReg) and (3) the subtraction of input-DNA (input-DNA, noinput-DNA).

	input-DNA, nomapReg, GC	input-DNA, mapReg, GC	noinput-DNA, mapReg, GC	noinput-NA, nomapReg, GC	input-DNA, nomapReg, noGC	input-DNA, mapReg, noGC	noinput-DNA, nomapReg, noGC	noinput-DNA, mapReg, noGC
input-DNA, nomapReg, GC								
input-DNA, mapReg, GC								
noinput-DNA, mapReg, GC								
noinput-DNA, nomapReg, GC								
input-DNA, nomapReg, noGC								
input-DNA, mapReg, noGC								
noinput-DNA, nomapReg, noGC								
noinput-DNA, mapReg, noGC								

Table A.5.: Friedman-Nemenyi hypothesis test results for the construction of the genomic signal based on histone modification experiments. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.

AUC	
poisson1	2.5
binomial	2.8125
poisson3	3.1563
poisson4	3.1875
poisson2	3.3438

Table A.6.: Results based on TF experiments and emission distributions applied by ODIN. We constrain our analysis on chromosome 1 and use $h = 50$ and $H = 500$ for the DAGE score. Friedman ranking: for each metric, the methods are displayed in decreasing order with their respective Friedman ranking.

	poisson1	binomial	poisson3	poisson4	poisson2
poisson1					
binomial					
poisson3					
poisson4					
poisson2					

Table A.7.: Results based on TF experiments and emission distributions applied by ODIN. Friedman-Nemenyi hypothesis test results for the AUC metric. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.

AUC	
poisson4	2.4286
poisson3	2.5714
poisson2	3.0
poisson1	3.1429
binomial	3.8571

Table A.8.: Results based on histone experiments and emission distributions applied by ODIN. We constrain our analysis on chromosome 1 and use $h = 50$ and $H = 500$ for the DAGE score. Friedman ranking: for each metric, the methods are displayed in decreasing order with their respective Friedman ranking.

	poisson4	poisson3	poisson2	poisson1	binomial
poisson4					
poisson3					
poisson2					
poisson1					
binomial					

Table A.9.: Results based on histone experiments and emission distributions applied by ODIN. Friedman-Nemenyi hypothesis test results for the AUC metric. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.

AUC	
MAnorm-macs	2.375
MAnorm-quest	3.0625
MAnorm-peakseq	3.125
DBChIP-quest	3.8125
DBChIP-macs	4.125
DBChIP-peakseq	4.5

Table A.10.: Results based on TF experiments. We consider different combinations of two-stage DPC and underlying SPC. We constrain our analysis on chromosome 1 and use $h = 50$ and $H = 500$ for the DAGE score. Friedman ranking: for each metric, the methods are displayed in decreasing order with their respective Friedman ranking.

	MAnorm-macs	MAnorm-quest	MAnorm-peakseq	DBChIP-quest	DBChIP-macs	DBChIP-peakseq
MAnorm-macs						
MAnorm-quest						
MAnorm-peakseq						
DBChIP-quest						
DBChIP-macs	+					
DBChIP-peakseq	*					

Table A.11.: Results based on TF experiments. We consider different combinations of two-stage DPC and underlying SPC. Friedman-Nemenyi hypothesis test results for the AUC metric. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.

AUC	
MAnorm-macs	1.3929
MAnorm-quest	2.4286
MAnorm-peakseq	3.1429
DESeq-quest	4.2143
DESeq-macs	4.75
DESeq-peakseq	5.0714

Table A.12.: Results based on histone experiments. We consider different combinations of two-stage DPC and underlying SPC. We constrain our analysis on chromosome 1 and use $h = 50$ and $H = 500$ for the DAGE score. Friedman ranking: for each metric, the methods are displayed in decreasing order with their respective Friedman ranking.

	MAnorm-macs	MAnorm-quest	MAnorm-peakseq	DESeq-quest	DESeq-macs	DESeq-peakseq
MAnorm-macs						
MAnorm-quest						
MAnorm-peakseq						
DESeq-quest	$*$					
DESeq-macs	$*$	$*$				
DESeq-peakseq	$*$	$*$	$+$			

Table A.13.: Results based on histone experiments. We consider different combinations of two-stage DPC and underlying SPC. Friedman-Nemenyi hypothesis test results for the AUC metric. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.

AUC	
ODIN	1.1875
ODIN-deseq	2.3125
ODIN-edgeR	2.5

Table A.14.: Results based on TF experiments. We constrain our analysis on chromosome 1 and use $h = 50$ and $H = 500$ for the DAGE score. Friedman ranking: for each metric, the methods are displayed in decreasing order with their respective Friedman ranking.

AUC	
ODIN	1.2143
ODIN-deseq	1.8571
ODIN-edgeR	2.9286

Table A.15.: *Results based on histone experiments. We constrain our analysis on chromosome 1 and use* $h = 50$ *and* $H = 500$ *for the DAGE score. Friedman ranking: for each metric, the methods are displayed in decreasing order with their respective Friedman ranking.*

A.1.3. DAGE

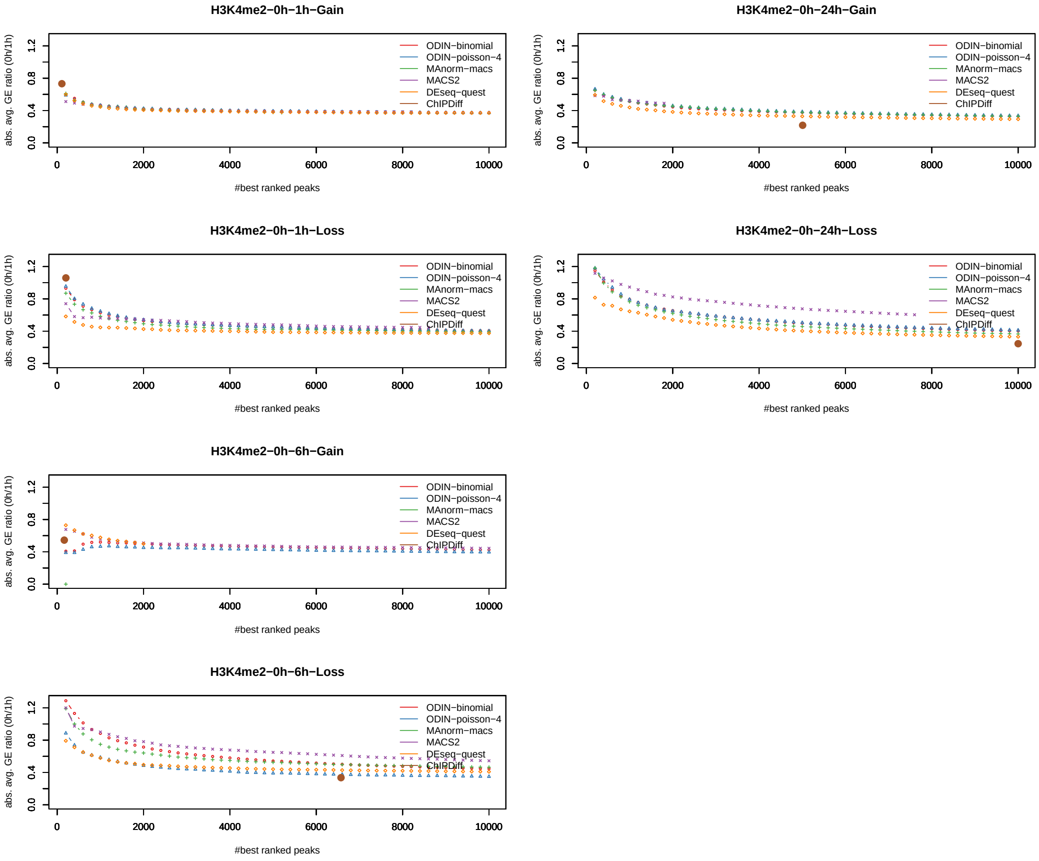


Figure A.1.: DAGE results for histone experiments (based on Kaikkonen et al. (2013))

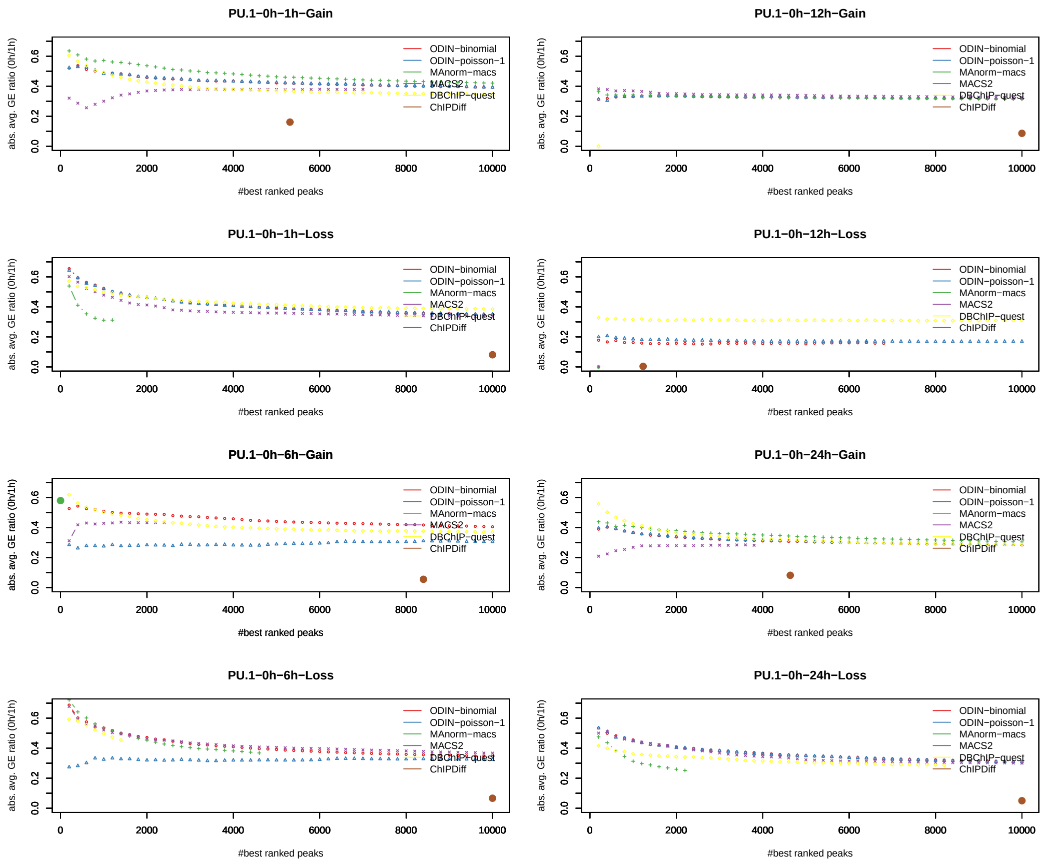
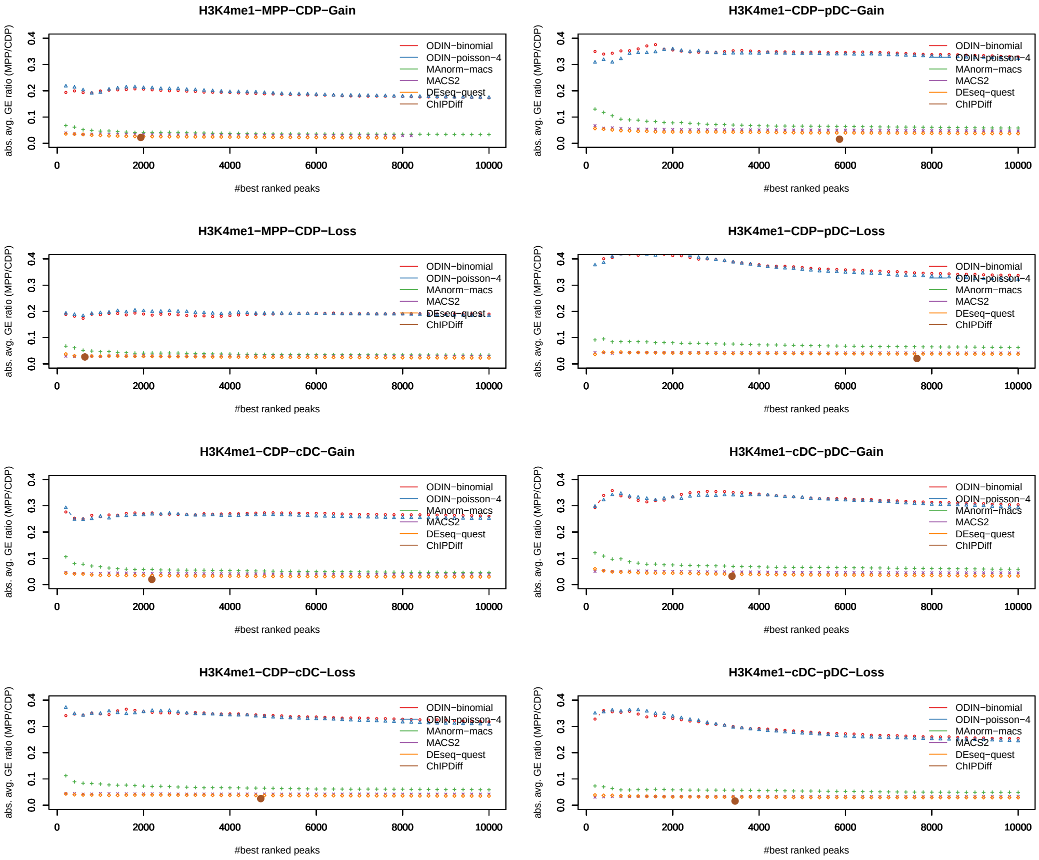


Figure A.2.: DAGE results for TF experiments (based on Kaikkonen et al. (2013))



H3K4me1−MPP−CDP−Gain

H3K4me1−CDP−pDC−Gain

H3K4me1−MPP−CDP−Loss

H3K4me1−CDP−pDC−Loss

H3K4me1−CDP−cDC−Gain

H3K4me1−cDC−pDC−Gain

H3K4me1−CDP−cDC−Loss

H3K4me1−cDC−pDC−Loss

Figure A.3.: *DAGE results for histone experiments (based on Lin et al. (2015)*

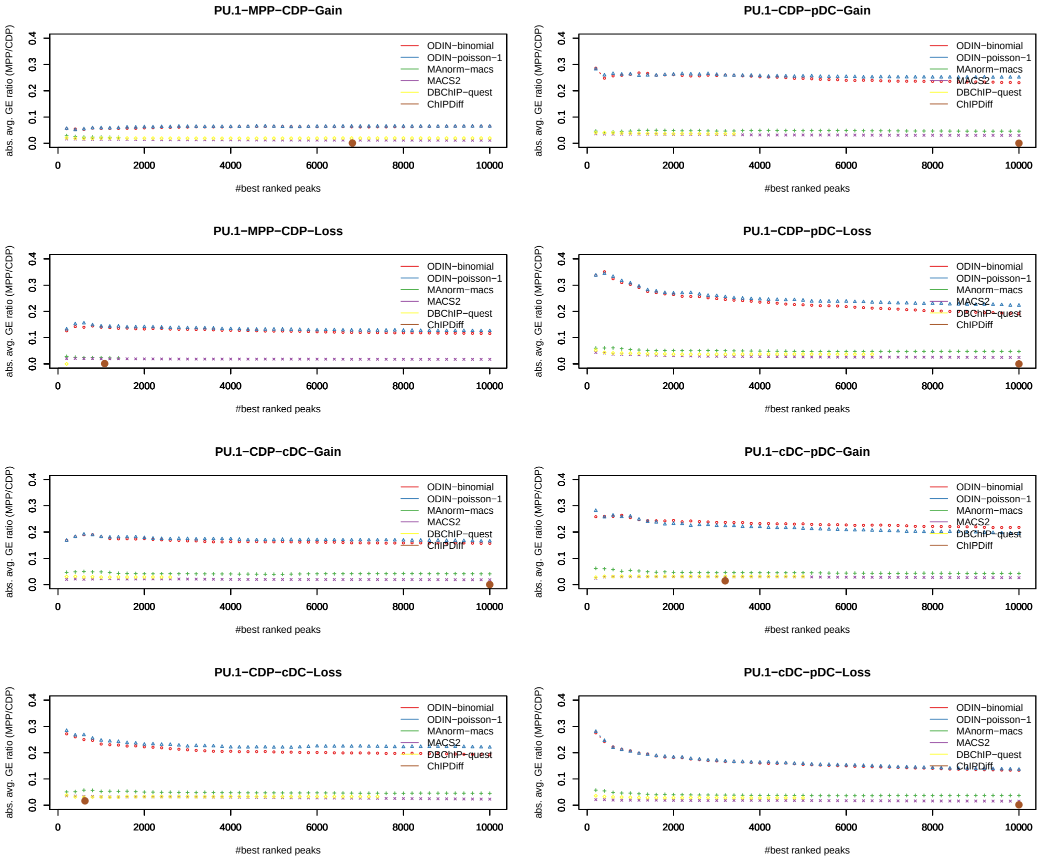


Figure A.4.: DAGE results for TF experiments (based on Lin et al. (2015)

AUC	
ODIN-poisson-1	1.875
ODIN-binomial	2.1875
MAnorm-macs	3.4063
MACS2	3.6563
DBChIP-quest	3.875

Table A.16.: DAGE results based on TF experiments. We use $h = 200$ and $H = 10000$ for the DAGE score. Friedman ranking: for each metric, the methods are displayed in decreasing order with their respective Friedman ranking.

AUC	
ODIN-binomial	1.6429
ODIN-poisson-4	2.0
MAnorm-macs	3.2857
MACS2	3.3571
DEseq-quest	4.7143

Table A.17.: DAGE results based on histone experiments. We use $h = 200$ and $H = 10000$ for the DAGE score. Friedman ranking: for each metric, the methods are displayed in decreasing order with their respective Friedman ranking.

AUC	
ODIN	2.1875
ODIN-poisson-1	2.3125
MACS2	3.75
ChIPDiff	4.0
MAnorm-macs	4.125
DBChIP-quest	4.625

Table A.18.: DAGE results with ChIPDiff based on TF experiments. We use $h = 200$; H equals the number of DPs called by ChIPDiff. Friedman ranking: for each metric, the methods are displayed in decreasing order with their respective Friedman ranking.

	ODIN	ODIN-poisson-1	MACS2	ChIPDiff	MAnorm-macs	DBChIP-quest
ODIN						
ODIN-poisson-1						
MACS2						
ChIPDiff	+					
MAnorm-macs	∗	+				
DBChIP-quest	∗	∗				

Table A.19.: *DAGE results with ChIPDiff based on TF experiments. Friedman-Nemenyi hypothesis test results for the AUC metric. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

AUC	
ODIN-poisson-4	2.5
ODIN	2.5714
MAnorm-macs	3.4286
ChIPDiff	3.6429
MACS2	3.7143
DESeq-quest	5.1429

Table A.20.: *DAGE results with ChIPDiff based on histone experiments. We use $h = 200$; H equals the number of DPs called by ChIPDiff. Friedman ranking: for each metric, the methods are displayed in decreasing order with their respective Friedman ranking.*

	ODIN-poisson-4	ODIN	MAnorm-macs	ChIPDiff	MACS2	DESeq-quest
ODIN-poisson-4						
ODIN						
MAnorm-macs						
ChIPDiff						
MACS2						
DESeq-quest	∗	∗				

Table A.21.: DAGE results with ChIPDiff based on histone experiments. Friedman-Nemenyi hypothesis test results for the AUC metric. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.

A.2 Results with THOR

A.2.1. Simulation

AUC	
THOR	1.0652
MACS2	3.0598
DESeq-JAMM	3.9185
DiffReps	4.087
DESeq-IDR	4.4891
DiffBind	5.0761
Poisson-THOR	6.3043

Table A.22.: Friedman ranking of simulated data for all parameter settings based on the AUC statistic (see main document Section 4.3.3 for details). The methods are displayed in decreasing order with their respective Friedman ranking.

	THOR	DESeq-IDR	MACS2	DiffReps	DiffBind	DESeq-JAMM	Poisson-THOR
THOR							
DESeq-IDR							
MACS2							
DiffReps	*						
DiffBind	*	*	+				
DESeq-JAMM	*	*	*				
Poisson-THOR	*	*	*	*			

Table A.23.: Friedman-Nemenyi hypothesis test results for the AUC metric. We consider the case with 2 replicates, low within-condition variance, and moderate peak size variability. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.

	THOR	DESeq-IDR	MACS2	DiffReps	DiffBind	DESeq-JAMM	Poisson-THOR
THOR							
DESeq-IDR							
MACS2							
DiffReps	∗						
DiffBind	∗	∗	+				
DESeq-JAMM	∗	∗	∗				
Poisson-THOR	∗	∗	∗	∗			

Table A.24.: *Friedman-Nemenyi hypothesis test results for the AUC metric. We consider the case with 2 replicates, medium within-condition variance, and moderate peak size variability. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

	THOR	DESeq-IDR	MACS2	DiffReps	DiffBind	DESeq-JAMM	Poisson-THOR
THOR							
DESeq-IDR							
MACS2							
DiffReps	∗	+					
DiffBind	∗	∗					
DESeq-JAMM	∗	∗	∗				
Poisson-THOR	∗	∗	∗				

Table A.25.: *Friedman-Nemenyi hypothesis test results for the AUC metric. We consider the case with 2 replicates, high within-condition variance, and moderate peak size variability. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

	THOR	DESeq-IDR	MACS2	DiffReps	DESeq-JAMM	Poisson-THOR	DiffBind
THOR							
DESeq-IDR							
MACS2							
DiffReps	∗	+					
DESeq-JAMM	∗	∗					
Poisson-THOR	∗	∗	∗				
DiffBind	∗	∗	∗	∗			

Table A.26.: *Friedman-Nemenyi hypothesis test results for the AUC metric. We consider the case with 2 replicates, low within-condition variance, and high peak size variability. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

	THOR	DESeq-IDR	MACS2	DiffReps	DESeq-JAMM	Poisson-THOR	DiffBind
THOR							
DESeq-IDR							
MACS2							
DiffReps	∗						
DESeq-JAMM	∗	∗	+				
Poisson-THOR	∗	∗	∗				
DiffBind	∗	∗	∗	∗			

Table A.27.: *Friedman-Nemenyi hypothesis test results for the AUC metric. We consider the case with 2 replicates, medium within-condition variance, and high peak size variability. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

	THOR	DESeq-IDR	MACS2	DiffReps	Poisson-THOR	DESeq-JAMM	DiffBind
THOR							
DESeq-IDR							
MACS2	∗						
DiffReps	∗	+					
Poisson-THOR	∗	∗					
DESeq-JAMM	∗	∗					
DiffBind	∗	∗	∗	+			

Table A.28.: *Friedman-Nemenyi hypothesis test results for the AUC metric. We consider the case with 2 replicates, high within-condition variance, and high peak size variability. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

	THOR	DESeq-JAMM	MACS2	DiffReps	DiffBind	Poisson-THOR	DESeq-IDR
THOR							
DESeq-JAMM							
MACS2							
DiffReps	∗						
DiffBind	∗	∗					
Poisson-THOR	∗	∗	∗				
DESeq-IDR	∗	∗	∗	∗			

Table A.29.: *Friedman-Nemenyi hypothesis test results for the AUC metric. We consider the case with 4 replicates, low within-condition variance, and moderate peak size variability. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

	THOR	DESeq-JAMM	MACS2	DiffReps	DiffBind	Poisson-THOR	DESeq-IDR
THOR							
DESeq-JAMM							
MACS2							
DiffReps	∗						
DiffBind	∗	∗					
Poisson-THOR	∗	∗	∗				
DESeq-IDR	∗	∗	∗	∗			

Table A.30.: *Friedman-Nemenyi hypothesis test results for the AUC metric. We consider the case with 4 replicates, medium within-condition variance, and moderate peak size variability. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

	THOR	DESeq-JAMM	MACS2	DiffReps	DiffBind	Poisson-THOR	DESeq-IDR
THOR							
DESeq-JAMM							
MACS2							
DiffReps	∗						
DiffBind	∗	∗					
Poisson-THOR	∗	∗	∗				
DESeq-IDR	∗	∗	∗	∗			

Table A.31.: *Friedman-Nemenyi hypothesis test results for the AUC metric. We consider the case with 4 replicates, high within-condition variance, and moderate peak size variability. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

	THOR	DESeq-JAMM	MACS2	DiffReps	DiffBind	Poisson-THOR	DESeq-IDR
THOR							
DESeq-JAMM							
MACS2							
DiffReps	*						
DiffBind	*	*					
Poisson-THOR	*	*	*				
DESeq-IDR	*	*	*	*			

Table A.32.: *Friedman-Nemenyi hypothesis test results for the AUC metric. We consider the case with 4 replicates, low within-condition variance, and high peak size variability. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

	THOR	DESeq-JAMM	MACS2	DiffReps	DiffBind	Poisson-THOR	DESeq-IDR
THOR							
DESeq-JAMM							
MACS2							
DiffReps	*						
DiffBind	*	*					
Poisson-THOR	*	*	*				
DESeq-IDR	*	*	*	*			

Table A.33.: *Friedman-Nemenyi hypothesis test results for the AUC metric. We consider the case with 4 replicates, medium within-condition variance, and high peak size variability. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

	THOR	DESeq-JAMM	MACS2	DiffReps	DiffBind	Poisson-THOR	DESeq-IDR
THOR							
DESeq-JAMM							
MACS2							
DiffReps	∗						
DiffBind	∗	∗					
Poisson-THOR	∗	∗	∗				
DESeq-IDR	∗	∗	∗	∗			

Table A.34.: *Friedman-Nemenyi hypothesis test results for the AUC metric. We consider the case with 4 replicates, high within-condition variance, and high peak size variability. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

A.2.2. Parameter Selection

AUC	
THOR-1.6/95	2.2857
THOR-1.3/95	2.4286
THOR-1.6/99	2.5
THOR-1.3/99	2.7857

Table A.35.: *Friedman ranking based on DCA score ($h = 100, H = 1000$). We evaluate the initial parameter setting of THOR, that is, $t_1 \in \{\langle x \rangle^{.95}, \langle x \rangle^{.99}\}$ and $t_2 \in \{1.3, 1.6\}$ where t_1 is the fold change criteria and t_2 the minimum difference between signals based on percentile estimates (see main document Section 4.3.4 for details). The analysis is restricted to chromosome 1. For each metric, the methods are displayed in decreasing order with their respective Friedman ranking.*

	THOR-1.6/95	THOR-1.3/95	THOR-1.6/99	THOR-1.3/99
THOR-1.6/95				
THOR-1.3/95				
THOR-1.6/99				
THOR-1.3/99				

Table A.36.: *Friedman-Nemenyi hypothesis test results for the DCA score ($h = 100, H = 1000$) restricted to chromosome 1. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

A.2.3. DCA

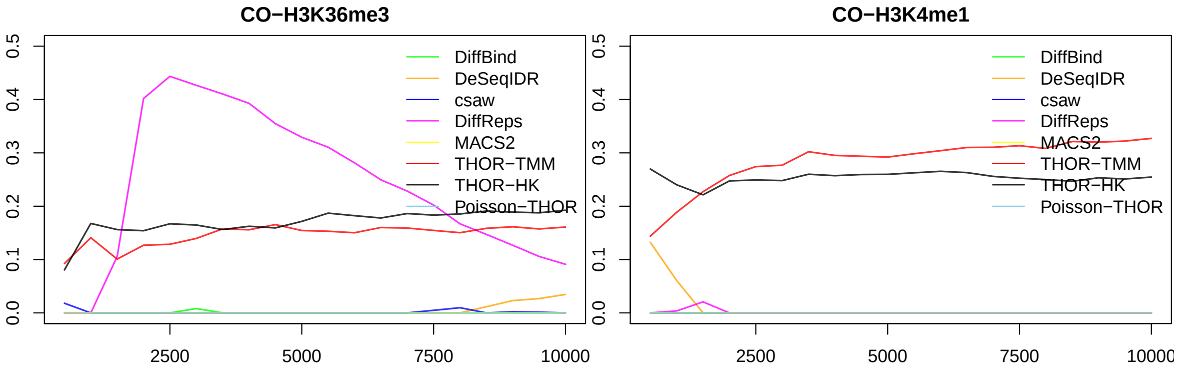


Figure A.5.: DCA curves for CO study. We run THOR with TMM and housekeeping genes normalization approach. PePr required input-DNA and is therefore unable to call DPs.

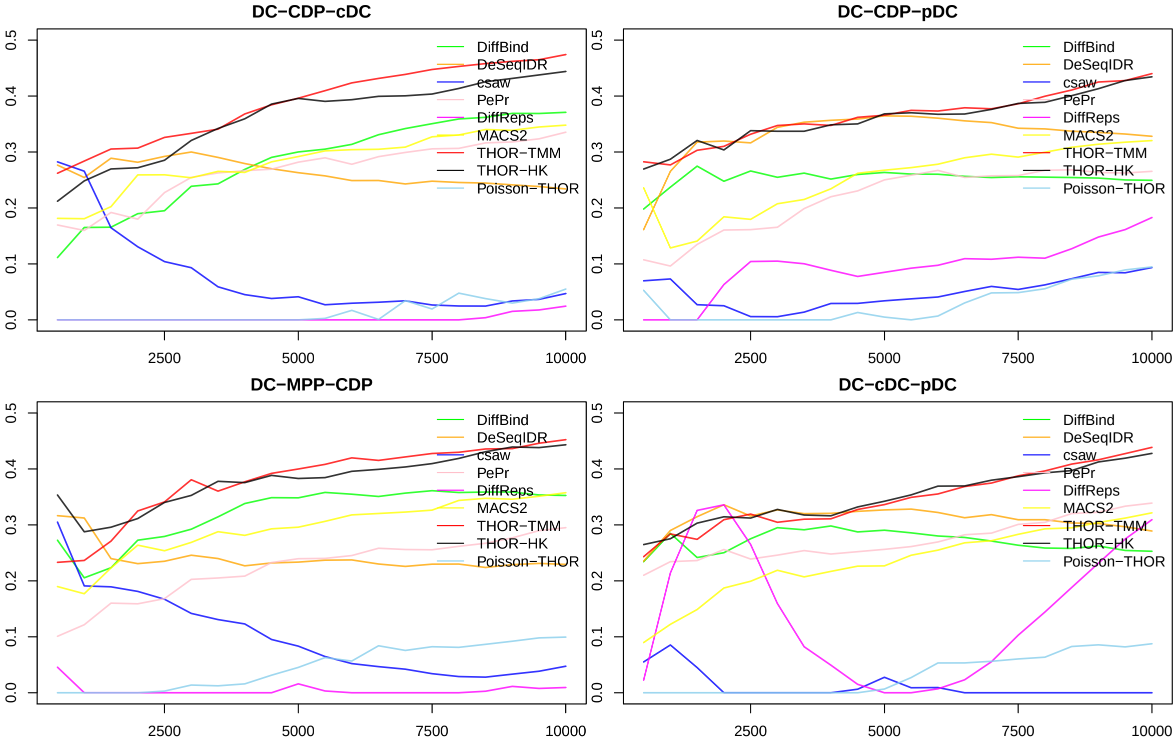


Figure A.6.: DCA curves for DC study. We run THOR with TMM and housekeeping genes normalization approach.

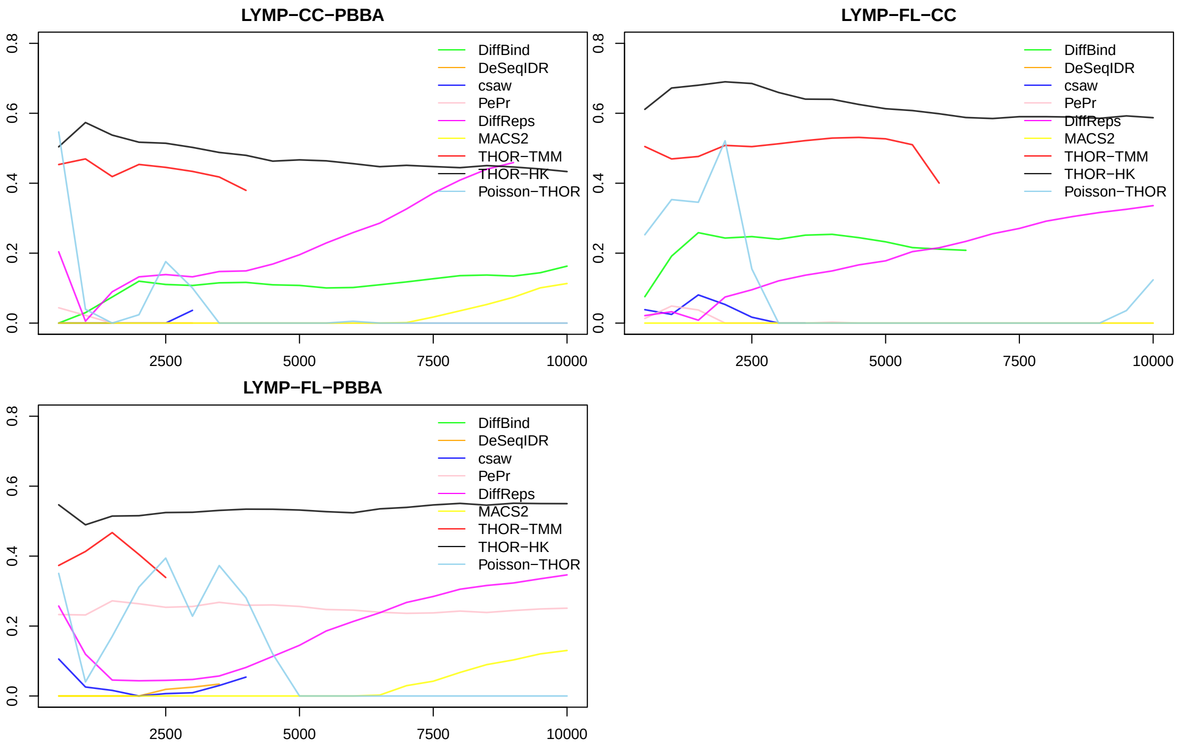


Figure A.7.: DCA curves for LYMP study. We run THOR with TMM and housekeeping genes normalization approach.

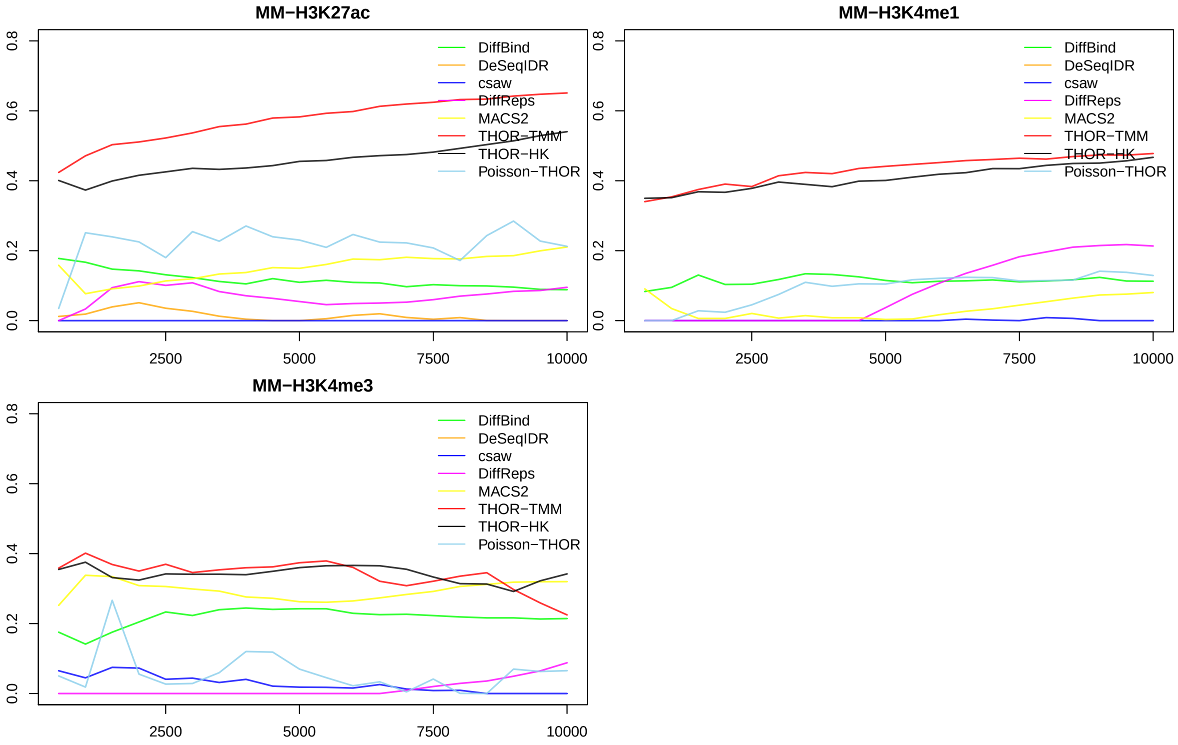


Figure A.8.: DCA curves for MM study. We run THOR with TMM and housekeeping genes normalization approach. PePr required input-DNA and is therefore unable to call DPs.

AUC	
THOR-HK	2.0
THOR-TMM	2.4286
macs2	3.7857
DiffReps	4.4286
DiffBind	5.2143
DESeqIDR	5.6429
Poisson-THOR	5.7143
csaw	6.7857

Table A.37.: Friedman ranking based on DCA score ($h = 500, H = 10000$) for all datasets (CO, DC, LYMP and MM). The methods are displayed in decreasing order with their respective Friedman ranking.

AUC	
THOR-HK	2.3333
THOR-TMM	2.7778
macs2	4.3333
PePr	4.6667
DiffReps	5.0
DiffBind	5.7778
DESeqIDR	6.0
Poisson-THOR	6.3333
csaw	7.7778

Table A.38.: Friedman ranking based on DCA score ($h = 500, H = 10000$) for datasets DC and LYMP. We restrict the analysis to DC and LYMP as PePr requires input-DNA which is not provided by CO and MM. The methods are displayed in decreasing order with their respective Friedman ranking.

	THOR-HK	THOR-TMM	macs2	PePr	DiffReps	DiffBind	DESeqIDR	Poisson-THOR	csaw
THOR-HK									
THOR-TMM									
macs2									
PePr									
DiffReps									
DiffBind									
DESeqIDR									
Poisson-THOR	+								
csaw	*	*							

Table A.39.: *Friedman-Nemenyi hypothesis test results for the DCA score ($h = 500, H = 10000$). The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1.*

A.2.4. Use Cases of THOR

rank	chrom	pos	dbSNP rID	$-\log_{10}$ p-value	dir	Gene upstream	Gene downstream	Figure 5.11?
1	chr3	16553883	rs2346911	99.8922475655	-	RFTN1 (+1330)	OXNAD1 (+247178)	
2	chr1	192776414	rs1418718	77.6115468364	-	RGS2 (-1757)	RGS13 (+171140)	X
3	chr9	127562788	rs750691	73.6809538923	+	OLFML2A (+23239)	RPL35 (+61452)	
4	chr9	127562973	rs913232	73.6809538923	+	OLFML2A (+23424)	RPL35 (+61267)	
5	chr11	71752160	rs7115200	69.3486017723	+	LRTOMT (-39222)	IL18BP (+42052)	X
6	chr1	192577986	rs4130930	65.9782743138	-	RGS13 (-27289)	RGS1 (+33130)	X
7	chr1	192578763	rs7538087	65.9782743138	-	RGS13 (-26512)	RGS1 (+33907)	X
8	chr17	74524652	rs8077736	53.8122623004	+	RHBDF2 (-27164)	CYGB (+9335)	
9	chr10	6391031	rs12416248	49.3236112959	-	PFKFB3 (+146138)	PRKCQ (+231232)	
10	chr11	65197393	rs674485	43.6775100799	-	SCYL1 (-95155)	FRMD8 (+43324)	X
11	chr21	45615896	rs2838520	36.4901273214	-	ICOSLG (+44932)	C21orf33 (+62410)	
12	chr21	45615917	rs2838521	36.4901273214	-	ICOSLG (+44911)	C21orf33 (+62431)	
13	chr8	27247339	rs34947559	26.0782531932	+	PTK2B (+78341)	CHRNA2 (+89474)	X
14	chr17	41400290	NA	25.3212748249	+	ARL4D (-76037)	TMEM106A (+36397)	
15	chr17	41400913	NA	25.3212748249	+	ARL4D (-75414)	TMEM106A (+37020)	
16	chr11	128496565	rs949097	24.244683318	+	FLI1 (-67325)	ETS1 (-39129)	
17	chr17	56709222	rs444393	24.1492427636	-	SEPT4 (-91044)	TEX14 (+60162)	
18	chr1	182558137	rs10911102	21.5391751611	+	RNASEL (+254)	RGSL1 (+138857)	
19	chr6	167527097	rs6909252	20.9400452035	-	CCR6 (-9160)	FGFR1OP (+114428)	X
20	chr1	147806874	rs2999607	20.6529101183	+	NBPF24 (-207326)	PPIAL4A (+148545)	
21	chr1	147807277	rs481176	20.6529101183	+	NBPF24 (-207729)	PPIAL4A (+148142)	
22	chr17	6659146	rs955462	20.4274431857	+	SLC13A5 (-42261)	XAF1 (-13)	
23	chr1	151031667	rs3806386	18.8647292719	+	MLLT11 (+1434)	CDC42SE1 (+11134)	
24	chr3	115377254	rs13100660	18.168913574	+	GAP43 (+34898)	LSAMP (+787124)	
25	chr9	134144806	rs7861111	17.4971632608	+	PPAPDC3 (-20275)	NUP214 (+143859)	
26	chr16	56946804	rs711746	17.3017972487	+	HERPUD1 (-19156)	SLC12A3 (+47686)	
27	chr17	45213047	NA	17.1375858395	-	RPRML (-156434)	CDC27 (+53495)	
28	chrX	15693367	rs4830979	17.0962250807	+	CA5B (-63026)	TMEM27 (-10214)	
29	chrX	15693461	rs4830980	17.0962250807	+	CA5B (-62932)	TMEM27 (-10308)	
30	chr3	56591508	rs73079894	15.0492637401	+	CCDC66 (+308)	ARHGEF3 (+521828)	
31	chr18	46549675	rs4939571	14.2426089689	+	SMAD7 (-72595)	DYM (+437497)	
32	chr6	88182439	rs2273129	14.2079766474	+	SLC35A1 (-256)	C6orf163 (+127869)	
33	chr2	44588941	rs698775	14.0871764676	+	PREPL (-309)	CAMKMT (-162)	
34	chr20	56056342	rs1001752	14.0111470481	-	CTCFL (+43821)	RBM38 (+89880)	
35	chr7	1979750	rs10950456	13.7186584489	+	ELFN1 (+251996)	MAD1L1 (+293128)	
36	chr17	67323781	rs333938	13.3851108104	+	MAP2K6 (-87058)	ABCA5 (-540)	
37	chr7	22862192	rs2270106	13.313357982	+	TOMM7 (+278)	IL6 (+96690)	
38	chr22	48494758	rs5768350	12.1807933237	+	FAM19A5 (-390514)		
39	chr1	43418026	rs2297972	11.1639466495	+	SLC2A1 (+6475)	ZNF691 (+105720)	
40	chr4	64378	NA	11.0619722048	+	ZNF595 (+11169)	ZNF732 (+234732)	
41	chr22	46984098	rs1883193	10.9788053627	+	CELSR1 (-51032)	GRAMD4 (-32201)	
42	chr22	46984100	rs1883192	10.9788053627	+	CELSR1 (-51034)	GRAMD4 (-32199)	
43	chr22	46984268	rs909558	10.9788053627	+	CELSR1 (-51202)	GRAMD4 (-32031)	
44	chr5	149793457	rs1560661	10.6980883333	+	CD74 (-1144)	RPS14 (+35853)	
45	chr6	32634104	NA	10.4960726902	+	HLA-DQB1 (+357)	HLA-DQA1 (+28971)	
46	chr3	28390351	rs1870259	10.4450508705	+	AZI2 (+267)	CMC1 (+107266)	
47	chr8	47829990	rs13259304	10.441488493	+	SPIDR (-343177)		
48	chr8	47829991	rs13259305	10.441488493	+	SPIDR (-343176)		
49	chr17	70025931	rs2193053	9.895586428	+	SOX9 (-91230)		
50	chr22	22400882	rs4145408	9.3687977404	-	VPREB1 (-198205)	TOP3B (-63736)	
51	chr22	24142330	rs738795	9.3254710848	-	SMARCB1 (+13170)	DERL3 (+38863)	
52	chr17	33905468	rs321600	9.2026740212	+	SLFN14 (-20352)	PEX12 (+180)	
53	chr15	52528193	rs6493549	8.9018777619	+	GNB5 (-44628)	MYO5C (+59802)	
54	chr9	126101008	rs10114139	8.8987861442	+	STRBP (-70154)	CRB2 (-17531)	
55	chr7	45025720	rs3213658	8.279020729	-	CCM2 (-40905)	MYO1G (-7024)	
56	chr9	137029841	rs28650068	8.1713392696	-	RXRA (-188585)	WDR5 (+28632)	
57	chr1	32355180	rs593133	8.0917910077	+	SPOCD1 (-73529)	PTP4A2 (+48808)	
58	chr12	6570966	rs1045548	7.5019308863	+	VAMP1 (+8877)	TAPBPL (+9717)	
59	chr5	156700461	rs62383003	7.2378390446	+	CYFIP2 (+7324)	FNDC9 (+72268)	
60	chr5	130588550	rs6596007	7.0902630177	+	CDC42SE2 (-11243)	LYRM7 (+82048)	
61	chr7	2750918	rs10252130	6.8566385722	+	AMZ1 (+31763)	GNA12 (+133040)	
62	chr5	43313178	rs10039048	6.7651996991	+	HMGCS1 (+417)	ENSG00000177453 (+120225)	
63	chr1	172412995	rs3213563	6.7522493823	+	PIGC (+231)	DNM3 (+602375)	
64	chr5	163342658	rs13184669	6.5468860152	+	MAT2B (+410105)		
65	chr5	163343803	rs12516138	6.5468860152	+	MAT2B (+411250)		
66	chr1	205601464	rs3088136	6.2865958958	+	ELK4 (-375)	SLC45A3 (+48123)	
67	chr1	242011406	rs1776179	6.1953416871	+	OPN3 (-207744)	EXO1 (-76)	
68	chr10	126289743	rs2104227	6.0646970367	+	LHPP (+139340)	FAM53B (+142876)	
69	chr6	116600774	rs3749895	6.0268617678	+	TSPYL4 (-25514)	TSPYL1 (+292)	
70	chr3	14473000	rs7620731	5.9079302255	+	C3orf20 (-243606)	SLC6A6 (+28925)	
71	chr8	135613597	rs894346	5.8709599503	+	ZFAT (+111684)		
72	chr8	135613624	rs894347	5.8709599503	+	ZFAT (+111657)		
73	chr2	109065858	rs2460947	5.8340918213	+	LIMS1 (-205635)	GCC2 (+842)	
74	chr11	1873950	rs2089908	5.7990375522	+	LSP1 (-12445)	TNNI2 (+13240)	
75	chr3	188108642	rs56046601	5.7612327644	-	TPRG1 (-781121)	LPP (+177922)	
76	chr5	40835088	rs2270625	5.6276004169	+	PRKAA1 (-36613)	RPL37 (+349)	
77	chr1	179051300	rs2296377	5.3746650821	-	TOR3A (+789)	ABL2 (+147436)	
78	chr9	6704188	rs820495	5.133890915	+	GLDC (-58539)	KDM4C (-53468)	
79	chr9	6704237	rs820494	5.133890915	+	GLDC (-58588)	KDM4C (-53419)	
80	chr2	225867366	rs281527	5.0464335748	+	CUL3 (-417302)	DOCK10 (+39793)	

Table A.40.: *List of candidate rSNPs ranked by the negative logarithm of the p-value of the DPs called by THOR. For each SNP we give the rank, the chromosome, the position, the dbSNP rID, the negative logarithm of the DP called by THOR the SNP lies within and the genes that lay in close vicinity. We also indicate whether the rSNP is pictured in Figure 5.11. We list the top* 80/137 *ranked rSNPs.*

rank	gene	chrom	start	end	strand	-log$_{10}$ p-value
1	**H2-AA**	chr17	34419688	34424772	-	9342.50545168
2	**H2-AB1**	chr17	34400153	34406363	+	8018.77571673
3	**H2-EB1**	chr17	34442821	34453144	+	4555.68454975
4	**ID2**	chr12	25778665	25780957	-	3379.91074335
5	**H2-EA-PS**	chr17	34342933	34344643	-	2380.2150288
6	**H2-EB2**	chr17	34462609	34477174	+	2380.2150288
7	KLRK1	chr6	129560340	129573882	-	2200.32296143
8	**CD74**	chr18	60963501	60972300	+	2101.33787491
9	IFI202B	chr1	175892699	175912975	-	2080.14008533
10	OLFR433	chr1	175972063	175973067	+	2080.14008533
11	MARCKS	chr10	36853180	36858726	-	1815.84060863
12	CCND1	chr7	152115835	152125774	-	1591.72987841
13	**ADAM19**	chr11	45869493	45960845	+	1559.08891211
14	**IRF8**	chr8	123260257	123280594	+	1530.25808614
15	CST3	chr2	148697457	148701428	-	1458.44354685
16	**H2-DMB1**	chr17	34290016	34297175	+	1412.96685804
17	**H2-DMB2**	chr17	34280251	34288498	+	1412.96685804
18	TNKS	chr8	35892232	36028744	-	1380.13798257
19	NF1	chr11	79153194	79395114	+	1372.2157044
20	CD83	chr13	43880475	43898499	+	1291.08471599
21	AMZ1	chr5	141200080	141237393	+	1286.78303837
22	P2RY10	chrX	104283830	104300313	+	1163.99255182
23	HERPUD1	chr8	96910337	96919277	+	1056.72241201
24	SLC12A3	chr8	96853091	96890113	+	1056.72241201
25	BC051142	chr17	34535764	34597679	+	1049.4840957
26	EGR3	chr14	70477251	70479964	+	1039.28587015
27	AHRR	chr13	74348565	74429779	-	1017.25699529
28	**CIITA**	chr16	10488270	10527657	+	1006.89398006
29	PRKAR2A	chr9	108594473	108651843	+	975.004304756
30	PMEPA1	chr2	173049958	173102034	-	968.09089682
31	HIAT1	chr3	116334081	116384178	-	946.531575531
32	A330009N23Rik	chr15	101055056	101055069	-	929.20633982
33	GRASP	chr15	101054637	101063186	+	929.20633982
34	HRH1	chr6	114347929	114433290	+	919.54865411
35	ZFP800	chr6	28189930	28348005	-	913.228573003
36	GRAMD3	chr18	56591785	56663446	+	895.274922571
37	**H2-OB**	chr17	34375847	34382852	+	855.004458933
38	WDR86	chr5	24217555	24236545	-	845.047427982
39	P2RX5	chr11	72973922	72986187	+	844.537705199
40	NCOA7	chr10	30365389	30522913	-	843.985543271
41	ACTB	chr5	143664793	143668433	-	840.4456036
42	COL25A1	chr3	129883808	130302795	+	831.737274956
43	5830416P10RIK	chr19	53526024	53526024	-	826.028290432
44	SMNDC1	chr19	53453703	53465063	-	826.028290432
45	SLFN5	chr11	82764850	82776443	+	820.070737402
44	GM8817	chrX	163526888	163553394	-	817.998789877
47	**KIT**	chr5	75970940	76052747	+	794.821851563
48	DIS3L2	chr1	88600382	88946670	+	791.794724159
49	FAM46A	chr9	85214045	85220955	-	788.360435944
50	ANKRD55	chr13	113078658	113174210	+	784.448973727

Table A.41.: *List of genes in DC-CDP-cDC (cDC peaks) that are associated with DPs called by THOR. We rank the genes by the p-value of the assigned DP. For each gene, we give the chromosome, the start and end positions, the strand as well as the p-value the gene is assigned to. Genes that are highlighted in bold are specifically known to be associated with dendritic cells and in particular cDC cells.*

rank	gene	chrom	start	end	strand	-log$_{10}$ p-value
1	**SIGLECH**	chr7	63023547	63034295	+	9678.16863265
2	**IRF8**	chr8	123260257	123280594	+	9397.95545021
3	PECAM1	chr11	106515530	106611942	-	7721.79239662
4	TEX2	chr11	106363460	106474737	-	7721.79239662
5	PSAP	chr10	59740374	59765345	+	6767.12351276
6	MCTP2	chr7	79222715	79451481	-	6471.46286813
7	EGFR	chr11	16652205	16818161	+	6031.83979238
8	FBXO48	chr11	16851377	16854775	+	6031.83979238
9	**IFNAR1**	chr16	91485482	91507686	+	5777.69157412
10	**IL10RB**	chr16	91406408	91426079	+	5777.69157412
11	LDLRAD3	chr2	101790359	102026542	-	5257.7739313
12	SEMA4B	chr7	87331726	87371280	+	4773.62970215
13	ST8SIA4	chr1	97484258	97564148	-	4752.49374499
14	PPM1H	chr10	122115817	122382851	+	4708.84903533
15	PRKAG2	chr5	24368561	24606460	-	4318.68020523
16	OLFR164	chr16	19285835	19286930	-	4135.75145138
17	CDK20	chr13	64533860	64541028	+	4015.81891814
18	CTSL	chr13	64464521	64471614	-	4015.81891814
19	MTAP7D1	chr4	125933470	125933614	-	3892.15227629
20	STAMBPL1	chr19	34266718	34314823	+	3697.16215076
21	ATP1B1	chr1	166367397	166388486	-	3647.33666707
22	CYBASC3	chr19	10651929	10651951	+	3619.76719482
23	TMEM138	chr19	10644967	10651852	-	3619.76719482
24	ARL5C	chr11	97850891	97857495	-	3528.20546387
25	EPHA2	chr4	140857154	140885299	+	3428.42485528
26	MED16	chr10	79357452	79371668	-	3330.06424981
27	TMEM229B	chr12	80062781	80108614	-	3321.71551074
28	RPGRIP1	chr14	52730378	52783221	+	3282.52005794
29	CMAH	chr13	24419288	24569154	+	3276.26687301
30	LRP8	chr4	107474865	107549445	+	3227.01828285
31	RPL31	chr1	39424695	39428753	+	3154.91756008
32	TBC1D8	chr1	39428343	39535592	-	3154.91756008
33	KLHDC4	chr8	124320212	124353469	-	3036.19329771
34	SLC7A5	chr8	124405049	124431594	-	3036.19329771
35	HPSE2	chr19	42863436	43462801	-	3020.84986442
36	**PACSIN1**	chr17	27792453	27848051	+	2978.26587504
37	CCDC162	chr10	41258651	41429106	-	2881.08000048
38	BCR	chr10	74523640	74647668	+	2861.03007096
39	LY6E	chr15	74785480	74790335	+	2850.91050473
40	MED12L	chr3	58810899	59122332	+	2834.58051421
41	**TCF4**	chr18	69503799	69847621	+	2698.42429934
42	DGAT2	chr7	106302172	106331223	-	2670.94433911
43	UVRAG	chr7	106035252	106289654	-	2670.94433911
44	3300005D01RIK	chr17	5803242	5803242	+	2661.34040582
45	SNX9	chr17	5841327	5931033	+	2661.34040582
46	PMEPA1	chr2	173049958	173102034	-	2627.43181036
47	CD4	chr6	124814709	124838239	-	2614.28892241
48	LAG3	chr6	124854378	124861723	-	2614.28892241
49	**RUNX2**	chr17	44632935	44951746	-	2580.8334652
50	CD33	chr7	50782825	50788541	-	2563.80058712

Table A.42.: List of genes in DC-CDP-pDC (pDC peaks) that are associated with DPs called by THOR. We rank the genes by the p-value of the assigned DP. For each gene, we give the chromosome, the start and end positions, the strand as well as the p-value the gene is assigned to. Genes that are highlighted in bold are specifically known to be associated with dendritic cells and in particular pDC cells.

Bibliography

L. G. Acevedo, L. A. Iniguez, H. L. Holster, X. Zhang, R. Green, and P. J. Farnham. Genome-scale chip-chip analysis using 10,000 human cells. *Biotechniques*, 43(6):791, 2007.

M. Adli and B. E. Bernstein. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nature Protocols*, 6(10):1656–1668, 2011.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 4th edition, 2002.

M. Allhoff, A. Schönhuth, M. Martin, I. G. Costa, S. Rahmann, and T. Marschall. Discovering motifs that induce sequencing errors. *BMC Bioinformatics*, 14(Suppl 5):S1, 2013.

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

C. D. Allis, T. Jenuwein, D. Reinberg, and M.-L. Caparros. *Epigenetics*. Cold Spring Harbor Laboratory Press, 2007.

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

H. Ashoor, A. Hérault, A. Kamoun, F. Radvanyi, V. B. Bajic, E. Barillot, and V. Boeva. HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. *Bioinformatics*, 29(23):2979–2986, 2013.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Y. Benjamini and T. P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):e72, 2012.

J. A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4(510):126, 1998.

S. D. Briggs, M. Bryk, B. D. Strahl, W. L. Cheung, J. K. Davie, S. Y. Dent, F. Winston, and C. D. Allis. Histone H3 lysine 4 methylation is mediated by Set1 and required for cell growth and rDNA silencing in Saccharomyces cerevisiae. *Genes & Development*, 15(24):3286–3295, 2001.

A. C. Cameron and P. K. Trivedi. Essentials of count data regression. *A companion to theoretical econometrics*, 331, 2001.

R. Cao, L. Wang, H. Wang, L. Xia, H. Erdjument-Bromage, P. Tempst, R. S. Jones, and Y. Zhang. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science*, 298(5595):1039–1043, 2002.

Y. Chen, N. Negre, Q. Li, J. O. Mieczkowska, M. Slattery, T. Liu, Y. Zhang, T.-K. Kim, H. H. He, J. Zieba, et al. Systematic evaluation of factors influencing chip-seq fidelity. *Nature Methods*, 6(6):609–614, 2012.

C. Cheng, K.-K. K. Yan, K. Y. Yip, J. Rozowsky, R. Alexander, C. Shou, and M. Gerstein. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biology*, 12(2):R15, 2011.

B. Cisse, M. L. Caton, M. Lehner, T. Maeda, S. Scheu, R. Locksley, D. Holmberg, C. Zweier, N. S. den Hollander, and S. G. Kant. Transcription factor E2-2 is an essential and specific regulator of plasmacytoid dendritic cell development. *Cell*, 135(1):37–48, 2008.

P. Collas. The current state of chromatin immunoprecipitation. *Molecular biotechnology*, 45 (1):87–100, 2010.

C. Couvreur. Hidden Markov Models and Their Mixtures. Master's thesis, Université Catholique de Louvaine Faculté des Sciences - Département de Mathématiques, Leven, Belgium, 1996.

M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, and P. A. Sharp. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936, 2010.

F. Cunningham, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, et al. Ensembl 2015. *Nucleic Acids Research*, 43(D1): D662–D669, 2015.

M. De Andrea, R. Ravera, D. Gioia, M. Gariglio, and S. Landolfo. The interferon system: an overview. *European Journal of Paediatric Neurology*, 6:A41–46, 2002.

J. Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

A. Diaz, K. Park, D. A. Lim, and J. S. Song. Normalization, bias correction, and peak calling for ChIP-seq. *Statistical Applications in Genetics and Molecular Biology*, 11(3), 2012.

R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

E. Eisenberg and E. Y. Levanon. Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574, 2013.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

T. Fawcett. ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, 31:1–38, 2004.

P. Felker, K. Seré, Q. Lin, C. Becker, M. Hristov, T. Hieronymus, and M. Zenke. TGF-beta1 accelerates dendritic cell differentiation from common dendritic cell progenitors and directs subset specification toward conventional dendritic cells. *The Journal of Immunology*, 185(9):5326–5335, 2010.

J. Feng, M. Wilkinson, X. Liu, I. Purushothaman, D. Ferguson, V. Vialou, I. Maze, N. Shao, P. Kennedy, J. Koo, C. Dias, B. Laitman, V. Stockman, Q. LaPlant, M. Cahill, E. Nestler, and L. Shen. Chronic cocaine-regulated epigenomic changes in mouse nucleus accumbens. *Genome Biology*, 15(4):R65, 2014.

M. Friedman. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

T. S. Furey. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics*, 13(12):840–852, 2012.

P. A. Grant, A. Eberharter, S. John, R. G. Cook, B. M. Turner, and J. L. Workman. Expanded lysine acetylation specificity of Gcn5 in native complexes. *Journal of Biological Chemistry*, 274(9):5895–5900, 1999.

R. Guinamard, M. Okigaki, J. Schlessinger, and J. V. Ravetch. Absence of marginal zone B cells in Pyk-2-deficient mice defines their role in the humoral response. *Nature Immunology*, 1(1):31–36, 2000.

L. Guo, Y. Du, S. Chang, K. Zhang, and J. Wang. rSNPBase: a database for curated regulatory SNPs. *Nucleic Acids Research*, 42(D1):D1033–1039, 2014.

T. Gutschner, M. Hämmerle, M. Eißmann, J. Hsu, Y. Kim, G. Hung, A. Revenko, G. Arun, M. Stentrup, M. Groß, et al. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Research*, 73(3):1180–1189, 2013.

S.-B. Han, C. Moratz, N.-N. Huang, B. Kelsall, H. Cho, C.-S. Shi, O. Schwartz, and J. H. Kehrl. Rgs1 and Gnai2 regulate the entrance of B lymphocytes into lymph nodes and B cell motility within lymph node follicles. *Immunity*, 22(3):343–354, 2005.

R. D. Hawkins, G. C. Hon, and B. Ren. Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, 11(7):476–486, 2010.

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

P. Humburg. *ChIPsim: Simulation of ChIP-seq experiments*, 2011. R package version 1.18.0.

M. M. Ibrahim, S. A. Lacadie, and U. Ohler. JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics*, 31(1):48–55, 2015.

N. Ismail and A. A. Jemain. Handling overdispersion with negative binomial and generalized poisson regression models. *Casualty Actuarial Society Forum*, pages 103–158, 2007.

J. T. Jackson, Y. Hu, R. Liu, F. Masson, A. D'Amico, S. Carotta, A. Xin, M. J. Camilleri, A. M. Mount, A. Kallies, et al. Id2 expression delineates differential checkpoints in the genetic program of CD8α+ and CD103+ dendritic cell lineages. *The EMBO Journal*, 30(13):2690–2704, 2011.

V. Jackson. Studies on histone organization in the nucleosome using formaldehyde as a reversible cross-linking agent. *Cell*, 15(3):945–954, 1978.

H. Jaiswal, M. Kaushik, R. Sougrat, M. Gupta, A. Dey, R. Verma, K. Ozato, and P. Tailor. Batf3 and Id2 have a synergistic effect on Irf8-directed classical CD8α+ dendritic cell development. *The Journal of Immunology*, 191(12):5993–6001, 2013.

D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, 2007.

M. U. Kaikkonen, N. J. Spann, S. Heinz, C. E. Romanoski, K. A. Allison, J. D. Stender, H. B. Chun, D. F. Tough, R. K. Prinjha, C. Benner, and C. K. Glass. Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription. *Molecular Cell*, 51(3):310–325, 2013.

R. Karlić, H.-R. Chung, J. Lasserre, K. Vlahoviček, and M. Vingron. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, 107(7):2926–2931, 2010.

P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology*, 26(12):1351–1359, 2008.

E. E. Khrameeva and M. S. Gelfand. Biases in read coverage demonstrated by interlaboratory and interplatform comparison of 117 mRNA and genome sequencing experiments. *BMC Bioinformatics*, 13(Suppl 6):S4, 2012.

O. I. Koues, R. A. Kowalewski, L.-W. Chang, S. C. Pyfrom, J. A. Schmidt, H. Luo, L. E. Sandoval, T. B. Hughes, J. J. Bednarski, A. F. Cashen, J. E. Payton, and E. M. Oltz. Enhancer Sequence Variants and Transcription-Factor Deregulation Synergize to Construct Pathogenic Regulatory Circuits in B-Cell Lymphoma. *Immunity*, 42(1):186–198, 2015.

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

I. Lappalainen, J. Almeida-King, V. Kumanduri, A. Senf, J. D. Spalding, S. ur Rehman, G. Saunders, J. Kandasamy, M. Caccamo, R. Leinonen, et al. The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics*, 47 (7):692–695, 2015.

K. Levenberg. A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, 2:164–168, 1944.

H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

Q. Li, J. B. Brown, H. Huang, and P. J. Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, 2011.

K. Liang and S. Keleş. Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics*, 28(1):121–122, 2012.

Q. Lin, H. Chauvistré, I. G. Costa, E. G. Gusmao, S. Mitzka, S. Hänzelmann, B. Baying, T. Klisch, R. Moriggl, B. Hennuy, H. Smeets, K. Hoffmann, V. Benes, K. Seré, and M. Zenke. Epigenetic program and transcription factor circuitry of dendritic cell development. *Nucleic Acids Research*, 43(20):9680–9693, 2015.

H. Lodish, A. Berk, C. A. Kaiser, M. Krieger, M. P. Scott, A. Bretscher, H. Ploegh, and P. Matsudaira. *Molecular Cell Biology*. W. H. Freeman, 6th edition, 2007.

D. Loughrey, K. E. Watters, A. H. Settle, and J. B. Lucks. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Research*, 42(21):000, 2014.

J. B. Lucks, S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, and A. P. Arkin. Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences*, 108(27):11063–11068, 2011.

A. T. Lun and G. K. Smyth. 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, 2014.

A. Mammana, M. Vingron, and H.-R. Chung. Inferring nucleosome positions with their histone mark annotation from chip data. *Bioinformatics*, 29(20):2547–2554, 2013.

T. Marschall, I. G. Costa, S. Canzar, M. Bauer, G. W. Klau, A. Schliep, and A. Schönhuth. Clever: clique-enumerating variant finder. *Bioinformatics*, 28(22):2875–2882, 2012.

A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C.-y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 42 (D1):D142–D147, 2014.

I. Maze, L. Shen, B. Zhang, B. A. Garcia, N. Shao, A. Mitchell, H. Sun, S. Akbarian, C. D. Allis, and E. J. Nestler. Analytical tools and current challenges in the modern era of neuroepigenomics. *Nature Neuroscience*, 17(11):1476–1490, 2014.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano. GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5):495–501, 2010.

C. A. Meyer and X. S. Liu. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, 15(11):709–721, 2014.

J. C. Miller, B. D. Brown, T. Shay, E. L. Gautier, V. Jojic, A. Cohain, G. Pandey, M. Leboeuf, K. G. Elpek, J. Helft, et al. Deciphering the transcriptional network of the dendritic cell lineage. *Nature Immunology*, 13(9):888–899, 2012.

K. B. Mullis and F. A. Faloona. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in Enzymology*, 155:335–350, 1987.

P. Nemenyi. Distribution-free multiple comparisons. In *Biometrics*, volume 18, page 263, 1962.

A. T. Nguyen and Y. Zhang. The diverse functions of Dot1 and H3K79 methylation. *Genes & Development*, 25(13):1345–1358, 2011.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

S. Pepke, B. Wold, and A. Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, 6(Suppl 11):22–32, 2009.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. 77(2):257–286, 1989.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

K. Robasky and M. L. Bulyk. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, 39(Suppl 1):D124–D128, 2011.

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

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

C. S. Ross-Innes, R. Stark, A. E. Teschendorff, K. A. Holmes, H. R. Ali, M. J. Dunning, G. D. Brown, O. Gojis, I. O. Ellis, A. R. Green, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481(7381):286, 2012.

J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology*, 27(1):66–75, 2009.

S. Saeed, J. Quintin, H. H. Kerstens, N. A. Rao, A. Aghajanirefah, F. Matarese, S.-C. Cheng, J. Ratter, K. Berentsen, M. A. van der Ent, et al. Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science*, 345(6204):1251086, 2014.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

C. M. Sawai, V. Sisirak, H. S. Ghosh, E. Z. Hou, M. Ceribelli, L. M. Staudt, and B. Reizis. Transcription factor Runx2 controls the development and migration of plasmacytoid dendritic cells. *The Journal of Experimental Medicine*, 210(11):2151–2159, 2013.

Z. Shao, Y. Zhang, G.-C. Yuan, S. H. Orkin, and D. J. Waxman. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biology*, 13(3):R16, 2012.

L. Shen, N.-Y. Shao, X. Liu, I. Maze, J. Feng, and E. J. Nestler. diffReps: Detecting Differential Chromatin Modification Sites from ChIP-seq Data with Biological Replicates. *PLoS One*, 8 (6):e65598, 2013.

J. Shendure and H. Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, 2008.

G.-X. Shi, K. Harrison, G. L. Wilson, C. Moratz, and J. H. Kehrl. RGS13 regulates germinal center B lymphocytes responsiveness to CXC chemokine ligand (CXCL)12 and CXCL13. *The Journal of Immunology*, 169(5):2507–2515, 2002.

R. J. Sims Iii and D. Reinberg. Processing the H3K36me3 signature. *Nature Genetics*, 41(3): 270–271, 2009.

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

C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904.

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

H. J. Szerlong and J. C. Hansen. Nucleosome distribution and linker dna: connecting nuclear function to dynamic chromatin structure. *Biochemistry and Cell Biology*, 89(1):24–34, 2010.

J. P.-Y. Ting and J. Trowsdale. Genetic control of MHC class II expression. *Cell*, 109(2):S21–S33, 2002.

A. Valouev, D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, and A. Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, 5(9):829–834, 2008.

A. Weiner, A. Hughes, M. Yassour, O. J. Rando, and N. Friedman. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Research*, 20 (1):90–100, 2010.

E. G. Wilbanks and M. T. Facciotti. Evaluation of algorithm performance in ChIP-seq peak detection. *PloS One*, 5(7):e11471, 2010.

H. Xu, C.-L. Wei, F. Lin, and W.-K. Sung. An hmm approach to genome-wide identification of differential histone modification sites from chip-seq data. *Bioinformatics*, 24(20):2344–2349, 2008.

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

T. Yoshimoto, K. Takeda, T. Tanaka, K. Ohkusu, S.-i. Kashiwamura, H. Okamura, S. Akira, and K. Nakanishi. IL-12 up-regulates IL-18 receptor expression on T cells, Th1 cells, and B cells: synergism with IL-18 for IFN-gamma production. *The Journal of Immunology*, 161 (7):3400–3407, 1998.

J. Zhang, A. Raper, N. Sugita, R. Hingorani, M. Salio, M. J. Palmowski, V. Cerundolo, and P. R. Crocker. Characterization of Siglec-H as a novel endocytic receptor expressed on murine plasmacytoid dendritic cell precursors. *Blood*, 107(9):3600–3608, May 2006.

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

Z. D. Zhang, J. Rozowsky, M. Snyder, J. Chang, and M. Gerstein. Modeling chip sequencing in silico with applications. *PLoS Computational Biology*, 4(8):e1000158, 2008.