Robust Normalization of Next Generation Sequencing Data

Dissertation

zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von Johannes Helmuth

Berlin, 2017

Erstgutachter: Prof. Dr. Martin Vingron
Zweitgutachter: Prof. Dr. Uwe Ohler

Tag der Disputation: 15.05.2017

Preface

This dissertation introduces a robust normalization method to uncover signals in noisy next generation sequencing data. The genesis of the described approach is the observation that next generation sequencing resembles a sampling process that can be modeled by means of discrete statistics. The specific and sensitive detection of signals from sequencing data pushes the field of molecular biology forward towards a comprehensive understanding of the functional basic unit of life – the cell. The thesis is structured in three parts:

Part I provides a background on molecular biology and statistics that is needed to understand Part II. I describe the fascinating subject of gene regulation and how diverse next generation sequencing techniques have been developed to study cellular processes at the molecular level. The data generated in these experiments are naturally modeled with statistical models. To accurately quantify sequencing data, I propose the computational program "bamsignals" which was developed in collaboration with Alessandro Mammana [1].

Part II introduces a novel sequencing data normalization method which was developed under supervision of Dr. Ho-Ryun Chung from the Epigenomics laboratory at the Max Planck Institute for Molecular Genetics in Berlin. A manuscript describing the approach is deposited on bioRxiv [2] and the method was also featured as a journal article in Springer Press BioSpektrum [3]. The method was implemented as an open-source software [4].

Part III provides the conclusion. The findings of Part II are wrapped up. Furthermore, I give future directions for research in the field of next generation sequencing data analysis.

Acknowledgments

Along the way to write this dissertation, many people helped me in pursuing my goal to become a computational biologist. First of all, I would like to thank Dr. Ho-Ryun Chung for his supervision and financial support. I always cherish his advice and critic on all my scientific endeavors. His comments on the manuscript greatly improved this thesis. Fruitful discussions with my referees Prof. Dr. Martin Vingron and Prof. Dr. Uwe Ohler helped me understand the science of computational biology.

I would like to give thanks to all members of the "Otto-Warburg-Laboratories: Computational Epigenomics" group and the department of "Computational Molecular Biology" at the Max Planck Institute for Molecular Genetics in Berlin. The work environment was always inspiring and gave rise to multiple successful collaborations that did not make it into the book [5–7]. In particular, I thank Dr. Brian Caffrey (*"Bhuaigh Èire!"*) and Marcos Torroba (*"¡Sigue así!"*) for proofreading the manuscript.

Also, I am thankful to be given the opportunity to represent the PhD students of the Max Planck Institute for Molecular Genetics in "Student Association" in the year 2013. The organization of scientific and social events was fun and provided me with feedback on my scientific work.

Finally, I would like to thank my darling Sina (*"Ich liebe Dich!"*), my parents Ingo and Charlotte and my brother Christian for their motivation along the way.

List of Publications

Kinkley, S* & **Helmuth, J***, Polansky, JK, Dunkel, I, Gasparoni G., Fröhler, S., Chen, W., Walter, J., Hamann, A. & Chung, HR.: "reChIP-seq reveals widespread bivalency of H3K4me3 and H3K27me3 in CD4+ memory T cells". *Nature Communications* (2016) 7:12514 `doi:10.1038/ncomms12514`

Helmuth, J. & Chung, HR.: "Auswertung von Histonmodifikations-ChIP-Seq-Datensätzen". *Biospektrum* (2016) 22: 568. `doi:10.1007/s12268-016-0728-6`

Helmuth, J., Li, N., Arrigoni, L., Gianmoena, K., Cadenas, C., Gasparoni, G., Sinha, A., Rosenstiel, P., Walter, J., Hengstler, JG., Manke, T. & Chung, HR.: "normR: Regime enrichment calling for ChIP-seq data". *bioRxiv* (October 2016). `http://dx.doi.org/10.1101/082263`

Corwin, T., Woodsmith, J., Apelt, F., Fontaine, JF., Meierhofer, D., **Helmuth, J.**, Grossmann, A., Andrade-Navarro MA., Ballif, B., Stelzl, U.: "Interaction networks mediate human tyrosine kinase specificity". *Cell Systems* (accepted)

Yang, J., Moeinzadeh, MH., Kuhl, H., **Helmuth, J.**, Xiao, P., Liu, MG., Zheng, JL., Sun, Z., Fan, W., Deng, MG., Wang, H., Hu, F., Fernie, A., Timmermann, B., Zhang, P. & Vingron, M.: "The haplotype-resolved genome sequence of hexaploid Ipomoea batatas reveals its evolutionary history". *Nature Plants* (accepted)

Contents

List of Figures

List of Tables

Part I

Background

Chapter 1

The Basics of Molecular Biology

This chapter gives a general introduction to molecular biology with references for in depth exploration of the subtopics. Section 1.1 explains the genome and its organization into the epigenome mediated by a dynamic biopolymer structure called chromatin. Herein, I describe how the information encoded in the genome is dynamically readout in the context of chromatin. Section 1.2 describes how next generation sequencing techniques are used to measure molecular quantities like gene readout and chemical modifications to the chromatin.

1.1 The Genome and the Readout of Genetic Information

The cell is sometimes referred to as the functional basic unit of life. Aside from unicellular organisms like bacteria, complex multicellular organisms consist of hundreds to billions of cells with diverse functions and phenotypes. Astonishingly, all the instructions to build such diverse cell types in an organisms are stored as heritable information in the genome – a linear polymer of deoxyribonucleotides, *i.e.* the DNA. In turn, the genome resides in the nucleus of the cell and is organized together with histone proteins in a higher-order structure called "chromatin". Apart from packaging the genome, the histone proteins are subject to diverse chemical modifications that dynamically adjust the readout of the genetic information. Diverse environmental stimuli require a living cell to vigorously adapt which genetic information are read out and set into operation. This section describes how the dynamic nature of the chromatin facilitates the dynamic regulation of the readout of the genetic information – a property that is required to specify adequate cellular responses to stimuli and, moreover, to built distinct cell types.

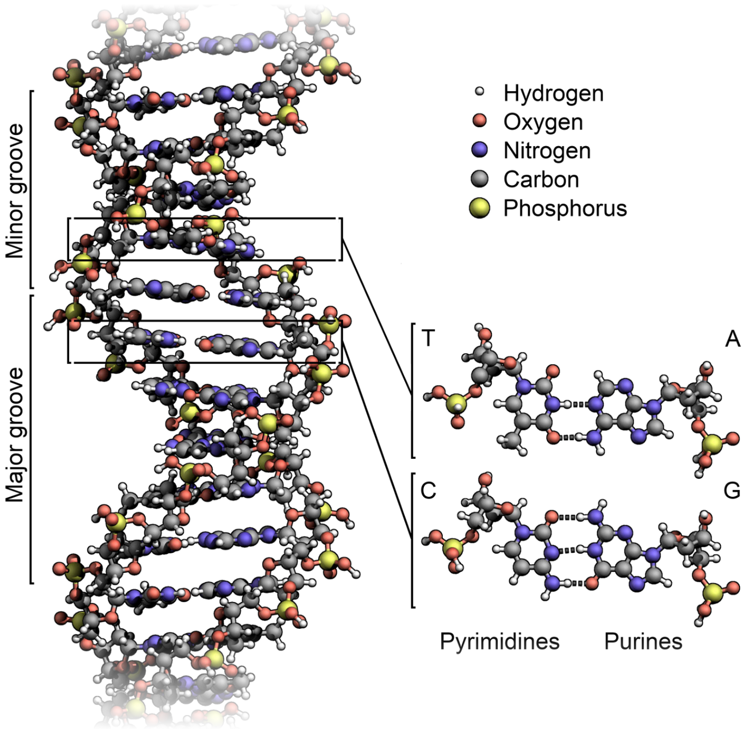

Fig. 1.1 – **The DNA Double Helix.** Balls denote atoms and edges are bonds. Bases are paired by hydrogen bonds (dashed lines). Base pairs form a deoxyribose back- bone with phosphodiester bonds. Twists of double strand form two different types of grooves, *i.e.* minor and major groove. Illustration adapted from Richard Wheeler [9].

1.1.1 The DNA

The seminal discovery of the DNA double helix by James D. Watson and Francis H.C. Crick in 1953 [8] paved the way for molecular genetics. The deoxyribonucleic acid (DNA) is a double stranded helix composed of a deoxyribose backbone and four nucleotides, namely adenine (A), cytosine (C), guanine (G) and thymine (T; Fig. 1.1). These nucleotides are sometimes called bases and are faced towards the fiber axis and build complementary base pairs (bp) by forming hydrogen bonds. A and T pair with two hydrogen bonds whereas G pairs with C by forming three bonds. In consequence, the DNA strands exhibit reverse complementarity. The deoxyribose backbone is established by phosphodiester bonds between the 3’ carbon atom of a deoxyribose and the 5’ carbon of the adjacent deoxyribose. A dinucleotide of C and G is denoted as CpG, where “p” refers to the phosphodiester bond linking the two. This results in a 5’ and 3’ end of the DNA, where the 5’ to 3’ direction is referred to as “downstream” and the 3’ to 5’ direction is called “upstream”.

The sequence of nucleotides builds the “genome” – a text that encodes for the heritable information of the cell. This static heritage is sometimes referred to as the book of life and contains all the instructions to build a functioning cell. Even when the human genome was completely sequenced in 2001 [10], it is still under intensive investigation which information is encoded exactly and how the readout of that information is regulated to create distinct cell types in a complex organism. Moreover, the genome slightly differs between individuals of a species creating a plethora of distinct "books" (see for example [11] for a study on the variations in the human genome). The genomes of complex organisms like eukaryotes are organized in chromosomes, *e.g.* 23 in human, which reside in an organelle of the cell called the "nucleus". The tight compaction of the DNA, *e.g.* 2 meters in human, into the ∼6 µm-sized nucleus is facilitated by a heteropolymer of DNA and histone proteins called the "chromatin".

1.1.2 The Chromatin

Eukaryotic genomes are packaged into chromatin whose basic repeating unit is the nucleosome. Nucleosomes form upon the association of two copies of each core histone, namely H2A, H2B, H3 and H4, with ∼147bp of DNA [12, 13]. In consequence, the sequence of nucleosomes forms the so called "beads-on-a-string" structure which is found at genomic regions that are read (*active*; Fig. 1.2). A further compaction into 30nm fibres is found at loci that are not read out (*inactive*). Thus, the chromatin enables the tight but dynamic packaging of the genome into the cell nucleus. The different levels of the chromatin compaction allow, at times, certain genome regions to be read while other regions are made inaccessible. For example the sole presence of a nucleosome can render the DNA less accessible. Nevertheless, how is this dynamic compaction of the genome achieved? The unstructured amino-terminal parts of the histones are frequently chemically modified by enzymes - a process that provides a regulation of the compaction with high fidelity. Those chemical modifications to histones include acetylation, methylation and phosphorylation. For instance, acetylation results in a positively charged histone that destabilizes its

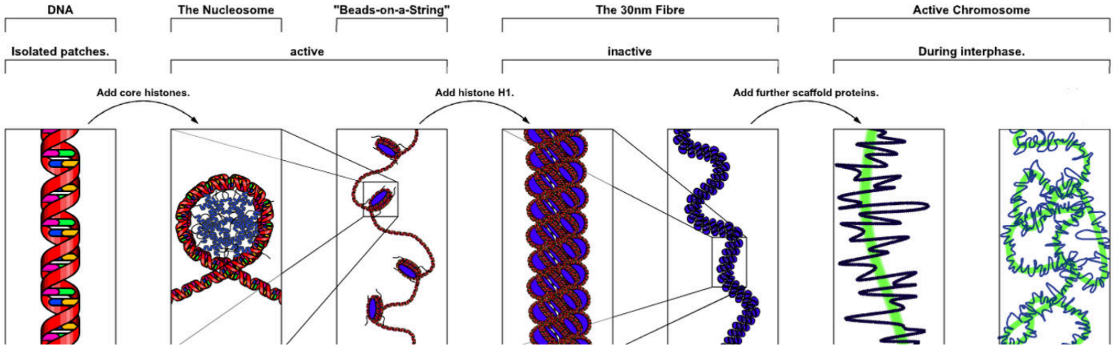

Fig. 1.2 – **Levels of Dynamic Chromatin Compaction.** The chromatin consists of consecutive nucleosomes containing DNA and histone proteins on a "beads-on-a-string" structure in genome regions whose information is currently read ("active"). Further compaction to the 30nm fibre marks genomic regions that are not read out ("inactive"). A chromosome is then a composite of active and inactive states. Adapted from Richard Wheeler [14].

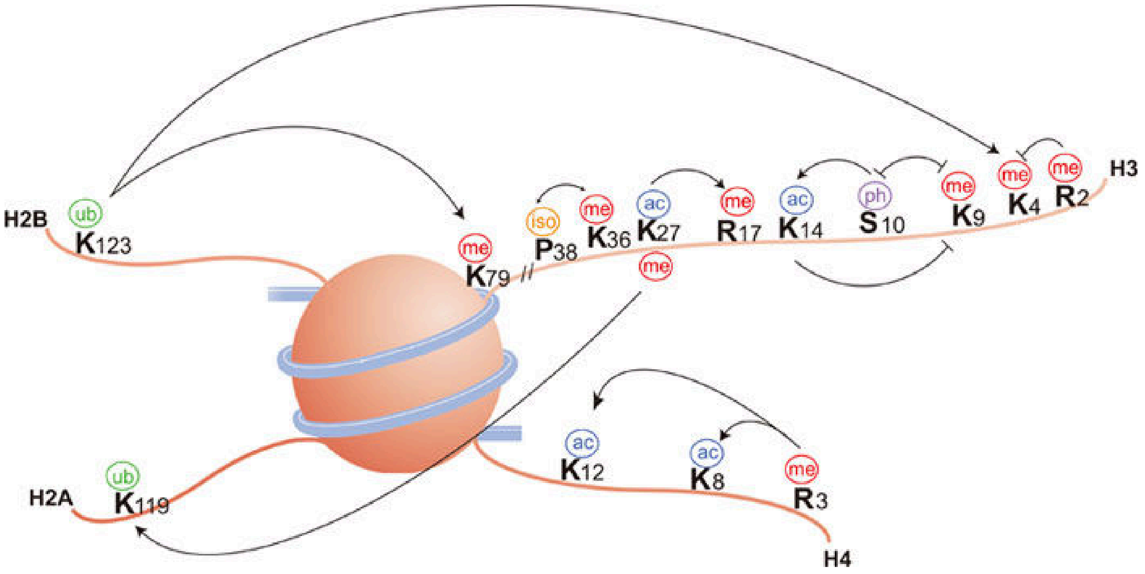

Fig. 1.3 – The Combinatorial Action of Histone Modifications. DNA (blue) is wrapped around a heterodimer of histones H2A, H2B, H3 and H4. Unstructured amino-terminal tails of the histones (red) are subject to chemical modifications. Genome-wide, those chemical modifications either do (arrows) or do not (blunt ends) coincide. Adapted from [17]

contact to the negatively charged DNA and, in consequence, results in less compaction [15, 16]. Traditionally, histone modifications are denoted after the histone they reside on and the identity and position of the amino acid being modified and, finally, the molecule that is attached to that amino acid (Fig. 1.3). From a simplistic view point certain histone modifications such H3 lysine 4 tri-methylation (H3K4me3) and the aforementioned acetylations facilitate accessibility to the DNA, whereas others such as H3K9me3 and H3K27me3 facilitate compaction [17]. However, the true complexity of this histone "language" is thought to be delivered through the combination of histone modifications [18, 19] (Fig. 1.3). Thus, the static information of the genome is interpreted on a dynamic level in context of chromatin, referred to as the "epigenome".

Apart from chemical modifications to the histones, the DNA template itself can be chemically modified. The most prevalent DNA modification is the methylation of CpG dinucleotides that is catalyzed by DNA methyltransferases and predominantly found in less accessible compacted chromatin regions (reviewed in [20]). Whether the histone and DNA modifications are generally passed on to the progeny is still a matter of debate (*e.g.* [21–24]). This begs the question how epigenetic information is preserved or re-established.

1.1.3 The Readout of Genetic Information

In 1970, Francis Crick proclaimed the general flow of genetic information to the functional level in the "Central Dogma of Molecular Biology" [25]:

> "The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid."

In essence, every protein in a cell originates from a specific genomic region referred to as "gene". Genes are probably the best characterized elements in the genome and their information is read out by an enzyme called RNA polymerase (RNAP) in a process called "transcription". The number and identity of genes transcribed is related to the status of the cell and depends on many factors, *e.g.* cell type and environmental conditions. How gene transcription is regulated has been studied thoroughly in the last decades and, in the recent years, this regulation was mostly studied in conjunction with the dynamic nature of the chromatin. In this book I will focus on the regulation of the readout of genetic information in transcription but, note, there exist regulatory steps past this layer (for example reviewed in [26, 27]).

The Transcription Cycle

The process of transcription transfers the information encoded by a gene into a ribonucleic acid template (RNA). During transcription, the RNAP proceeds through distinct states [28]. In the *pre-initiation step*, the TATA-binding protein (TBP) binds a characteristic DNA sequence called "TATA-box" in the "promoter" which describes the region around the "transcription start site" (TSS). However, the majority of eukaryotic promoters do not contain a TATA-box. Instead TBP-related factors (TRFs) recognize other still poorly described core promoter sequences. The binding of TBP and/or TRFs induces the formation of the "pre-initiation complex" which, in turn, facilitates the *RNAP recruitment.* In the *initiation step,* the RNAP starts synthesizes $\sim$50 nucleotides of a reverse complementary RNA and passes them to add a guanine via a 5' to 5' triphosphate bond to the RNA transcript which prevents enzymatic degradation. After the *transition into the elongation*, the RNAP leaves the promoter region and continues to read the gene in 5' to 3' direction to synthesize an RNA as a reverse complement to the non-coding DNA strand in the *elongation step.* In the *termination step*, the RNAP cleaves the transcript at the "transcript termination site". If a polyA signal, *i.e.* a characteristic sequence motif, exists, the RNAP adds up to 250 consecutive Adenines to the transcript. After RNAP release it can re-initiate transcription again. If the transcribed gene encodes for a protein, the RNA is exported to the cytosol and translated into a protein by ribosomes. Note that RNAs themselves can regulate gene transcription (reviewed in [29]).

The Regulation of Transcription

In molecular biology, the study of gene regulation deals with how genes are transcribed at different rates in different cells. In a multicellular organism every cell has the same genome, yet the genetic information is readout differently giving rise to distinct phenotypes like brain neurons or liver cells. Essentially, a cell type can be defined to a large extend by its transcriptional program but how exactly are these transcriptional programs detailed?

Proteins play a pivotal role in the regulation of transcription. For example, the described TATA-binding protein facilitates the initiation of transcription through binding to a characteristic DNA sequence and the recruitment of the pre-initiation complex. In fact, there exists a huge class of proteins called "transcription factors" or "*trans*-acting factors" which are all attracted by specific nucleotide sequences that are referred to as "motifs". A *trans*-acting factor can directly or indirectly (*i.e.* through other proteins called "co-factors") regulate the transcription of a gene after binding a specific genomic locus. This regulatory genomic region is referred to as a "*cis*-regulatory element" of a gene. Promoters are generally enriched for these DNA elements to allow for their targeted transcriptional control. Yet, some *cis*-regulatory elements can also be located far away from their targets and only the dynamic looping of the DNA brings these elements in spatial proximity of their targets [30]. These regulatory regions are referred to as "enhancers". Recently, efforts were made to comprehensively catalogue transcription factor motifs [31] and to identify the occurrences of those motifs in the static genome [32]. While these approaches paved the way to study gene regulation, they neglect a central compartment of the eukaryotic genome that can fundamentally influence the DNA binding of proteins – the dynamics of the chromatin.

The Dynamic Chromatin

Apart from packaging the DNA, the chromatin serves essential roles in the regulation of transcription and is traditionally classified in two forms: The "euchromatin" harbors accessible DNA and is rich in actively transcribed genes as well as *cis*-regulatory elements. The euchromatic regions tend to localize towards the center of the cell nucleus. On the other hand, the "heterochromatin" is characterized by higher degree of compaction which renders the DNA less accessible. Heterochromatic regions localize to the exterior of the nucleus, referred to as "lamina", and contain only a few genes which are mostly inactive.

Aside from the mere level of compaction, there are indications that the role of the chromatin in gene regulation is multi-faceted. Firstly, the modifications to histones serve as dynamically deposited binding residues for proteins. So called chromatin modifiers can read, catalyze or remove histone modifications resulting in a natural language of histone modifications. For example, the "Enhancer of zeste homolog 2" (EZH2) enzyme as part of the polycomb repressive complex 2

(PRC2) catalyzes the tri-methylation of H3K27 (H3K27me3) which, itself, propagates and stimulates EZH2 activity (reviewed in [33]). By binding multiple H3K27me3-modified nucleosomes through its Pc subunit, PRC1 facilitates a stable compaction of the chromatin which results in gene silencing and the putative preclusion of transcription factor binding.

Secondly, histone modifications are related to the transcriptional status of the DNA [34]. For example, histone acetylation and H3K4me3 are found at accessible genomic loci where transcription can be initiated whereas H3K27me3 and H3K9me3 are found in heterochromatic repressed regions [35]. The observed co-occurrence of certain histone modifications lead to the notion of a "chromatin state" [36] which is a genome-wide re-occurring pattern of coinciding histone modifications. This concept allows for a compact and still comprehensive description of the epigenome with a small number (*e.g.* $\leq$15) of chromatin states.

A last facet of the role of the chromatin in gene regulation is distinct hierarchies of chromatin folding facilitating a spatial organization of the chromatin in the nucleus. Some regions, *e.g.* telomeres, are stably associated to the nuclear lamina, *i.e.* "lamina associated domains" (LADs), whereas some segments are organized in "topologically associated domains" in nuclear interior (see [37] for review).

1.2 Measuring the Cell by Next Generation Sequencing

For the last decade, next generation sequencing (NGS) has been the experimental technique of choice to quantify molecular properties genome-wide. The NGS technique is standardized and, compared to polymerase chain reaction [38], it achieves a higher throughput with million of data points generated. Yet, similar to the low-throughput polymerase chain reaction, it follows a "sequencing-by-synthesis" approach where a short segment of DNA is reverse complementary synthesized to a DNA fragment isolated from a sample of cells. These short segments (36 to 50bp) are called "reads" and their original location in the genome is determined to the bp in a computational process called "mapping". In this section, I will explain how NGS can be used to quantify the level of transcription and to identify genome-wide protein binding sites.

1.2.1 Gene Expression

RNA-seq refers to the capture of the transcriptome via NGS (for review see [39]). There exist different protocols which generally follow these steps: Firstly, a sample of cells is lysed, *e.g.* by the reagent TRIzol. Secondly, the RNA species of interest are selected, *e.g.* polyA selection for mRNAs. Alternatively, some protocols simply deplete gratuitous RNA species, *e.g.* depletion of ribosomal RNA ($\geq$90% of the transcriptome). Thirdly, RNA is reverse transcribed to clonal DNA (cDNA) and then fragmented. Finally, the resulting fragments are end-sequenced to generate reads which are then mapped to the reference genome. The number of reads overlapping a specific region is called “coverage” and reflects a quantitative measurement of the steady state RNA abundance.

A derivate of RNA-seq is represented by an RNA species selection called Cap Analysis of Gene Expression (CAGE) [40]. Therein, after fragmentation, one enriched for 5’-capped RNA molecules which are then reverse transcribed and end-sequenced ($\sim$ 27 nucleotides) [41]. When aligned to the reference genome the generated short reads are indicative for transcriptional start sites.

1.2.2 Chromatin Modifications

ChIP-seq

Chromatin Immunoprecipitation followed by high-throughput sequencing (ChIP-seq) [42] has become a standard method to determine the localization of DNA-associated proteins, like transcription factors or histone modifications. In brief, after proteins are crosslinked with formaldehyde to the DNA, the chromatin is sheared and the resulting chromatin fragments are enriched by immunoprecipitation for the protein of interest. This precipitate is reverse-crosslinked to obtain DNA fragments, which are amplified and then end-sequenced. The reads generated in this way are mapped to a reference genome and genomic loci bound by the antigen are inferred by an accumulation of sequencing reads. The accurate identification of these loci will be the subject of this thesis.

Due to genome-wide scalability and cost-efficiency of ChIP-seq, hundreds of distinct proteins and their modifications have been assayed to study underlying mechanisms of molecular function in different cell types [43, 44]. Previously, ChIP-seq data have been used to characterize transcription factor binding sites [45] and chromatin states [46]. As a derivative to antibodies, histone modification-specific interaction domains from chromatin binding proteins have been used for precipitation [47].

WGBS

In Whole Genome Bisulfite Sequencing (WGBS) the isolated DNA is treated with bisulfite to convert unmethylated cytosines into thymines prior to amplification. When aligned to the reference genomes, this substitution is detected and relative DNA methylation levels at this nucleotide in the sample population can be quantified.

Chapter 2

Mathematical Concepts

In the previous chapter I introduced the concept of read counts to investigate molecular properties such as gene expression and the location of DNA-associated proteins by Next Generation Sequencing. This chapter provides mathematical prerequisites aiding the identification of biological phenomena by means of modeling discrete read count distributions. Discrete statistics provide an appropriate framework to infer biological properties of the data in the presence of uncertainty. When many statistical tests are performed on the same data set a careful multiple testing correction is essential. In my thesis I use mixture models and focus on maximum likelihood parameter estimation via the Expectation-Maximization algorithm. Together the introduced concepts provide the basis for the studies described in the following chapters of the thesis.

2.1 Statistical Prerequisites

In Section 1.2 I described how a Next Generation Sequencing (NGS) experiment generates a pool of DNA fragments that are then end-sequenced. The alignment of reads to the reference genome gives rise to characteristic read count patterns across the genome. When dealing with count data discrete statistics is the natural language to model processes that generate these random variables.

2.1.1 Statistical Inference

Given a statistical model, hypotheses can be tested on the data. For example in a NGS experiment we could ask if an observed non-negative read count at a genomic locus is substantially larger than expected under a statistical model H_0 on the read count distribution. Formally, a realization x of a random variable X is assigned the probability of observing a value at least as extreme as x under H_0, *e.g.* for "at least as great as x" then this probability is given by

$$P = P(X \geq x|H_0 \; is \; valid) = 1 - \sum_{k=0}^{x-1} P(X = k|H_0 \; is \; valid), \tag{2.1}$$

where H_0 is also referred to as the "null model" or "null hypothesis" and the calculated probability P is called "p-value". The p-value is the probability of sampling another observation from the null hypothesis that is as far or farther away from the value of x. Usually, a threshold probability, referred to as "significance level" or α, is used to identify observations that reject H_0 as their generative process. For example, with $\alpha = 0.05$ the probability that H_0 gets falsely rejected is 5%, referred to as "type-1-error" or "false discovery".

An appropriate framework to model read count data is provided by probability theory. As a formal description of a statistical hypothesis a probabilistic model is interpretable. The interpretability allows for the sampling of new observations and to reason from the data. Various discrete distribution families have been used to model read count distributions, *e.g.* the Poisson [48] or the Negative Binomial [49, 50] distribution. For example, a researcher could encode expected NGS read count patterns into a Poisson distribution. Let X denote a random variable that follows a Poisson distribution with $\lambda \geq 0$: $X \sim \text{Pois}(\lambda)$. The probability of observing a non-negative integer x is:

$$P(X = x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \tag{2.2}$$

where e denotes *Euler's number.* This function is also referred to as the "probability mass function". The main properties of a Poisson model are the **independence** and **homogeneity** of observations. For details the reader is referred to [51].

The Poisson null model H_0 can be used to identify substantial deviations in observed read count patterns from an expected outcome. In our example, the researcher could approximate λ by $\hat{\lambda} = \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$, the arithmetic mean of read counts for all genomic regions $i = 0, \ldots, m$ (Figure 2.1). Together with Equation (2.2) and (2.1) a p-value is assigned to each observation x_i:

$$\begin{aligned} P(X \geq x_i|\lambda = \hat{\lambda}) &= 1 - P(X \leq (x_i - 1)|\lambda = \hat{\lambda}) \\ &= 1 - \sum_{k=0}^{x_i-1} \frac{\lambda^k e^{-\lambda}}{k!} \\ &= 1 - e^{-\hat{\lambda}} \sum_{k=0}^{x_i-1} \frac{\hat{\lambda}^k}{k!}, \end{aligned}$$

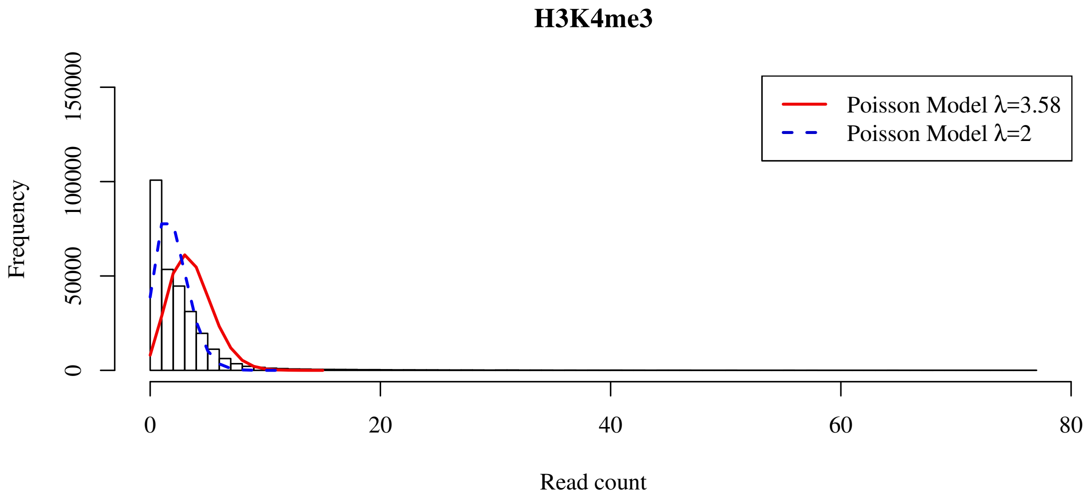

Fig. 2.1 – A Simple Poisson Model for ChIP-seq Read Count Data. H3K4me3 ChIP-seq read counts in 500bp genomic windows (histogram) and Poisson model fits for $\lambda = \bar{x} = 3.58$ (red solid line) and $\lambda = \text{median}(x) = 2$ (blue dashed line).

where $P_\lambda(X \leq (x_i - 1)) = F_\lambda(x_i - 1)$ is the Poisson "distribution function" evaluated at $(x_i - 1)$.

Using this Poisson framework the researcher could identify regions with an observed read count distant from the read count that would be expected given $\hat{\lambda}$. For example, a genomic region with significantly higher RNA-seq read coverage than expected by the model could be indicative of a highly transcribed gene – or even just a certain exon of a gene that is very highly transcribed. In ChIP-seq, a read count that is far from the expectation indicates the binding of a protein. Also, highly DNA-accessible regions show much higher DNaseI-seq read coverage than the genomic average. In summary statistical inference enables the reasoning of putative biological phenomena generating observed NGS read count patterns.

2.1.2 Multiple Testing Correction and the T Method

The p-value is the probability that a test is going to produce a statistic at least as extreme assuming the truthfulness of the null hypothesis. In computational biology hundreds or even thousands of statistical tests are performed on the data set, *e.g.* testing each of ∼20,000 genes for differential expression. Every statistical test is deemed significant for a threshold α, say 0.05. However, with an increase in the number of performed tests the chance of incorrectly rejected H_0 for one or more tests also increases, referred to as **type-1-error accumulation**. Say $m = 20,000$ independent tests are performed, and it is known that all null hypotheses are true $m_0 = m$, on average $1,000$ tests are incorrectly called significant at $\alpha = 0.05$ (Figure 2.2A). Thus, the nominal p-values are misleading due to the type-1-error accumulation (see [52] for review).

An early and naïve approach is the Bonferroni method [53] which controls the family-wise error rate (FWER), *i.e.* the chance that at least one true null was falsely rejected. To this end the significance level α is transformed by

$$\alpha^* = \frac{\alpha}{m_0} \tag{2.3}$$

with $m_0 = m$, which guarantees to control the FWER. However, this approach reduces power in detecting true non-nulls if many tests are performed because generally m_0 is smaller than m.

Instead of the strict FWER control the Benjamini-Hochberg procedure [54] controls the expected proportion of discoveries that are false given at least one discovery, referred to as the **false discovery rate** (FDR). This approach assumes that not all H_0 are true by controlling the FDR at a level α. Given a vector of sorted p-values p the method finds the largest k such that

$$p_i \leq \frac{k}{m_0} \cdot \alpha \tag{2.4}$$

with $m_0 = m$. Alternatively, the procedure can adjust sorted p-values p via

$$p_{BHi} = \min \left\{ m_0 \cdot \frac{p_i}{i}, 1 \right\}. \tag{2.5}$$

This method implicitly accounts for the fact that $m_0 \leq m$ by penalizing p-values according to their sorted index. Yet, an explicit account for the number of true null hypotheses m_0 would increase statistical power.

The proportion of true null hypotheses $\pi_0 = m_0/m$ is estimated by adaptive FDR controlling methods [55,56]. In general, the distribution of p-values is continuous and uniform on the interval $[0, 1]$ if all H_0 are true (Figure 2.2A). In contrast, non-null p-values are skewed towards small values (Figure 2.2B). In a real scenario the distribution of p-values is a composite of these two p-value populations which impedes the determination of m_0 (Figure 2.2C). In principle, m_0 can be estimated by modeling these mixtures as mixture models (see [57] for review), *e.g.* a composite of a Uniform distribution for H_0 tests and a Beta distribution for H_1 tests [58,59]. Mixture models will also be discussed in Section 2.1.5 of this book. Adaptive FDR controlling methods make use of the fact that greater p-values most likely originate from true H_0. For example, Storey's method [56] estimates $\hat{\pi}_0$ by counting the number of p-values m_λ that are greater than a cut-off λ. Because of the uniformity of null p-values

$$\hat{\pi}_0 = \frac{m_\lambda}{m \cdot (1 - \lambda)}$$

can be calculated, traditionally for any $\lambda \geq 0.5$. The estimated number of true null hypotheses is then $\hat{m}_0 = \hat{\pi}_0 \cdot m$ and can be plugged into Equations (2.3), (2.4) or (2.5) leading to an adaptive and, thus, less strict multiple testing correction.

Alternatively to Benjamini-Hochberg corrected p_{BHi} defined by Equation (2.5) FDR-adjusted

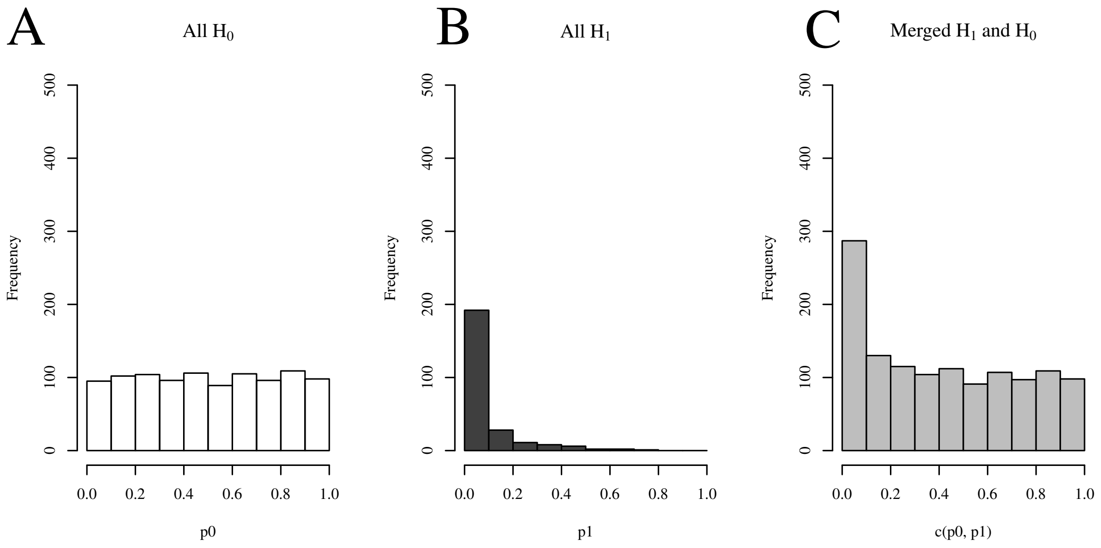

Fig. 2.2 – The Anatomy of the P-Value Distribution. Three types of P-value distributions are given: (A) If all H_0 are true, the P-value distribution resembles a uniform distribution in the interval $[0, 1]$. (B) If all H_1 are true, the P-value distribution is skewed towards smaller values. (C) In a real scenario the P-value distribution is a mixture of H_0 and H_1.

p-values can be calculated with $\hat{m}_0$ by the expectation that at level α a type-1-error occurs at a rate of $\alpha \cdot m_0$. To this end, Storey [56] defines $R(\gamma)$ as the number of nominal p-values less than or equal tho γ. The **q-value** is then given by

$$q_1 = p_1 \cdot \frac{\hat{m}_0}{R(p_1)} \text{ and}$$
$$q_{i+1} = \max\left\{q_i, p_{i+1} \cdot \frac{\hat{m}_0}{R(p_{i+1})}\right\}.$$

Hence, these multiple testing corrected p-values q's can now be filtered based on a significance level α.

Note that these multiple testing correction procedures have been proven to perform well for continuous data like microarray chip data. However, statistical inference on discrete NGS counts with, for example, a Poisson model is impeded by the fact that there exists only a finite number of achievable p-values. The methods outlined above may not perform as required without some adjustment. This adjustment is achieved by the “T method” which will be described in the next section.

The T Method: Estimate π_0 when Statistics Are Discrete

In statistical tests every hypothesis i is associated to a test statistic $\mathcal{X}_i$ dependent on the statistical model. For continuous data the test statistic $\mathcal{X}_i$ and the p-value p_i are continuous. If H_0 is uniform on $[0, 1]$ the methods described in the previous section apply readily to this scenario. Yet, for discrete data, the test statistic $\mathcal{X}_i$ and, hence, the p-value p_i are discrete, *i.e.* there exists only a fixed set $\mathcal{S}_i$ of j achievable p-values dependent on the ancillary statistic $\mathcal{A}_i$. For example, the contingency table margin represents the ancillary statistic in Fisher's exact test [60] which, in fact, varies with i. In this case the correct estimation of π_0 is complicated by the dependence of $\mathcal{X}$ and p on i which impedes multiple testing correction. Specifically, π_0 gets over-estimated [61] which leads to less statistical power in detecting true differences.

Various methods have been proposed to estimate π_0 in discrete testing problems [62–64]. The "T method" [61] represents a straight forward filtering approach on p to improve on downstream multiple testing correction: Remove all tests where the test statistic can never be rejected at a nominal level α. Those tests have zero power – even with increasing effect size. Formally, the T method generates a reduced list of p-values p' for a significance level α with

$$p' = \{p_i \in p : \exists x \in \mathcal{A}_i \; P(X \geq x | \; H_0 \; is \; valid) \leq \alpha\} \, .$$

The so filtered p' can then be used for multiple testing correction, *e.g.* in Benjamini-Hochberg because its distribution is more uniform on $[0.5, 1]$.

In Computational Biology, zero power can result from a low NGS read coverage. For example, a conditional read count difference of 5 to 10 and of 50 to 100 both constitute a fold change of 2, yet one intuitively attributes a higher confidence to the latter. Even a greater fold change (effect size) does not affect this implication. Please refer to Chapter 3 for the interplay of power and effect size in the statistical analysis of NGS read count data.

2.1.3 The Binomial Distribution

Aside from the Poisson distribution introduced in section 2.1.1 the **Binomial distribution** is a possible model for discrete count data. Let X denote a random variable that follows a Binomial distribution with a number of trials $n \geq 0$ and a success probability $0 \leq \theta \leq 1$: $X \sim \text{Bin}(n, \theta)$. The probability of observing k successes with $0 \leq k \leq n$ is given by the probability mass function

$$f(k|n, \theta) = P(X = k|n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}, \tag{2.6}$$

where $\binom{n}{k} = n!(k!(n-k)!)^{-1}$, referred to as "binomial coefficient". The main property of the Binomial distribution is that trials are **independent** and **identical**, *i.e.* all trials have the same success probability. The variable X is an independent and identically distributed (*iid*) random variable.

Each of the n trials is essentially an independent Bernoulli trial – an experiment with the outcomes success or failure. This experiment is often illustrated as drawing balls with replacement from an urn containing two populations of balls, *e.g.* of black and red color. The Binomial distribution is the joint probability distribution of k successes after n independent Bernoulli trials. Thus, if $n = 1$ and $k \in \{0, 1\}$, the Binomial distribution is a Bernoulli distribution: $f(k|1, \theta) = \theta^k(1-\theta)^{1-k}$.

Alternatively to the Poisson framework in the previous section a researcher could test every observed read count x_i in m genomic regions for deviations from the expected value under a Binomial model. Here, the researcher could approximate the number of trials n by the total number of reads sequenced in the NGS experiment and the success probability θ as m^{-1}. Based on Equations (2.1) and (2.6) a binomial test can be performed with

$$\begin{aligned} P(X \geq x_i|n, \theta) &= 1 - \sum_{k=0}^{x_i-1} P(X = k|n, \theta) \\ &= 1 - \sum_{k=0}^{x_i-1} \binom{n}{k} \theta^k (1-\theta)^{n-k}. \end{aligned}$$

The Binomial framework yields similar results to the Poisson framework for large n and small θ with $\lambda = n \cdot \theta$. This approximation of the Binomial distribution is good for $n \geq 20$ and $\theta \leq 0.05$ which is usually true for NGS experiments.

The first two moments of the binomial distribution are

$$\begin{aligned} \mathrm{E}[X] &= \mu = n\theta \\ \mathrm{Var}[X] &= \sigma = n\theta(1-\theta) \end{aligned}$$

where, both, the μ and σ are not necessarily discrete numbers. It shows that variance σ is dependent on the mean μ, referred to as **heteroskedasticity**, which leads to a mis-specification of the second momemt, *i.e.* the variance. Hastie *et al.* [65] provides details on how Heteroskedasticity affects models that assume uncorrelated and uniform modeling errors, *e.g.* regression analysis or analysis of variance (ANOVA). NGS read count data are inherently heteroskedastic [50, 66] which makes the binomial distribution (among other distributions) an appropriate model for read counts which will be discussed in Chapter 3.

2.1.4 Sampling from Binomial Distributions

A probabilistic model is interpretable by means of sampling new observations from the underlying distribution, called “prediction”, and reasoning on the data, called “inference”. On the other hand, the model encodes assumptions about the data generation process which are needed for a meaningful statistical analysis. The generative process for discrete data is sometimes referred to as “sampling”. Sampling is useful for parameterizing a statistical model (section 2.2) and also determining confidence intervals (see [51] for details).

In discrete statistics, unrestricted sampling corresponds to the Poisson model whereas sampling from a fixed sample size n is described by the Binomial model. In Binomial sampling a fixed number of observations n are collected and classified according to a categorical given by Equation (2.6). For example in NGS data, one of n reads either originates from a certain region i, *i.e.* "success", or it does not, *i.e.* "failure". Thus, every region i provides a sample p of the true success probability θ. Because every binomial trial is independent the **"Rule for Sample Proportions"** applies which states that the distribution of these sample proportions can be approximated by a normal distribution given the numbers of successes k and failures $n - k$ are sufficiently large, *i.e.* $n \geq 10$ and $(n - k) \geq 10$. The sample distribution is approximately normal with $\mathrm{E}[p] = \theta$ and $\mathrm{Var}[p] = n \cdot \theta(1 - \theta)$. This property is one of the basic foundations of parameter inference which will be discussed in Section 2.2.

If the observations are according to more than two labels, say k, a generalization of Binomial sampling called **Multinomial Sampling** can be used to model the data. Let X be a random variable distributed under a Multinomial Model with : $X \sim \mathrm{Mult}(n, \theta)$,

$$f(x|\theta) = \frac{n!}{x_1!x_2!\ldots x_k}\theta_1^{x_1}\theta_2^{x_2}\ldots\theta_k^{x_k} \tag{2.7}$$

where n is fixed and $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ is a vector of population proportions. Each X_j is the count of occurrences for population j in the sample which itself has a binomial margin distribution. A nice example for a Multinomial is a ChIP-seq experiment: What happens if the pool of sequenced fragments is generated by multiple generative processes, *i.e.* two populations "ChIP-enriched" and "non-ChIP-enriched"? I will leave this question to be discussed in Chapter 3. For now, I will introduce mixture models as a model for a heterogeneous statistical population and, later, describe how sampling data is used to estimate an underlying model in Section 2.2.

2.1.5 Mixture Models

In many computational biology studies the measured data are of complex nature and composed of observations generated by different statistical processes. For example, ChIP-seq read count data is obtained for the genomic regions devoid of the antigen (background) and those bound by the antigen (signal). In consequence the statistical population contains two or even more sub-populations. If the researcher does not account for the latent sub-populations in his modeling efforts, the subsequent statistical inference most likely leads to false results. A sound formulation in those cases is achieved by mixture modeling.

A mixture model is a mathematical formulation to model hidden sub-populations within a data set in probabilistic terms. The mixture density for the total population consists of a weighted combination of component densities. The sub-populations in the data are not labeled and, thus, the unknown labels need to be inferred from the data itself.

Let X be a random variable that follows a mixture distribution with K mixture components each weighted by non-negative **mixing proportions** $\pi = (\pi_1, \ldots, \pi_K)$ and parametrized by $\theta = (\theta_1, \ldots, \theta_K)$: $X \sim MD(X|K, \pi, \theta)$. The probability mass function f is given by

$$f(x|K, \pi, \theta) = P(X = x|K, \pi, \theta) = \sum_{k=1}^{K} \pi_k P_k(X = x|\theta_k),$$

where $0 \leq \pi_i \leq 1$ and $\sum_k \pi_k = 1$. The parameters π and θ can be estimated with techniques that are described in Section 2.2.2, *e.g.* Maximum Likelihood Estimation.

A desired key characteristic of mixture distributions is the increase in the variance of the overall population. Uncertainty in θ causes a further increase in the unconditional variance of the overall mixture model. Note that a mixture distribution can contain mixture components of various parametric families. The herein described mixture models are distinct from convolutional models where one observation is a combination of multiple underlying random variables.

Many NGS data analysis problems require a classification of genomic regions into distinct classes. Mixture models fulfill this task by means of the assignment of labels to each data point. The model provides an estimate $\hat{r}_{ij}$ of the probability that an observation i belongs to a component j, called **responsibility**,

$$\hat{r}_{ij} = \frac{\pi_j P_j(X = x_i|\theta_j)}{\sum_{k=1}^{K} \pi_k P_k(X = x_i|\theta_k)}, \tag{2.8}$$

which is also called "posterior probability". Based on the probability defined by Equation (2.8) every observation i can be assigned to the latent (hidden) component z_i that generated i most likely via

$$z_i = \underset{k \in 1, \ldots, K}{\arg\max}\, \hat{r}_{ik},$$

where $z \in \{1, \ldots, K\}$.

The explicit modeling of K sub-populations facilitates statistical inference based on only one component of the mixture model. For example, a researcher could test if an observation i originates from component k via $P_k(X \geq x_i|\theta_k)$ (see Equation (2.1)). In this case H_0 is described by one component of the mixture model, *e.g.* an assumed background component.

When dealing with mixture models there shall exist a unique characterization for any model in the family being considered, referred to as **identifiability**. Formally, a model $f(X|\theta)$ is identifiable if

$$\theta_1 \neq \theta_2 \text{ implies } p(X|\theta_1) \neq p(X|\theta_2).$$

It follows that a model is non-identifiable if there are subspaces of the parameter space where the family is not identifiable, *e.g.* $\pi_1 = 1$. Identifiability is closely linked to the concept of parameter estimation and sufficient statistics which are described in the following section.

2.2 Model Parameter Estimation

In Section 2.1.1 the Poisson model for read counts was estimated simply by the arithmetic mean $\bar{x}$ of counts x in all genomic regions. Such an ad-hoc method is not always applicable. Model parameter estimation aims to fit a model based on the made observations. Here, I present the concept of sufficient statistics and describe its close ties to parameter inference.

2.2.1 Sufficient Statistics

The mean is the simplest sample statistic on the data to summarize the information contained in the sample into a single numerical value. No further information contained in the sample is needed to parametrize the Poisson model. This is referred to as a **Sufficient Statistic** – a concept which was introduced by the biologist and statistician Ronald Fisher in 1922 [67]:

> "A statistic satisfies the criterion of sufficiency when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter to be estimated."

Let X be a set of independent identically distributed (*iid*) variables conditional on a parameter θ, a statistic $T(X)$ is sufficient for θ if the probability density function $f_\theta(X)$ depends *solely* on X through $T(X)$. Formally **Fisher's factorization theorem** [67] describes this relation by

$$f_\theta(X) = h(X)g_\theta(T(X)), \tag{2.9}$$

where $g_\theta(\bullet)$ and $h(\bullet)$ are nonnegative functions. A statistic $T(x)$ can now be tested for sufficiency by testing $h(x)$ for independence of θ via $h(x) = \frac{f_\theta(x)}{g_\theta(t)}$. From the factorization theorem it follows that for two observations x_1 and x_2 the estimate $\hat{\theta}$ is identical if $T(x_1) = T(x_2)$.

The Poisson distribution is a distribution of the exponential family which is a prevalent set of probability distributions including also the Normal or Binomial distribution. For every exponential family distribution with parameter θ its probability mass function f can be written as

$$f_\theta(X|\theta) = h(X)\ \exp(\eta(\theta)\ T(X) - A(\theta)),$$

where $T(\bullet)$ is a sufficient statistic, $\eta(\bullet)$ is called the "natural parameter function" and $A(\bullet)$ is called the "log-partition function". When this form is compared to Equation (2.9) it becomes evident that it is a readily applicable framework for sufficient statistics. In fact, for every exponential family distribution there exist sufficient statistics. Below I derive a sufficient statistic for the Binomial distribution because it is relevant for the following chapters of this thesis.

A Sufficient Statistic for the Binomial Family

The Binomial distribution family is an **exponential family distribution** described by two parameters, namely the number of trials n and the success probability θ. In most cases n is fixed and known and only θ has to be estimated.

In contrast to the Poisson distribution the arithmetic mean $\bar{x}$ is not a sufficient statistic for θ. The Factorization theorem shows that $h(x)$ still depends on θ for $T(x) = t = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $g_\theta(t) = f_\theta(t)$:

$$\begin{aligned}
h(x) &= \frac{f_\theta(x)}{g_\theta(t)} \\
&= \frac{\prod_{i=1}^{m}\binom{n}{x_i}\theta^{x_i}(1-\theta)^{n-x_i}}{\binom{n}{t}\theta^{t}(1-\theta)^{1-t}} \\
&= \frac{\binom{n}{x_1}\theta^{x_1}(1-\theta)^{n-x_1}\cdot\binom{n}{x_2}\theta^{x_2}(1-\theta)^{n-x_2}\cdot\ldots\cdot\binom{n}{x_m}\theta^{x_m}(1-\theta)^{n-x_m}}{\binom{n}{t}\theta^{t}(1-\theta)^{1-t}} \\
&= \frac{\binom{n}{x_1}\cdot\binom{n}{x_2}\cdot\ldots\cdot\binom{n}{x_m}\cdot\theta^{\sum x}(1-\theta)^{m-\sum x}}{\binom{n}{\frac{1}{m}\sum x}\theta^{\frac{1}{m}\sum x}(1-\theta)^{1-\frac{1}{m}\sum x}} \\
&= \binom{n}{x_1}\cdot\binom{n}{x_2}\cdot\ldots\cdot\binom{n}{x_m}\cdot\theta^{\frac{m-1}{m}\sum x}(1-\theta)^{m-1-\frac{m-1}{m}\sum x}.
\end{aligned}$$

On the other hand the sum of all successes in all observed trials $T(x) = t = \sum_{i=1}^{m} x_i$ is a sufficient statistic for θ because the sum of the successes is also distributed as a binomial $t \sim \mathrm{Bin}(mn, \theta)$. In this case it can be shown that $h(x)$ is independent of θ

$$\begin{aligned}
h(x) &= \frac{f_\theta(x)}{g_\theta(t)} \\
&= \frac{\prod_{i=1}^{n}\binom{n}{x_i}\theta_i^{x}(1-\theta)^{n-x_i}}{\binom{mn}{t}\theta^{t}(1-\theta)^{n-t}} \\
&= \frac{\binom{n}{x_1}\theta_1^{x}(1-\theta)^{n-x_1}\ldots\binom{n}{x_m}\theta_m^{x}(1-\theta)^{n-x_m}}{\binom{mn}{t}\theta^{t}(1-\theta)^{n-t}} \\
&= \frac{\binom{n}{x_1}\ldots\binom{n}{x_m}}{\binom{mn}{t}}.
\end{aligned}$$

Sufficient Statistics for Mixtures of Exponential Family Distributions

In Section 2.1.5 Mixture Models were introduced. Because every exponential family distribution has sufficient statistics they are mathematically amenable and commonly used as mixture components. In consequence, the log density of a K-mixture model parameterized by ν can be expressed as

$$\log p(x|\nu) = \eta_x + \sum_k t_k(x)\nu_k, \tag{2.10}$$

where η_x is a normalization constant and t_k is a sufficient statistic for component k. It can be seen that $\log p(x|\nu)$ depends on ν *solely* through t which makes t a sufficient statistic for $p(x)$. Details are given in Bishop, 2007 [68]. In the next sections parameter inference based on sufficient statistics will be explained.

2.2.2 Maximum Likelihood Estimation

In some cases simple statistics on the data can yield a convenient model parametrization. However in more sophisticated cases like mixture models a function can be defined to compute *the probability of the measured data given a certain "model realization"*, referred to as **"likelihood"**. Let X be an *iid* random variable with observations $x = x_1, \cdots, x_m$ under a parametric model defined by θ: $x_i \sim f_\theta(x)$. The function $f_\theta(x_i)$ describes how likely x_i is observed given θ. On the other hand, if x_i is fixed, $f_\theta(x_i)$ describes how likely a model defined by θ could give rise to x_i. The likelihood $\mathcal{L}$ for a parametrization θ is given by the **likelihood function**

$$\mathcal{L}(\theta|x) = \prod_{i=1}^{m} f_\theta(x_i).$$

The **log-likelihood** is defined by

$$\begin{aligned} \log \mathcal{L}(\theta|x) = \ell(\theta) &= \log \prod f_\theta(x) \\ &= \sum_{i=1}^{m} \log f_\theta(x_i) \\ &= \sum_{i=1}^{m} \ell(\theta|x_i) \end{aligned} \tag{2.11}$$

where $\ell(\theta|x_i) = \log f_\theta(x_i)$ is called a "log-likelihood component".

Essentially, for exponential family distributions, the log-likelihood function is a θ-linear combination of the *sufficient statistics* of the model. This results in a mathematically manageable analysis by partial derivatives. To estimate a parameter θ_j the derivative of the *log-likelihood* with respect to θ_j is set to 0 in

$$\frac{\partial}{\partial \theta_j}\ell = \frac{\partial}{\partial \theta_j} \log f_{\theta_j}(x_i) \overset{!}{=} 0. \tag{2.12}$$

Conveniently this derivative can be solved analytically, referred to as a **closed form solution** (see [68] for details).

The method of **maximum likelihood estimation** (MLE) (see, for example, [65]) finds a value $\theta = \hat{\theta}$ that maximizes Equation (2.11) in the parameter space by the criterion

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta|X).$$

Alternatively, by setting the Derivative (2.12) to 0 stationary points can be identified with

$$0 = \frac{\partial}{\partial\hat{\theta}}\ell = \frac{\partial}{\partial\hat{\theta}} \log f_{\hat{\theta}}(x_i).$$

The extremum estimator is consistent, *i.e.* it converges to the true value θ_0 for infinitely large sample sizes. In NGS experiments millions of data points are generated which makes MLE a method of choice. However, there exist more or less similar alternatives: If the sample size is small **Bayesian inference** can encode prior information into maximimum *a posteriori* estimation (detailed in [69]). If used with a non-informative prior distribution this estimation is essentially analogous to MLE. Another simple alternative called **Bootstrapping** [65] allows for parametric and also model-free estimation. It is based on resampling of observations which makes it less consistent than MLE or Bayesian inference.

MLE for Mixture Models

Mixture modeling is a missing data problem where the sub-population membership of the data points is unknown. The latent (hidden) labels of data points have to be estimated together with the parameters of K distributions in the mixture distribution. Denote the *iid* random variable by X which follows a K-mixture distribution with mixing proportions π and parameter θ. Let K be fixed and $\nu = \{\pi, \theta\}$ the log-likelihood is defined by

$$\begin{aligned}\ell(\nu) &= \sum_{i=1}^{N} \ell(\nu|x_i) \\ &= \sum_{i=1}^{m} \log \left\{ \sum_{k=1}^{K} \pi_k f_{\theta_k}(x_i) \right\}.\end{aligned}$$

Next, the derivative with respect to a parameter θ_j is

$$\begin{aligned}\frac{\partial}{\partial\theta_j}\ell &= \sum_{i=1}^{m} \frac{1}{\sum_k \pi_k p(x_i|\theta_k)} \pi_j \frac{\partial}{\partial\theta_j} p(x_i|\theta_j) \\ &= \sum_{i=1}^{m} \underbrace{\frac{\pi_j p(x_i|\theta_j)}{\sum_k \pi_k p(x_i|\theta_k)}}_{w_{ij}} \frac{\partial}{\partial\theta_j} \log p(x_i|\theta_j).\end{aligned} \tag{2.13}$$

Note that this derivative of the log-likelihood is simply the ordinary likelihood of the parametric model of component j weighted by w_{ij} (see Equation 2.12). This is an advantageous feature of mixture models for maximum likelihood estimation as will be discussed in the next section.

2.2.3 The Expectation-Maximization Algorithm

Sometimes the likelihood equations can not be solved directly. For example, hidden sub-population memberships in mixture modeling render the MLE intractable, *i.e.* X is dependent on a latent discrete variable Z. The popular Expectation-Maximization (EM) algorithm [70] represents an elegant solution for MLE when the model depends on hidden variables. The goal is to maximize the likelihood function of the parameter ν for Z given the variables X

$$\ell(\nu) = \log p(X|\nu) = \sum_{Z} \log p(X, Z|\nu),$$

where Z is unknown and has to be inferred from the data (see below). The EM algorithm is an iterative method that rotates between two steps: In the **expectation step** the algorithm calculates the expectation of the log-likelihood for the current parameter estimate. In the following **maximization step** the parameter ν is updated to maximize the expected log-likelihood. The algorithm terminates after a given number of steps or if the likelihood can not be further increased subject to a number $\varepsilon > 0$.

The maximization problem in EM amounts to iterative updates of ν to achieve a step wise improvement on the likelihood. However, in this missing data problem the true value of Z is unknown and can only be inferred based on the knowledge of X and ν. Formally, the posterior distribution $p(Z|X, \nu)$ provides an expected value for Z which is used to estimate ℓ in the expectation step. At iteration t, the *expectation* $\dot{\ell}$ can be calculated conditional on $\nu^{(t-1)}$ and X via the objective function

$$\dot{\ell}(\nu, \nu^{(t-1)}) = \sum_{Z} p(Z|X, \nu^{(t-1)}) \log p(X, Z|\nu). \tag{2.14}$$

In the maximization step, $\nu^{(t)}$ is calculated by *maximizing* $\dot{\ell}$

$$\nu^{(t)} = \arg\max_{\nu} \dot{\ell}(\nu, \nu^{(t-1)}). \tag{2.15}$$

In the beginning ν is initialized with arbitrary values $\nu^{(0)}$. More detail on the EM algorithm is provided in [68].

The EM algorithm proves useful and performant especially for distributions of the exponential family: Sufficient statistics have to be added up to calculate the expectation, such that the maximization is done on a linear function. In consequence closed form updates at each iteration can be formulated which speeds up the EM algorithm substantially. A major drawback of the EM method constitutes its convergence to local maxima – dependent on the initial parameter $\nu^{(0)}$. A local maximum is not necessarily the maximum likelihood estimator if the likelihood distribution is multimodal which is usually the case for mixture models [71]. However, results of multiple instances with different initialization can be compared with respect to $\dot{\ell}$. Alternatively to EM, Markov Chain Monte Carlo performs a posterior sampling via Bayes' theorem but is susceptible to non-identifiability.

The EM Algorithm for Mixture Models

Section 2.2.2 described an advantageous property in MLE for mixture models: The log-likelihood derivative given by Equation (2.13) with respect to θ_j is simply a weighted ordinary likelihood of the model component j. It follows that, if the model components are from the exponential family, closed form updates are possible. The problem remains that the weights w_{ij} depend themselves on the latent mixing proportions π and the parameter θ. The EM algorithm calculates an expected value for π_j by the posterior probability $P(Z = j|X = x, \pi, \theta)$ which breaks the cycle.

The EM algorithm for mixture models is initialized with the parameters $\theta^{(0)} = \theta_1^{(0)}, \dots, \theta_K^{(0)}$ and $\pi^{(0)} = \pi_1^{(0)}, \dots, \pi_K^{(0)}$. At every iteration t the weights w_{ij} are computed based on $\nu^{(t-1)}$ via

$$w_{ij} = p(Z = j|X = x_i, \pi, \theta) = \frac{\pi_j^{(t)} p(x_i|\theta_j^{(t)})}{\sum_k \pi_k^{(t)} p(x_i|\theta_k^{(t)})}.$$

Based on w_{ij} the log-likelihood is maximized

$$\ell(\theta) = \sum_{i=1}^{N} \sum_{j} w_{ij} \log p(x_i|\theta_j).$$

The EM algorithm iterates until no further improvement in the log-likelihood ℓ is achieved, *i.e.* $\Delta\ell \leq \epsilon$.

Part II

Normalization of NGS Read Count Data

The robust normalization method described in this part was developed by me and Ho-Ryun Chung from the Otto-Warburg-Laboratory: Epigenomics at Max Planck Institute for Molecular Genetics in collaboration with researchers from the Max Planck Institute of Immunobiology and Epigenetics, Freiburg; the Leibniz Research Centre for Working Environment and Human Factors, TU Dortmund, Dortmund; the Department of Genetics and Epigenetics, University of Saarland, Saarbrücken and the Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel. A preprint is published in bioRxiv [2] as

Johannes Helmuth, Na Li, Laura Arrigoni, Kathrin Gianmoena, Cristina Cadenas, et al. normR: Regime Enrichment Calling for ChIP-seq Data. *bioRxiv*, page http://dx.doi.org/10.1101/082263, October 2016. `http://biorxiv.org/content/early/2016/10/25/082263`

Chapter 3

The normR Framework – Robust Normalization of Read Count Data with Mixture Models

This chapter presents the "normR Framework" – a data-driven computational framework to normalize read count data with a binomial mixture model. The model accounts for the effects on the overall read statistics caused by the presence of signal, *e.g.* read accumulations in certain regions, and technical biases, *e.g.* sequencing depth. In this thesis the "normR" approach is mainly applied to the analysis of protein binding sites from ChIP-seq read count data, yet it is not limited to this type of NGS read counts, but can also be applied to RNA-seq, DNaseI-seq or ATAC-seq data.

3.1 Motivation

Section 1.2 described how Next Generation Sequencing (NGS) based techniques can be used to measure distinct molecular properties in a population of cells. For example in ChIP-seq, antibodies are used to preferentially enrich for specific protein-DNA complexes. Therein, the probability of selecting a DNA fragment depends on the presence or absence of the protein of interest. In this chapter I will describe how sequencing can be seen as a sampling process where the final pool of fragments contains *quantitative* information on the investigated molecular property across the genome. In ChIP-seq, this quantity is the genome-wide DNA binding pattern of a protein. The spatial distribution of the sampled fragments is estimated by mapping the reads to the genome which results in a characteristic read count pattern. For example, more ChIP-seq reads are observed at genomic loci bound by the protein than at loci devoid of the protein. However, some ChIP-seq reads still map to the regions devoid of the protein because the ChIP enriches rather than selects for the protein-containing fragments.

The analysis of the read count patterns involves the *in silico* identification of genomic regions characterized by a biological signal of interest – a task that requires the discrimination of signal from background. Intuitively, the "signal-regions" should be characterized by a read count which is greater than the read count in "background-regions". For example in RNA-seq, a signal could be the elevated transcription of a gene X in diseased subjects if compared to a healthy control whereas other genes remain unchanged. An increased number of mRNA transcripts results in more RNA-seq reads accumulating in the exons of gene X (signal-regions). The identification of this conditionally differential transcription is also referred to as **"difference calling"**. However, the identity of the signal-regions and, ultimately, the expected read enrichment therein is unknown *a priori*, *i.e.* the data are not labeled.

A meaningful interpretation of read count patterns requires the mitigation of the effects of technical biases which influence the expected read counts. These biases arise, for example, from copy number variations, sequencing biases or mapping ambiguities (reviewed in [72, 73]). Read counts obtained in a control sequencing run without specific signals are one way to correct for biases in the experiment. The adjustment of the read count pattern to the control, however, requires a **"normalization"** to account for differences in sequencing depths and the presence of signal in the experiment. To this extent, a normalization factor can correct the average ratio between the control and the experiment. Most importantly, this factor has to be estimated based on background-regions *only*, since an estimation based on all genomic regions results in a bias towards the prevalent read enrichment in the signal-regions. In other words, the effect of the read accumulations in signal-regions on the overall read statistics has to be accounted for. If this effect is not taken into account, the normalization and, consequently, the difference calling suffer from low sensitivity (see below). Thus, a proper normalization requires the identity of background-regions.

The two tasks of difference calling and normalization in read count data are mutually dependent (Figure 3.1): On the one hand, the discrimination of signal- and background-regions requires normalization to account for technical artifacts but, on the other hand, the normalization requires the knowledge of the regions that remained unchanged, *i.e.* background-regions. In consequence, normalization and difference calling are inseparable – they are two faces of the same coin.

In my thesis I illustrate and tackle this interlinked dependency in the problem of the identification of protein binding events from ChIP-seq read count patterns. In a ChIP-seq experiment, a population of chromatin fragments is obtained by the sonication of the chromatin. Next, an-

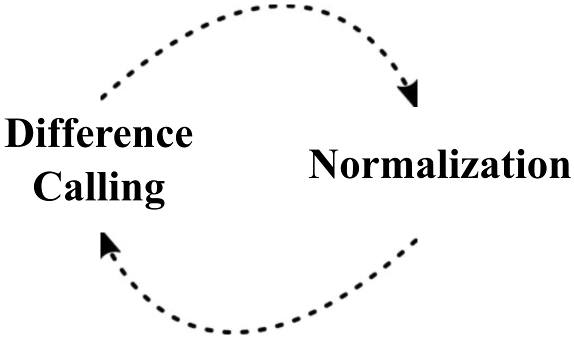

Fig. 3.1 – The Inter-Dependency of Difference Calling and Normalization. Difference Calling and Normalization in the analysis of read count patterns are mutually dependent and interlinked. On the one hand, Difference Calling requires normalization to discriminate signal- from background-regions and, on the other hand, a proper normalization requires the knowledge of background-regions.

tibodies are used to enrich for fragments carrying a protein of interest (see Section 1.2.2 for details). The signal-regions correspond to genomic loci bound by the protein which is reflected by an accumulation, *i.e.* **"enrichment"**, of ChIP-seq reads. Because the ChIP only enriches rather than selects for protein containing fragments, the probability of observing a read at genomic loci devoid of the protein is *low but not zero*. Those loci represent the background-regions. Bearing this in mind, I will show how a ChIP-seq experiment can be seen as multinomial sampling trial where the average read enrichment in signal-regions affects the average read count in background-regions (Section 3.2.1). Moreover, the more regions are enriched by the ChIP, the lower the signal-to-noise ratio (S/N) becomes at a fixed sequencing depth (Section 3.2.2).

The discrimination of enriched regions from background-regions in ChIP-seq read count patterns is an unsupervised learning problem, referred to as **"enrichment calling"** (sometimes also "peak calling"). Enrichment calling depends on an appropriate normalization with respect to a control to mitigate aforementioned biases. The ChIP-seq control is obtained, for example, by sequencing the sonicated chromatin without specific enrichment (Input). A correct normalization factor should be estimated based *only* on the regions devoid of the protein. The normR approach uses a mixture model (Section 2.1.5) to break the mutual dependence as I will outline in the next section. A detailed comparison to previously developed methods [74–80] in Chapter 4 illustrates the superior performance of the normR approach.

3.2 The normR Approach

Based on the thoughts on NGS read count patterns above I developed a robust and extendable framework for joint normalization and difference calling, called **"normR"** (recursive acronym: *"normR obeys regime mixture rules"*). Here, I explain the normR approach by taking the example of ChIP-seq data enrichment calling. Firstly, I show that a ChIP-seq experiment relates to

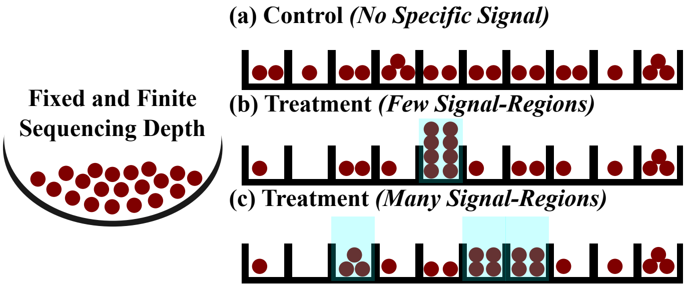
(a) Control *(No Specific Signal)*
Fixed and Finite Sequencing Depth
(b) Treatment *(Few Signal-Regions)*
(c) Treatment *(Many Signal-Regions)*

Fig. 3.2 – A Sequencing Experiment Constitutes a Multinomial Sampling Trial. A finite number of sequenced reads are generated in a sequencing experiment (left; 20 reads depicted as red balls) and mapped to the genome (right). The number of reads is usually quantified in fixed-size genomic bins, exemplified as 10 black buckets. (a) If there are no signal-regions present, all regions are background-regions and their expected read count is 2. (b) In few signal-regions (blue overlay; 1 bucket, 10% of genome) the expected read count is high (8) but it is decreased in background-regions ($\sim$1.3). (c) With more signal-regions (3 buckets, 30% of genome) the expected read count in background-regions is $\sim$1.3 but the expected read count in signal-regions is decreased ($\sim$3.7).

sampling of chromatin fragments which can be seen as a multinomial sampling trial where read counts in background- and signal-regions are inter-dependent. Secondly, I discuss how the overall number of signal-regions affects the S/N and how statistical power can be increased in low S/N settings. Finally, the normR model is described. It uses a binomial mixture model which, in its simplest incarnation, uses two mixture components corresponding to background B and enrichment E. normR [4] was implemented as an R package [81, 82] with performance-critical routines optimized for performance as C++ code.

3.2.1 Sequencing is a (Multinomial) Sampling Trial

During a ChIP experiment a population of chromatin fragments are obtained by sonication of chromatin. Antibodies preferentially bind chromatin fragments that carry an antigen of interest. Note, these antibodies bind not exclusively antigen-DNA complexes – ChIP only enriches rather than selects antigen containing chromatin fragments. More specifically the probability to draw a fragment depends on the presence or absence of an antigen. If present, the probability is high, if absent, the probability is *lower but not zero*. The spatial distribution of these sampled fragments is then estimated by end-sequencing and mapping the corresponding reads to the genome (see Section 1.2). The sequencing of the ChIP library is a multinomial sampling process (described in Section 2.1.4) which induces dependencies between the regions: As the total number of reads obtained from one sequencing run is fixed and finite, the increase of reads in some regions due to the ChIP enrichment (*antigen present*; “signal-regions”) leads to a decrease in all remaining regions (*devoid of antigen*; “background-regions”). Figure 3.2 illustrates this idea for proteins with few or numerous binding sites across the genome. An adequate normalization has to take into account this inter-dependency of read coverage.

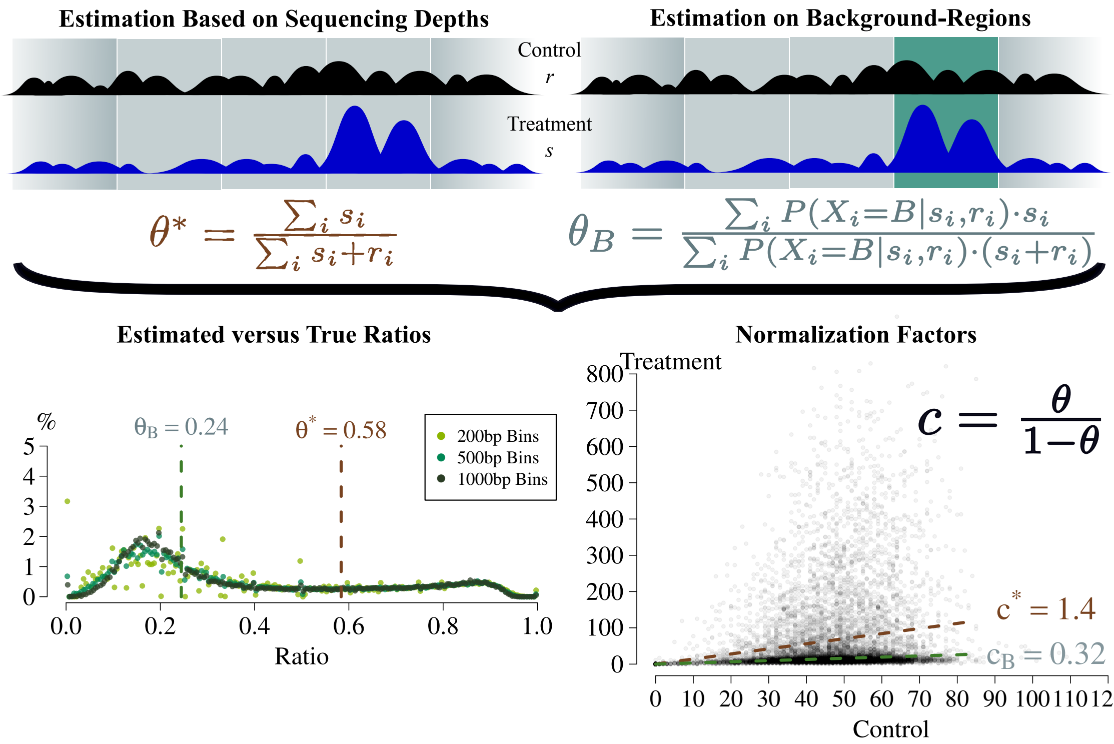

Fig. 3.3 – The Sequencing Depth Ratio Is a Background Estimation Biased towards Signal-Regions. The naïve background estimation based on the sequencing depth ratio θ^* is shown in the top left panel. The herein proposed background estimation based on the sequencing depth ratio θ_B is illustrated in the top right panel. θ_B is the ratio of sequencing depths weighted by the probability of each region to be background. The ratio $\frac{s_i}{s_i+r_i}$ of read counts in treatment s_i and control r_i (lower left panel) shows a characteristic bimodality representing background-regions (mode≈ 0.19) and signal-regions (mode≈ 0.9). When compared to θ^*, θ_B improves on the estimation of the background population. Consequently, the inferred normalization factor c_B approximates the read coverage in background-regions well (lower right panel).

To infer signal-regions the read densities obtained by ChIP-seq experiment are compared to the corresponding counts obtained by a control experiment, *e.g.* by sequencing the sonicated chromatin (Input). This approach addresses some systematic biases, like copy number variations, sequencing biases, mapping ambiguities or chromatin structure [72,83,84]. A region i should be called “enriched by ChIP” *only if* the number of reads from ChIP s_i is substantially greater than expected given the number of reads from Input r_i. To this extent, a normalization factor c_B is required to define a statistically sound Null hypothesis to test whether the observed ChIP read counts are significantly greater than expected given the control.

A simple example illustrates the dependency of c_B on the outcome of the ChIP experiment: If twice as many reads are sequenced in the ChIP than in the control, the read counts per region in the ChIP are on average expected to be twice as high as in the Input, *i.e.* the ratio of sequencing depths is 2. In fact, with no specific enrichment, the normalization factor c_B is $\sim$2, *i.e.* the number of reads in ChIP-seq s_i is twice the number in the Input r_i for every region i. However, with specific ChIP enrichment, the normalization factor c_B is ≤ 2 because it depends on the average ChIP enrichment and, also, on the total number of signal regions. In particular, c_B shrinks as the number of signal-regions and the level of enrichment in those regions increases. In this case a naïve ratio of sequencing depths is insufficient to estimate c_B correctly. Figure 3.3 illustrates how the ratio θ_B can be estimated based on a probability weighted sequencing depth ratio. The estimation of the normalization factor $c_B = \frac{\theta}{(1-\theta)}$ requires the identity of background-regions, albeit their identification requires normalization itself. Again this illustrates that normalization and the identification of signal-regions are two sides of the same problem – this problem is the motivation of the normR approach.

Previously developed approaches estimate c_B either by the ratio of sequencing depths [74, 76], by the ratio of ChIP- and control read counts summed over *ad hoc*-chosen background-regions with fixed width [77, 78, 80], or by the data-driven identification of background-regions [85, 86]. A detailed comparison of their performance to normR's c_B estimation is given in Chapter 4.

3.2.2 Deliberations on the Signal-to-Noise Ratio (S/N)

Apart from the inter-dependency of difference calling and normalization, a low S/N can lead to low power in statistical analysis of a sequencing experiment and, thus, reliable assertions about the signal-regions can not be made, *e.g.* protein binding sites in a ChIP-seq experiment. In sequencing experiments, the S/N depends mainly on two factors: the fraction of signal-regions and the overall number of sequenced reads. Figure 3.2 illustrates that the more signal-regions are present, the lower becomes the S/N at a fixed sequencing depth N (also reviewed in [87]).

The concept of S/N relates to the statistical concept of "effect size" which measures the strength of a phenomenon. The standardized effect size is the difference in means of s and r standardized by the pooled standard deviation given by

$$d = \frac{\bar{s} - \bar{r}}{\sigma}, \tag{3.1}$$

where $\sigma = \sqrt{\frac{(N_s-1)\sigma_s^2+(N_r-1)\sigma_r^2}{N_s+N_r-2}}$ is the pooled standard deviation for two independent samples of size N_s and N_r. This quantity is also referred to as "Cohen's d" [88] and relates to the concept of S/N. Note that with increasing sample size N the sample average $\bar{s}$ converges to the expected value $\mathrm{E}[s] = \mu_s$ and the sample variance σ^2 decreases, referred to as the "Law of Large Numbers" (detailed in [89]).

The effect size d forms a closed system of statistical power together with the significance level α and the sample size N (Figure 3.4). The more liberal α and/or the greater N and/or the

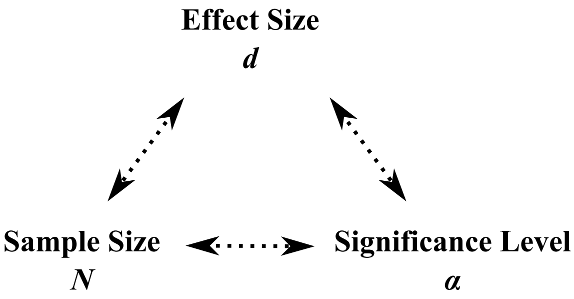

Fig. 3.4 – The Closed System of Statistical Power. The more liberal α and/or the greater the sample size N and/or the greater the observed effect d, the more likely a significant effect is detected.

greater d, the more likely a significant effect is detected, *i.e.* the more powerful the test. Assume α is held fixed to some value, say 0.05, there exist two set screws to improve the power of the test for a sequencing-based study, namely d and N: The effect size d may be increased by increasing the signal strength. For example, a more specific antibody could be used in the ChIP. However, an antibody with improved affinity for the antigen is not always available. On the contrary, the sample size N relates to the expected read counts per region which can easily be increased by sequencing deeper. The decreasing cost of NGS experiments allows for a convenient increase of the sequencing depth to boost statistical power in low S/N sequencing data (*i.e.* numerous signal-regions). Yet, it persists the need of a statistical sound null hypothesis based on a proper normalization factor c_B.

3.2.3 The normR Method

The normR method tackles the described problem of inter-dependency of difference calling and normalization by performing both tasks simultaneously to identify genomic regions harboring a statistically relevant signal. To this extent, it models the read counts from treatment, *e.g.* ChIP-seq, and control, *e.g.* Input, as a binomial mixture model. Given two vectors of integers r (control) and s (treatment) of identical length, normR models the read counts from the treatment s and control r by a binomial m-mixture model:

$$\begin{aligned} k_i &\sim \text{Categorical}(\pi) \\ N_i = s_i + r_i | k_i = j &\sim \text{Bin}(N_j, \theta_j) \end{aligned} \tag{3.2}$$

with $i = 1, \ldots, n$ and $\sum \pi_j = 1; \pi_j \in [0, 1]; j = 1, \ldots, m$. Given this model, normR follows a two step procedure: (i) A binomial m-mixture model is fit by the expectation maximization (EM) algorithm [70] (refer to Section 2.2.3) using the likelihood function,

$$\mathcal{L} = P(\pi, \theta, N_i | s_i, r_i) = \prod_{i=1}^{n} \binom{N_i}{s_i} \sum_{j=1}^{m} \pi_j \cdot \theta_j^{s_i} \cdot (1 - \theta_j)^{r_i}; \tag{3.3}$$

and (ii) each (r_i, s_i) is tested for significance against a fitted background component to label signal-regions.

In a preprocessing stage, the vectors r and s are filtered for entries where $r = s = 0$ because no assertion about their signal state can be made. Secondly, a hash of unique (r_i, s_i) tuples is created which improves run time substantially. Because there exist only a fixed number of discrete tuples and many tuples are observed multiple times, this approach vastly reduces the number of computations needed.

In the first step of the algorithm, the EM algorithm is run with initial values π sampled from $\mathrm{U}(0, 1)$ and θ sampled from $\mathrm{U}(0.001, \theta^*)$ where θ^* denotes the ratio of sequencing depth ratios (Section 3.2.1, Figure 3.3). Upon convergence of the EM algorithm, *e.g.* $\Delta\mathcal{L} \leq \epsilon = 0.001$, the background component B is determined to be the component with θ_B that is the smallest of $\{\theta_1, \dots, \theta_m\}$. By default the EM algorithm is run 10 times to find the parametrization with greatest $\mathcal{L}$.

In the second step, every region i is tested for significance against the fitted background component B with

$$P_i = P(s_i \geq x | N_i, \theta_B) = 1 - \sum_{k=0}^{s_i - 1} \binom{N_i}{s_i} \theta_B^{s_i} \cdot (1 - \theta_B)^{N_i - s_i}.$$

Obtained P-values are transformed to q-values for FDR correction [56] using the T method as described in Section 2.1.2. The applied T method ensures tests are performed only for observations with sufficient statistical power to make reliable assertions. Note that by nature the binomial mixture model assumes the independence between regions which is valid for a sufficiently large bin size that is greater than the average fragment size.

In a last step, the regularized and normalized enrichment based on the fitted background B is calculated for every region i. To account for noise in low count regions, s_i and r_i are adjusted by adding model-derived pseudo counts. Given the posterior probability for the background component B

$$P(X_i = B | s_i, r_i) = \frac{\pi_B \cdot \theta_B^{s_i} \cdot (1 - \theta_B)^{r_i}}{\sum_{j=1}^{m} \pi_j \cdot \theta_j^{s_i} \cdot (1 - \theta_j)^{r_i}}$$

the pseudo counts are taken to be the average read counts in Input α_r and ChIP α_s defined by

$$\alpha_r = \frac{\sum_{i=1}^{n} P(X_i = B | s_i, r_i) \cdot r_i}{\sum_{i=1}^{n} P(X_i = B | s_i, r_i)} \text{ and}$$
$$\alpha_s = \frac{\sum_{i=1}^{n} P(X_i = B | s_i, r_i) \cdot s_i}{\sum_{i=1}^{n} P(X_i = B | s_i, r_i)}.$$

The regularized enrichment e^* is calculated with

$$e_i^* = \log\left(\frac{s_i + \alpha_s}{r_i + \alpha_r} \cdot \frac{\alpha_r}{\alpha_s}\right),$$

where (α_r/α_s) regularizes e_i^*, *i.e.* shifts to 0, for background-regions. To account for the average signal in a component $j \neq B$, e_i^* is normalized by the "enrichment factor" $\langle f_j \rangle$, *i.e.* the average fold enrichment, given by

$$\langle f_j \rangle = \frac{\theta_j}{1-\theta_j} \cdot \frac{1}{c_B},$$

with $c_B = \frac{\theta_B}{1-\theta_B}$ which represents the fitted background normalization factor. The regularized and normalized enrichment $e_i^{(j)}$ is then obtained via

$$e_i^{(j)} = \frac{e_i^*}{\log\langle f_j \rangle}.$$

The normR enrichment can be used as a background normalized signal estimate in downstream analyses or for visualization as described in the following chapters of this book.

In its simplest incarnation normR has two components, *i.e.* $m = 2$, representing background B and signal E, *e.g.* ChIP-seq enrichment over Input. In this setting the model has three free parameters, *i.e.* θ_B, θ_E and π_B. θ_B and θ_E are the expected fraction of reads in the ChIP-seq over the sum of reads from ChIP and Input per region for the background-regions and signal-regions, respectively. π_B is the proportion of regions that belong to the background B. The proportion of signal-regions π_E is simply $(1 - \pi_B)$. From the deliberations in the previous sections it follows that, for the true background θ_B, $\theta_B \leq \theta^*$ where θ^* denotes the expected fraction of reads from ChIP taking into account only sequencing depth differences (Fig. 3.3). In the case of no enrichment one has $\pi_B = 1$ and $\theta_B = \theta^*$. The definition of regions is the last "implicit" parameter, *e.g.* promoter flanking regions or fixed width tiling windows across the whole genome. This approach is detailed in Section 4.2.2 where I show the applicability and performance of the normR approach for calling ChIP-seq enrichment.

3.2.4 Why Not Use a Negative Binomial or Multinomial Distribution?

The Negative Binomial distribution models the number of successes in a sequence of *iid* Bernoulli trials before a specified number of failures occurs (see also [89]). Denote a random variable X counting the number of trials n given s successes with a success probability θ one has

$$f(n; s, \theta) = P(X = n) = \binom{n-1}{s-1}\theta^s(1-\theta)^{n-s}.$$

Expect for the binomial coefficient, this density is equivalent to the Binomial distribution, *i.e.* $\theta^s(1-\theta)^{n-s}$. Recently, numerous studies [46, 49, 66] have shown the applicability of this distri-

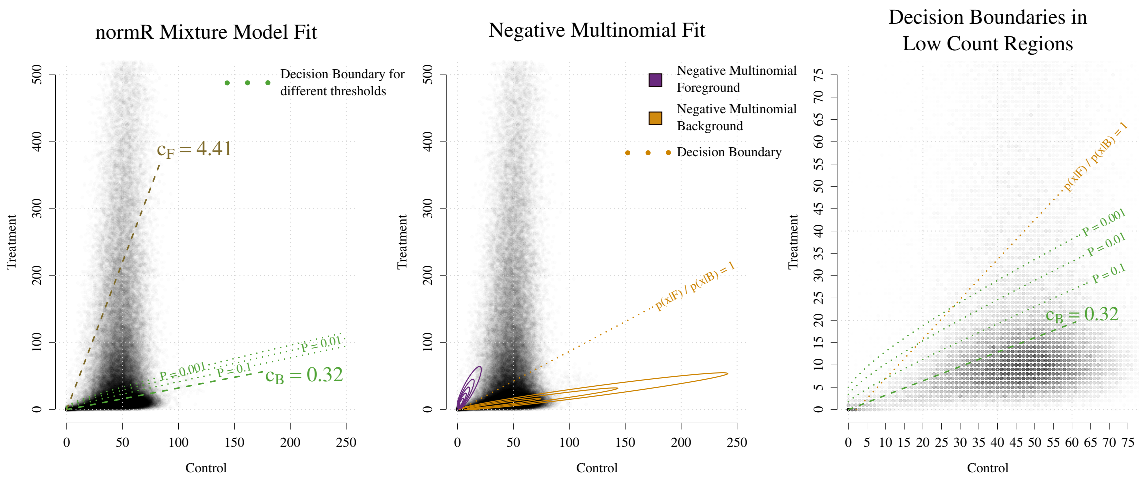

Fig. 3.5 – A Negative Multinomial Mixture Fit Does Not Model The Interrelation between Treatment and Control. The normR framework models the relation of treatment to control in background B and foreground F (left; Decision boundary based on the binomial test given). A 2-Mixture Model of Negative Multinomials models the density of read counts in two dimensions resulting in a separation of low and high count regions (middle; Decision boundary based on likelihood ratio given). Low count regions are not classified as background in the Negative Multinomial Mixture and the normR classification is more accurate (right). See also Supplementary Fig. A.1.

bution to model NGS counts. Despite the fact that this model is generally accepted as a natural formulation for the over-dispersed (*i.e.* heteroskedastic) NGS count data, I decided to use a mixture of Binomial distributions because of three reasons:

(i) The maximum likelihood estimation of a mixture of binomial distributions is computationally more malleable due to closed form updates for mixtures of the exponential family in the EM algorithm (see Section 2.2.2) than fitting a mixture of negative binomial distributions with numerical methods (see, for example, [69]).

(ii) The normR framework accounts for heteroskedasticity by encoding prior knowledge on the number of mixture components, *i.e.* background B and a finite set of signal components F. In Chapters 5 and 6, I explain how multiple predetermined signal components facilitate an accurate and meaningful model fit to ChIP-seq data. Here, the normR framework is able to deal with multimodality in the read count distribution, *i.e.* distinct *foregrounds*. The signal component could comprise an *a priori* unknown number of distinct signal "regimes" and, in the future, it might be desirable to encode this uncertainty in the mixture model, *e.g.* by modeling the signal component as a β-Binomial distribution. The Negative Binomial distribution can be expressed as a continuous mixture of Poisson distributions, *i.e.* the mean is not fixed, with the Poisson parameter λ being a random variable distributed as a gamma distribution. However, the Negative Binomial distribution is unimodal and it may be difficult to project the presence of distinct signal regimes that give rise to a multimodality in the read distribution. In principle, a mixture of

Negative Binomials or Negative Multinomials could be used (see below).

(iii) The normR framework uses a mixture of Binomial distributions to model the "interrelation" between the counts in treatment s and control r rather that explicitly modeling the density of points $N = r + s$. Consequently, the normR fit can be seen as a regression of the treatment versus the control conditional on the mixture component. As the multivariate generalization of the Negative Binomial distribution, the Negative Multinomial distribution, has recently been used to model read counts in ChIP-seq data [46]. Figure 3.5 compares the normR model to a Mixture Model of Negative Multinomials that models the read count density in two dimensions. Thereby the latter models the high density of points in regions with few treatment s counts (background-regions). This fit separates high and low count regions effectively but does not model the relation of treatment to control for background regions over a range of count values. In consequence, *putative* background regions with high and low control counts fall into two different components. Even with an increased number of Negative Multinomial components, the model does not find *one* foreground component that models the signal – most likely due to a high variance (see (ii); Supplementary Fig. A.1). The normR framework correctly estimates the relation of treatment to control and its statistical test results in an adequate decision boundary.

3.3 Outlook

In the following chapters I demonstrate the normR framework's suitability in three different scenarios published also in [2]:

(i) Chapter 4 introduces "enrichR" which facilitates the enrichment calling for high (H3K4me3) and low (H3K36me3) S/N data and shows better performance than previously published methods [74–80];

(ii) Chapter 5 describes "regimeR" which discovered two previously undescribed H3K27me3 and H3K9me3 heterochromatic regimes of broad and peak enrichment that are correlated to sequence features and binding of histone methyltransferase recruitment alike and are indicative for heterochromatin dynamics in the HepG2 human hepatocarcinoma cell line;

(iii) Chapter 6 explains "diffR" which calls differential H3K4me3 or H3K27me3-enrichment between HepG2 cells and primary human hepatocytes and performs well when compared to previous methodologies [90–92].

The normR framework is implemented in R [81] and C++ using bamsignals [1] for efficient read quantification. Its source code is freely available under GNU General Public License, Version 2 on Bioconductor [82] at `http://www.bioconductor.org/packages/normr` [4].

Chapter 4

enrichR– Enrichment Calling in ChIP-seq Data with the normR Framework

In this chapter the normR framework (Chapter 3) is used to call statistically significant enrichment in ChIP-seq data – a normR application referred to as "enrichR". When applied to high (*localized* H3K4me3) and low (*delocalized* H3K36me3) signal-to-noise ratio (S/N) ChIP-seq data enrichR calls genuine enrichment as valdidated by functional outputs such as gene expression, DNA methylation state and histone methyltransferase binding. A thorough comparison to enrichment calls of previously developed approaches [74–80] illustrates the superior sensitivity of enrichR accounted for by an adequate background estimation, especially in genomic regions with only minute ChIP-seq read enrichment over control. The enrichR normalized enrichment corresponds to the one estimated by other *in silico* approaches [85,86] and to the Histone Mark Density (HMD%) inferred from ICeChIP-seq experiments that use spiked-in semi-synthetic nucleosomes for normalization [93]. Based on the enrichR enrichment calls the chromatin segmentation by ChromHMM [36] is augmented by the identification of a previously undetected poised enhancer state as well as by the dissection of a large previously unresolved chromatin state.

4.1 Introduction

Chapter 1 described the ChIP-seq protocol which provides genome-wide localization data for DNA-associated proteins. By mapping sequenced fragments to a reference genome protein binding sites can be inferred by an accumulation of sequencing reads, *i.e.* **"enrichment"**. Due to the genome-wide scalability and cost-efficiency of ChIP-seq, hundreds of DNA-associated proteins have been assayed in different cell types, *e.g.* by the ENCODE [43] and Roadmap Epigenomics consortia [44]. This huge resource of ChIP-seq data sets paved the way for detailed genome-wide characterization of transcription factor binding sites [94], chromatin landscapes [36, 46] or *cis*-regulatory elements like enhancers [95,96]. Most ChIP-seq studies integrate protein binding with other functional outputs like gene regulation, *e.g.* in ES cell differentiation [97]. Furthermore, ChIP-seq signals of histone marks are predictive for promoter [34] and enhancer [98] activity. To enable for those studies an adequate ChIP-seq data normalization is crucial.

The discrimination of signal from background facilitates the identification of regions bound by a protein of interest, referred to as **"enrichment calling"**. However, this inference is complicated due to the "binding mode" of the protein of interest and apparent technical biases in the ChIP-seq experiment. The "binding mode" of a protein influences the S/N between signal- and background-regions in the ChIP-seq read densities (see also Section 3.2.2):

A high S/N is observed in ChIP-seq data of transcription factors that bind a DNA binding motif and also in ChIP-seq data of certain localized histone modifications such as H3K4me3 or H3K27ac (*localized enrichment*). The identification of signal-regions is usually easily achieved in this scenario due to a high S/N.

A low S/N is observed in ChIP-seq data of histone modifications with a delocalized read accumulation such as H3K27me3, H3K36me3 or H3K9me3 (*delocalized enrichment*). The determination of signal-regions is complicated in this scenario due to a low S/N.

In addition, *"technical biases"* introduced in the experiment lead to accumulation of reads in regions that are devoid of the antigen (see [72] for a review). The read densities obtained in the ChIP-seq experiment need to be compared to a corresponding read profile obtained by a control experiment to mitigate the effects of technical biases in ChIP-seq, *e.g.* copy number variations or chromatin structure [83,84]. Together, the S/N properties and the presence of technical biases complicate the enrichment calling in ChIP-seq data.

Earlier methodologies follow a two-step procedure to call enrichment: In the first step ChIP-seq read densities are normalized against a control experiment and, second, enriched regions are identified in normalized read counts. The initial normalization is achieved either by the ratio of sequencing depths [74, 76], by the ratio of ChIP and control read counts summed over *ad hoc*-chosen fixed width background-regions [77, 78, 80], or by the data-driven identification of background-regions [85, 86]. In a second step these approaches identify signal-regions that are characterized by a read enrichment and equate those to genomic loci bound by the antigen. The question remains as to which of those methods performs best given the mutual dependency of normalization and enrichment calling (see Section 3.1).

The **correctness** of a classification method is routinely assessed as a supervised learning problem with respect to a "gold-standard" that defines a true positive and a true negative set. However, the performance assessment of enrichment callers is aggravated because there exists, as yet, no universal gold-standard for ChIP-seq enrichment (see [99] for review). ChIP enrichment for a few dozen regions validated by low-throughput ChIP [100, 101] provides an initial performance assessment but is neither unbiased nor genome-wide scalable. Another approach represents the definition of a "*bona-fide* benchmark", *i.e.* a trustworthy validation set to score a classification method. This validation set can be obtained, for example, by combining conditionally independent classifiers to a meta-classifier [102] or by a consensus-vote strategy among classifiers [7]. Those *bona-fide* benchmarks can readily be used to derive a performance score for each method among a set of classification approaches.

Some ChIP-seq enrichment callers perform less well when the sequencing depth is reduced [103]. To demonstrate the **robustness** of an enrichment caller, a gold-standard can benchmark enrichment calls on *in silico* downsampled ChIP-seq and control libraries. Lower sequencing depth reduces the statistical power but, nevertheless, a robust enrichment caller ought to return consistent results over a range of sequencing depths.

An enrichment caller can be assessed by **auxiliary information**, *i.e.* its ability to recover known biological phenomena, as well as its correlation to signals that are measured by complementary or even more advanced experimental assays. For instance, nucleosomes trimethylated on H3K4 (H3K4me3) have been reported to be found at hypomethylated promoter regions [104–106] and CpG islands [7]. Moreover, the body of a transcribed gene is marked by the trimethylation of H3K36 (H3K36me3) [107] which associates with DNA-hypermethylation [108]. These reported biological insights aid to judge the correctness of an enrichment caller. Further, signals measured by CAGE or RNA-seq can be correlated to the normalized ChIP-seq read counts for a protein that associates with either promoter (H3K4me3) or gene (H3K36me3) activation, respectively. A correct normalization of histone modification ChIP-seq data ought to equate to the Histone Modification Density (HMD%) measured by the novel ICeChIP-seq technique [93]. Therein, a standardization is achieved by an advanced chromatin spike-in technique to infer the true normalization factor c_B experimentally. If a novel enrichment caller performs better than previous methods in these scenarios and under a *bona-fide* benchmark, it will augment previous studies that rely on sensitive enrichment identification like chromatin segmentation by chromHMM [36].

Here, the normR framework (see Chapter 3) is put to the test. The framework models NGS read count densities in the experiment and control as a multinomial sampling trial by means of a binomial mixture model. This data-driven approach follows the notion that normalization and calling of signal-regions are inseparable. To this extent, normR simultaneously performs the normalization against the Input (control) and the identification of regions enriched by the ChIP (treatment) – an application of normR referred to as **"enrichR"**. The robust normalization of ChIP-seq in enrichR is achieved by the estimation of a normalization factor based solely on regions that are putatively devoid of the protein. This normalization aids the sensitive identification of regions enriched by the ChIP, *i.e.* signal-regions / binding sites of the protein. In the following I will apply enrichR to high S/N (*localized* H3K4me3) and low S/N (*delocalized* H3K36me3) ChIP-seq data. A systematic comparison of enrichR enrichment calling to previously developed methods [74–78, 80] shows that enrichR outperforms its competitors' methods, especially for low S/N ChIP-seq data. The robust enrichR normalization improves on *in silico* normalization methods [85] and shows an astonishing agreement to the *in vitro* spike-in inferred normalization of ICeChIP-seq [93]. Furthermore, a substantially augmented resolution in chromatin segmentation by chromHMM [36] is achieved if enrichR enrichment calls are used as input. Taken together, these findings support the applicability of the normR approach in the normalization of ChIP-seq data.

4.2 Methods

Firstly, details on the processing and the quality of the sequencing data are provided. Secondly, the normR framework is adapted to the tasks of ChIP-seq enrichment calling which is referred to as "enrichR". Thirdly, a confidence-weighted beta-value for WGBS data is introduced to score the validity of enrichment calls. Fourthly, to compare enrichR enrichment calls to results of other approaches I introduce a binary classifier statistic that is based on a consensus vote among the set of approaches tested [74–78, 80]. Next, enrichR's normalization is compared to NCIS's normalization [85] and to results obtained by ICeChIP [93] which uses spiked-in modified nucleosomes to estimate a histone mark density (HMD%). Finally, based on chromHMM's [36] `LearnModel` routine, an enrichR-chromHMM hybrid is proposed and demonstrated to achieve an improved chromatin segmentation by constructing the input $(0, 1)$-matrix based on enrichR enrichment calls.

4.2.1 Data Sets

Primary Human Hepatocytes

ChIP-seq Data. Paired end reads from Input, H3K4me3, H3K27me3, H3K36me3 and H3K9me3 ChIP-seq for primary human hepatocytes were mapped with bwa (version 0.6.2, [109]) against human genome version hg19. Initial data quality was assessed for reads with mapping quality ≥ 20 with bamFingerprint tool in deepTools [110] version 2.3 (Supplementary Figure A.2). Fragment coverage tracks for browser display were generated with deepTools [110] in 25 bp windows (`-bs 25`), considering only reads with a mapping quality of at least 20 (`-MinMappingQuality 20`), normalized to the effective genome size (`-normalizeTo1x 2451960000`) and, for paired end data only, filtering for first reads in a properly mapped pair (`-samFlag 66`) with `bamCoverage -bam in.bam -o out.bw -of bigwig -bs 25 [-samFlag 66] -minMappingQuality 20`

-normalizeTo1x 2451960000.
For single end ChIP-seq data read counting, I shifted reads by 100 bp in 3' direction (shift = 100). For paired end ChIP-seq data read counting, I considered only reads with a mapping quality of at least 20 (mapq = 20). I regarded midpoints of properly mapped fragments (midpoint = TRUE) that were unduplicated (filteredFlag = 1024) and within 100 to 220 bp in length (tlenFilter = c(100, 220)) with normR's countConfigPairedEnd() function [4] (see Section 4.2.2 for necessary R code).

RNA-seq Data. Trizol extration was used for the preparation of Total RNA according to the manufacturer's guidelines and as described in [111]. An Agilent Bioanalyzer (Agilent, Santa Clara, USA) was used to check RNA integrity following the manufacturer's guidelines. Strand-specific sequencing libraries for mRNA and total-RNA were constructed for the human hepatocytes using the TruSeq stranded Total RNA kit (Illumina Inc, San Diego, USA) starting from 500 ng of the total RNA of the samples. Illumina HiSeq2000 was used to perform the sequencing (101-nucleotide paired-end reads for each library), resulting in the creation of about 100 million reads per library. The reads were aligned to the NCBI 37.1 (hg19) version of human genome using TopHat v2.0.11 [112] in the settings -library-type fr-firststrand and -b2-very-sensitive. Reads mapping to genes were counted using htseq-count from HTSeq-0.6.1p1 [113] in -f bam -s reverse -m union -a 20 setting. Annotation file for running htseq-count was downloaded from GENCODE release 19 (GRCh37.p13) [114].

CAGE Data. Primary human hepatocyte CAGE data was downloaded from http://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.primary_cell.hCAGE/Hepatocyte%252c%2520donor2.CNhs12349.11603-120I1.hg19.nobarcode.bam [115]. Reads with mapping quality of at least 20 were counted with bamsignals [1].

WGBS-seq Data. For primary human hepatocyte two types of whole-genome bisulfite sequencing NGS libraries were produced to achieve even read coverage. Firstly, 100ng of DNA was used in the TruSeq DNA methylation kit (Illumina, San Diego, USA) according to the manufacturer's protocol. The second type was performed as previously described [7]. Briefly, 2 g of DNA were sheared using a Bioruptor NGS device (Diagenode, Liege, Belgium) and cleaned-up using Ampure beads XP (Beckman Coulter, Brea, USA). Next, samples were subjected to end-repair, A-tailing and adaptor ligation steps using components of the TruSeq DNA PCR-Free Library Preparation Kit (Illumina). After bisulfite conversion involving the Zymo Gold kit (Zymo, Irvine, USA) the libraries were PCR amplified for 10-12 cycles. The amplified libraries were purified using Ampure beads XP and sequenced on three lanes of V3 paired-end flow cells (2x 100bp). Reads were mapped using bwa [109] and methylation levels were called with Bis-SNP37 [116].

GM12878 cells

ChIP-seq Data. GM12878 ChIP-seq alignment bam files for hg19 were downloaded from the UCSC ENCODE DCC repository (hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/

wgEncodeBroadHistone/) for CTCF, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3-K4me3, H3K9ac, H4K20me1 and Input (Whole Cell Extract; WCE). Data quality was assessed with deepTools [110] bamfingerPrint as described above (Supplementary Figure A.3). Based on this assessment, Input (WCE) bam files were merged.

Hek293 cells and mouse embryonic stem cells (mESC)

ICeChIP-seq Data. Downloaded ICeChIP-seq [93] paired end reads for mouse embryonic stem cells (Input, SRA-Accession: SRR1714013; H3K4me3, SRR1714008; H3K36me3, SRR1714011; H3K79me2,SRR1714012; H3K27me3, SRR1714010; H3K9me3, SRR1714009) were mapped with bowtie2 [117] against mm9:

```
bowtie2 -p 8 -x mm9 -1 Reads_1.fastq.gz -2 Reads_2.fastq.gz | \
 samtools view -bS - | samtools sort -@ 8 - out
samtools index out.bam
```

Furthermore, bigWigs containing the quantitative Histone Mark Density values (HMD%) were downloaded from Gene Expression Omnibus under accession GSE60378.

Transcription Start Site Definition

54,763 promoters (extend 750bp down- and upstream of TSS) of 54,849 GENCODE genes [114] obtained by using GenomicFeatures R package [118]:

```
require(GenomicFeatures)
gencode <- loadDb("data/gencode.v19.annotation.transcriptDb.sqlite")
genes <- genes(gencode)
proms <- unique(promoters(genes, upstream=750, downstream=750))
```

4.2.2 The normR Methods: enrichR

The normR framework (Chapter 3) was adapted to calling enrichment in ChIP-seq Data, referred to as **"enrichR"** (Figure 4.1). It uses two mixture components, *i.e.* background B and foreground F (enriched), to normalize and call enrichment over Input, *i.e.* control. There are now three free parameters, namely θ_B, θ_F and π_B:

θ_B) represents the expected fraction of reads in the ChIP over the sum of reads from ChIP and Input in a non-enriched region (*background-region*);

θ_F) represents the expected fraction of reads in the ChIP over the sum of reads from ChIP and Input in an enriched region (*signal-region*); and

π_B) is the proportion of background-regions over all regions, *i.e.* the proportion of signal-regions is $\pi_F = 1 - \pi_B$.

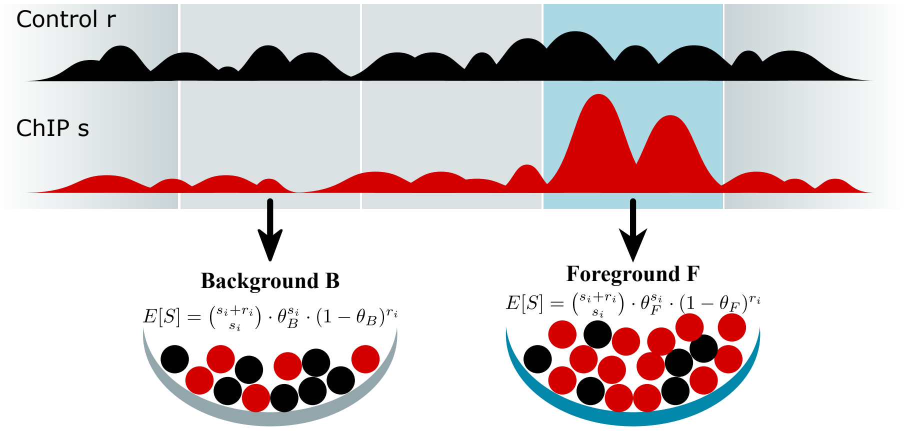
Control r
ChIP s
Background B
$E[S] = \binom{s_i+r_i}{s_i} \cdot \theta_B^{s_i} \cdot (1-\theta_B)^{r_i}$
Foreground F
$E[S] = \binom{s_i+r_i}{s_i} \cdot \theta_F^{s_i} \cdot (1-\theta_F)^{r_i}$

Fig. 4.1 – `enrichR`: The normR Method for Two Components. Reads in control r, *e.g.* Input, and ChIP s are modeled as a binomial mixture model with two components. Here, two components model the expected fraction of reads in the ChIP over the sum of reads from ChIP and control per region for background θ_B and the foreground θ_F, *i.e.* enriched.

Given the normR model in Equation (3.2) and the normR likelihood function defined by Equation (3.3) the following “enrichR” likelihood function can be derived:

$$\mathcal{L} = P(\pi_B, \theta_B, \theta_F | s_i, r_i) = \prod_i \binom{s_i + r_i}{s_i} \left(\pi_B \cdot \theta_B^{s_i} \cdot (1-\theta_B)^{r_i} + \pi_F \cdot \theta_F^{s_i} \cdot (1-\theta_F)^{r_i} \right),$$

where s_i (r_i) corresponds to the number of reads in the ChIP (Input) for non-overlapping, fixed size genomic bins $i = 1, \dots, n$.

The parameters θ_B, θ_F and π_B are then fitted by the EM algorithm [70] (see Section 2.2.3). At iteration t the posterior probability that a bin i is generated by the background B is

$$P(X_i = B | s_i, r_i) = \frac{\pi_B^{(t)} \cdot (\theta_B^{(t)})^{s_i} \cdot (1-\theta_B^{(t)})^{r_i}}{\pi_B^{(t)} \cdot (\theta_B^{(t)})^{s_i} \cdot (1-\theta_B^{(t)})^{r_i} + \pi_F^{(t)} \cdot (\theta_F^{(t)})^{s_i} \cdot (1-\theta_F^{(t)})^{r_i}}$$

at the values of parameters π_B^t, θ_B^t and θ_F^t. Next, the parameters are re-estimated as

$$\begin{aligned} \pi_B^{(t+1)} &= \frac{\sum_{i=1}^n P(X_i = B|s_i, r_i)}{n} \\ \theta_B^{(t+1)} &= \frac{\sum_{i=1}^n P(X_i = B|s_i, r_i) \cdot s_i}{\sum_{i=1}^n P(X_i = B|s_i, r_i) \cdot (s_i + r_i)} \\ \theta_F^{(t+1)} &= \frac{\sum_{i=1}^n (1 - P(X_i = B|s_i, r_i)) \cdot s_i}{\sum_{i=1}^n (1 - P(X_i = B|s_i, r_i)) \cdot (s_i + r_i)}. \end{aligned}$$

The algorithm continues until the likelihood converges, *i.e.* it does not change anymore, *e.g.* the change $\Delta\mathcal{L} \leq \epsilon = 0.001$. The EM algorithm converges to a local maximum. The chance to find a global maximum is increased by running the routine 10 times (per default) with θ_B and θ_F

randomly initialized close to $\theta^* = \frac{\sum s_i}{\sum s_i + r_i}$ (see Section 3.2.3).

To recover significantly enriched regions, the ChIP read count in each region is compared to the expected ChIP-seq read count under the fitted background component B with a binomial test (Figure 4.1). The distribution of the p-values from a binomial test is discrete and, thus, the correction for multiple testing is impeded (see Section 2.1.2). By filtering out low power tests, *i.e.* low count regions, with the T method [61], the p-value distribution becomes more uniform and p-values can readily be transformed to q-values [56]. Enriched regions are reported if they fall below a significance threshold α.

On the basis of the enrichR fit, a normalized and regularized enrichment is calculated. To account for noise that is generated by low count regions, counts are adjusted by adding pseudo counts for ChIP and Input. The pseudo counts are model-derived and taken to be the average read counts in the background component:

$$\alpha_r = \frac{\sum_{i=1}^n P(X_i = B|s_i, r_i) \cdot r_i}{\sum_{i=1}^n P(X_i = B|s_i, r_i)}$$
$$\alpha_s = \frac{\sum_{i=1}^n P(X_i = B|s_i, r_i) \cdot s_i}{\sum_{i=1}^n P(X_i = B|s_i, r_i)}.$$

The regularized ChIP-seq enrichment e^* is then calculated with

$$e_i^* = \log\left(\frac{s_i + \alpha_s}{r_i + \alpha_r} \cdot \frac{\alpha_r}{\alpha_s}\right),$$

where the second terms regularizes e_i^*, *i.e.* shifts to 0, for background regions. To account for the achieved ChIP enrichment, e^* is normalized with the $\log$ of the model-derived average enrichment factor

$$\langle f \rangle = \frac{\theta_F}{1 - \theta_F} \cdot \frac{1 - \theta_B}{\theta_B}$$

to obtain a regularized and normalized enrichment e:

$$e_i = \frac{e_i^*}{\log\langle f \rangle}.$$

These routines were implemented in the normR R package [4] as the `enrichR`()-function (see also R code snippet below).

For enrichment calling, only regions on regular autosomes (`chr1`-`chr22`; 2.9 Gigabases (Gb)) were considered:

```
require(GenomeInfoDb)
genome <- fetchExtendedChromInfoFromUCSC("hg19")
genome <- genome[which(!genome$circular &
                 genome$SequenceRole=="assembled-molecule"), 1:2]
genome <- genome[grep("X|Y|M", genome[, 1], invert=T), ]
require(GenomicRanges)
genome.gr <- GRanges(
 seqnames = genome[, 1],
 ranges = IRanges(start = 1, end = genome[, 2]),
 seqinfo = Seqinfo(
   seqnames = genome[,1],
   seqlengths = genome[,2],
   genome = "hg19"
 )
)
```

A binsize of 500bp (1000bp) was used for H3K4me3 (H3K36me3) because both θ and π were not robust for much smaller bin sizes (Supplementary Figure A.4). Read counts were modeled with `enrichR` and the fitted background component B was used for significance testing. Bins with q-value ≤ 0.05 (H3K4me3) and q-value ≤ 0.1 (H3K27me3/K36me3/K9me3) were called enriched and exported to bed tracks for display:

```
require(normr)
countConfig <- countConfigPairedEnd(
  binsize = 500, #1000
  mapqual = 20,
  midpoint = TRUE,
  filteredFlag = 1024,
  tlenFilter = c(100,220),
  shift = 0
)
fit <- enrichR(
  treatment = "ChIP.bam",
  control = "Input.bam",
  genome = genome,
  countConfig = countConfig,
  procs = 8
)
exportR(
  x = fit,
  filename = "enriched.bed",
  type = "bed",
  fdr = 0.05 #0.1
)
```

Finally, e was exported to bigWig tracks for browser display:

```
exportR(
  x = fit,
  filename = "enrichment.bigWig",
  type = "bigWig"
)
```

4.2.3 Confidence-Weighted Quantification of DNA-Methylation

To account for, both, the number of CpGs in a bin i and the read coverage at each CpG, I calculated confidence-weighted DNA-methylation β values. For each fixed width bin i and covered CpGs M β was calculated by

$$\beta_i = \frac{\sum_{j=0}^{M} \text{ReadCount}_j \cdot \text{FractionMethylated}_j}{\sum_{j=0}^{M} \text{ReadCount}_j}.$$

Only regions with at least 2 CpGs covered by reads were reported.

4.2.4 Comparison of Enrichment Callers

For comparison to enrichR, peaks in H3K4me3 and H3K36me3 ChIP-seq data in primary hepatocyte were called with six previously developed tools for enrichment (peak) calling [74–80]. To compare called peaks by above methods to enrichR called regions, overlap of peaks with 500bp (1,000bp) windows was calculated for H3K4me3 (H3K36me3). A comparison was achieved in two ways: (i) The overlap of enrichR results with third-party tools is analyzed for auxiliary information like DNA-methylation and expression, and (ii) Binary classification scores like precision, recall and F$_\beta$-score are calculated based on a validation set defined by a consensus vote strategy.

Enrichment Calling in Third-Party Tools

Peaks were called with MACS2 [74,75] (`v2.1.0.20150731`), DFilter [76] (`v1.6`), CisGenome [77], SPP [78], BCP [79] (`v1.1`) and MUSIC [80]. An FDR threshold of 0.99 was used where applicable to enable for subsequent filtering of results and the construction of precision-recall-curves. Firstly, duplicated fragments were removed and only reads with a mapping quality higher than 20 were extracted with samtools [119] (`v0.1.19-44428cd`) to allow for a fair comparison with enrichR:

```
samtools view -F 1024 -q 20 in.bam > out.bam
```

Secondly, peaks for H3K4me3 and H3K36me3 were called. MACS2 was run using the following commands:

```
macs2 callpeak -t ChIP.bam -c Control.bam -f BAMPE -g hs -q 0.99
```

For H3K36me3, results were merged with results from results with option “`-broad`”:

```
macs2 callpeak -t H3K36me3.bam -c Control.bam -f BAMPE -g hs --broad -q 0.99
```

DFilter was run using suggested configurations (`http://collaborations.gis.a-star.edu.sg/~cmb6/kumarv1/dfilter/tutorial.html`):

```
run_dfilter.sh -t=H3K4me3.bam -c=Control.bam -f=bam -pe -bs=100 -ks=100 \
  -lpval=0.001 -o=H3K4me3_result.bed
run_dfilter.sh -t=H3K36me3.bam -c=Control.bam -f=bam -pe -bs=100 -ks=20 \
  -lpval=0.001 -nonzero -o=H3K36me3_result.bed
```

CisGenome, SPP, BCP and MUSIC work on single end read alignments only. Here, only first reads in a proper mapped pair (`-f 66`) were considered for a fair comparison to peak callers working on paired end data:

```
samtools view -b -f 66 Input.bam > Input_SE.bam
samtools view -b -f 66 ChIP.bam > ChIP_SE.bam
```

For CisGenome, I generated *.aln files with piping bedtools bamtobed [120] and ran CisGenome’s SeqPeak routine using default parameters with a P-value cutoff of 0.99 on a generated filelist:

```
bedtools bamtobed -I Input_SE.bam > Input_SE.bed
cut -f 1,2,6 Input_SE.bed > Input_SE.aln
bedtools bamtobed -I ChIP_SE.bam > ChIP_SE.bed
cut -f 1,2,6 ChIP_SE.bed > ChIP_SE.aln
echo -n "Input_SE.aln\t0\nChIP_SE.aln\t1" > ChIP_filelist.txt \
  && ./seqpeak -i ChIP_filelist.txt -d . -o Result -bar 0 -lpcut 0.99
```

SPP was run in R using suggested configurations (`compbio.med.harvard.edu/Supplements/ChIP-seq/tutorial.html`) with a Z-score threshold of 0.5 and by removing chromosomes with no reads mapping to them:

```
chip.data <- read.bam.tags("ChIP_SE.bam")
input.data <- read.bam.tags("Control_SE.bam")
idx.notnull <- !sapply(chip.data[["tags"]], is.null)
chip.data <- lapply(chip.data, "[", idx.notnull)
input.data <- lapply(input.data, "[", idx.notnull)
bin.charac <- get.binding.characteristics(
  chip.data,srange = c(50,500),
  bin = 5,
  accept.all.tags = T
)
broad.clusters <- get.broad.enrichment.clusters(
  signal.data=chip.data[["tags"]],
  control.data=input.data[["tags"]],
  window.size=1e3,
  z.thr=0.5,
  tag.shift=round(bin.charac[["peak"]][["x"]]/2)
)
write.broadpeak.info(broad.clusters,file)
```

BCP was run using the following command:

```
./BCP_v1.1/BCP_HM -1 ChIP_SE.bed -2 Input_SE.bed -3 EnrichmentCalls.bed -p 0.9
```

For MUSIC, I downloaded mappability files for 50bp reads from (http://archive.gersteinlab.org/proj/MUSIC/multimap_profiles/hg19/hg19_50bp.tar.bz2) and ran the following command:

```
samtools view Input_SE.bam | ./MUSIC -preprocess SAM stdin Input/ && \
 ./MUSIC -sort_reads Input Input/sorted && \
 ./MUSIC -remove_duplicates Input/sorted 2 Input/dedup
samtools view ChIP_SE.bam | ./MUSIC -preprocess SAM stdin preprocessed && \
 ./MUSIC -sort_reads preprocessed sorted && \
 ./MUSIC -remove_duplicates sorted 2 dedup && \
 ./MUSIC -get_multiscale_punctate_ERs -chip dedup -control
```

Finally, to compare called peaks by above methods to enrichR enriched regions, overlap of reported peaks with 500bp (1,000bp) windows was calculated in R for H3K4me3 (H3K36me3) if a peak at FDR 0.05 (0.10) overlapped a window by at least 250bp:

```
binsize <- 500; fdr <- 0.05 #0.1
gr <- tileGenome(genome.gr, width = binsize)
ov <- matrix(0, nrow = length(gr), ncol = 7)
colnames(ov) <- c("enrichR", "MACS2", "DFilter", "CisGenome", "SPP",
                  "BCP", "MUSIC")
for (method in colnames(ov)) {
  peaks.sig <- peaks[[meth]][which(peaks[[meth]][["lqval"]] >= -log10(fdr))]
  ov[,method][countOverlaps(gr, peaks.sig, minoverlap = 250)> 0 )] <- 1
}
```

A *Bona-Fide* Benchmark Based on a Consensus-Vote among Peak Callers

Accuracy of Classification. Firstly, a "tool-specific *bona-fide* benchmark", *i.e.* a trustworthy validation set, for the evaluation of correctness of a tool was defined as follows: A bin is enriched under the tool-specific *bona-fide* benchmark if at least four out of six other methods (including enrichR) called this bin enriched. In R, I ran the following code:

```
gs <- lapply(colnames(ov), function(method) {
  which(apply(ov[,which(colnames(ov) != method)], 1, sum) >= 4)
})
names(gs) <- colnames(ov)
```

Secondly, I computed precision, recall and F_2-score under the "tool-specific *bona-fide* benchmark" in R:

```
getPrecRecall <- function(ov, gs) {
  mp <- which(ov == 1)
  tp <- sum(mp %in% gs)
  fn <- sum(!(gs %in% mp))
  fp <- sum(!(mp %in% gs))
  tn <- dim(ov)[1] - tp - fn - fp
  specificity <- tn / (fp + tn)
  precision <- tp / length(mp)
  recall <- tp / (tp + fn)
  f2 <- fscore(precision, recall, 2)
  return(c(
    "precision"=precision,
    "recall"=recall,
    "f2"=f2,
    "specificity"=specificity
  ))
}
stats <- mapply(getPrecRecall, as.list(ov), gs)
```

Thirdly, precision-recall-curves were computed for each tool under its own "tool-specific *bona-fide* benchmark" in R:

```
peaks <- lapply(peaks, function(p) {
  peaks[order(peaks[["lqval"]], decreasing=TRUE)]
})
nmax <- max(colSums(ov))
```

```
stats.sub <- lapply(seq(100, nmax, 100), function(n) {
  ov.sub <- lapply(peaks, function(p) {
    if (length(p) < n) {
      NULL
    } else {
      idx <- countOverlaps(gr, p[1:min(n, length(p))], minoverlap=250) > 0
      ov <- rep(0, length(gr))
      ov[idx] <- 1
      return(ov)
    }
  })
  return(getPrecRecall(ov.sub, gs))
})
```

Lastly, the obtained precision and recall values were used to plot a precision recall curve. Because some tools did not report the full spectrum of recall values I calculated a "PartAUC", *i.e.* the area under the curve ranging from the minimum to the maximum recall value. Thus, the "PartAUC" represents a lower bound of the full AUC.

Validity of Tool-Specific Classification. I catalogued "tool-specific regions" not represented by the "unified *bona-fide* benchmark", *i.e.* the union of seven "tool-specific *bona-fide* benchmark sets":

```
gsUnified <- unique(unlist(gs))
toolSpecCalls <- lapply(colnames(ov), function(method) {
  which(!(mat[,method] %in% gsUnified))
})
```

Robustness to Varying Sequencing Depth. For the saturation analysis I downsampled bam files with samtools to 5%, 10%, 20%, 30%, 50% and 75%:

```
for sub in .05 .1 .2 .3 .5 .75; do
 for f in *.bam; do
  of=${f/.bam/.Sub$sub}
  samtools view -u -s 1$sub $f | samtools sort - $of && samtools index $of.bam
 done
done
```

Next, peak calling and classification of enriched bins was performed as described above on the reduced libraries. The recovered fraction of "unified *bona-fide* benchmark" by each method was calculated in R:

```
stats.ds <- mapply(getPrecRecall, as.list(ov), rep(list(gsUnified), 7))
```

4.2.5 Correlating `enrichR`-estimated Enrichment to NCIS and HMD%

NCIS. The background normalization factor was calculated with NCIS [85] on the created single end bed files in R:

```
ncis <- NCIS(
  chip.data = "ChIP_SE.bed",
  input.data = "Input_SE.bed",
  data.type = "BED",
  chr.vec = seqnames(gr),
  chr.len.vec = seqlengths(gr)
)
```

ICeChIP. Similar to reported in [93] only midpoints of properly mapped fragments (`midpoint = TRUE`) were quantified. Furthermore, it was filtered for unduplicated fragments (`filteredFlag = 1024`) within 100 to 220 bp in length (`tlenFilter = c(100, 220)`). 500 (1,000) bp windows were used for H3K4me3 (H3K36me3/K79me2/K27me3/K9me3) using again normR's countConfigPairedEnd() function. Firstly, enrichR was run with this configuration to estimate the normalized enrichment genome-wide:

```
require(normr)
counConfig <- countConfigPairedEnd(
  binsize = 500,
  mapq = 20,
  midpoint = TRUE,
  filteredFlag = 1024,
  shift = 0,
  tlenFilter = c(0,220)
)
fit <- enrichR("ChIP.bam", "Input.bam", genome, procs = 8)
```

Secondly, to compare the estimated enrichR enrichment, I downloaded the ICeChIP histone mark density (HMD%) information for H3K4me3, H3K36me3, H3K79me2, H3K27me3 and H3K9me3 from Gene Expression Omnibus [121] under accession `GSE60378`. Because the original scaling factors $\langle \mathrm{IP}_{\mathrm{Ladder}} \rangle$ were not reported in [93] I inferred the average normalization factor based on

$$\langle \mathrm{IP}_{\mathrm{Ladder}} \rangle = \frac{1}{n} \sum_i \frac{100}{\mathrm{HMD\%}_i} \cdot \frac{\text{ChIP-coverage}_i}{\text{Input-coverage}_i}.$$

In this regard, the per bp fragment coverage was calculated from bam files and HMD% from bigWig files in R with the help of bamsignals [1] and rtracklayer [122], respectively:

```
gr <- reduce(fit.rep1@ranges)
require(bamsignals)
coverage <- lapply(c("ChIP.bam", "Input.bam"), function(b) {
  unlist(bamCoverage(
   bampath = b,
   gr = gr,
   mapqual = 20,
   shift = 0,
   paired.end = "extend",
   tlenFilter = c(0,220),
   filteredFlag = 1024
 ))
})
require(rtracklayer)
hmd <- unlist(import.bw("HMD.bigWig", which=gr, as="NumericList"))
ratio <- 100/hmd * coverage[[1]]/coverage[[2]]
ipladder <- mean(na.omit(ratio[!is.infinite(bs)]), na.rm=T)
```

I computed Pearson's r of enrichR standardized enrichment e and HMD%. Finally, I fitted a linear model for $f(x = e) = \text{HMD\%} = \alpha + \beta * x$ and residuals were studentized.

4.2.6 Chromatin Segmentation Based on enrichR Enrichment Calls

The chromHMM method uses a Hidden Markov Model to segment the genome in distinct epigenetic states based on enrichment calls in a set of ChIP-seq experiments. The input for chromHMM's Hidden Markov Model is a $(0, 1)$-matrix where 0 indicates a background and 1 an enriched bin in non-overlapping fixed size windows along the genome, respectively. The chromHMM developers provide a program called BinarizeBam to identify enriched windows based on a Poisson background model (see Section 2.1.1).

Here, I would like to test if enrichR can increase the quality of the input $(0, 1)$-matrix by a sensitive enrichment identification. Firstly, as a comparison set, I ran chromHMM BinarizeBam with default options (200bp bins, reads shifted by 100bp) in GM12878 for two replicates of CTCF, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac and H4K20me1:

```
java -mx80G -jar ChromHMM/ChromHMM.jar BinarizeBam -b 200 -n 100 \
 hg19_chroms.txt GM12878 cellmarkfiletable GM12878chromHMMInput
```

Secondly, enrichR() was used to call enrichment ChIP-seq over Input for pooled replicates in 200bp windows with reads shifted by 100bp:

```
require(normr)
getSumOfCounts <- function(bampaths, gr = genome) {
  require(bamsignals)
  l <- lapply(bampaths, function(bam) {
    unlist(bamProfile(
      bampath = bam,
      gr = genome,
      shift = 100,
      binsize = 200
    ))
  })
  return(as.integer(colSums(do.call(rbind, l))))
}
r <- getSumOfCounts("Input.bam")
enr <- mclapply(names(bamfiles), function(b) {
  s <- getSumOfCounts(bamfiles[[b]])
  fit <- enrichR(s, r, gr)
  return(fit)
}, mc.cores=8)
```

Next, enriched regions were classified under FDR=10% and exported the binary matrices for each chromosome to feed this information into chromHMM's Hidden Markov Model:

```
invisible(mclapply(seqlevels(genome), function(chr) {
  idx <- which(seqnames(gr) == chr)
  mat <- sapply(names(bamfiles), function(n) {
    x = getClasses(enr[[n]], fdr=0.1)[idx]
    x[is.na(x)] = 0 #background is all NA classes
    x
  })
  #header of binary file
  cat(paste0("GM12878\t", chr, "\n"), file=chr, append=F)
  cat(paste(names(bamfiles), collapse="\t"), file=chr, append=T)
  cat("\n", file=chr, append=T)
  #(0,1)-matrix
  write.table(x = mat, file = chr, col.names=F, row.names=F, eol="\n",
              sep="\t", quote=F, append=T)
}))
```

Finally, I applied the chromHMM method to both binarizations:

```
java -mx80G --jar ChromHMM/ChromHMM.jar LearnModel -stateordering emission \
 -holdcolumnorder -printposterior -printstatebyline -b 200 -p 8 \
 Input_chromHMM/ Output_chromHMM/ 15 hg19
java -mx80G --jar ChromHMM/ChromHMM.jar LearnModel -stateordering emission \
 -holdcolumnorder -printposterior -printstatebyline -b 200 -p 8 \
 Input_enrichR/ Output_enrichR/ 15 hg19
```

Later, hidden states were labeled based on a hierarchical clustering of the emission probabilities which gave rise to Fig. 4.11. Fold enrichments for genomic features were taken from chromHMM `LearnModel` output and $\log_2$ transformed.

4.3 Results – Enrichment Calling in High and Low S/N

To illustrate the enrichment calling based on a robust background estimation, I applied enrichR to ChIP-seq experiments for *localized* and *delocalized* histone modifications in primary human hepatocytes (PHH). In a first analysis, auxiliary information is used to verify the valdidity of calls: The *localized* trimethylation of H3K4 (H3K4me3) correlates with promoter activity and DNA-hypomethylation [104–106]. The *delocalized* H3K36me3 is associated to transcriptional elongation in the body of transcribed genes [107] as well as DNA-hypermethylation [108]. H3K4me3 exhibited a substantially higher S/N ratio than H3K36me3 (Supplementary Fig. A.2). In a second instance, I show that the enrichR-based enrichment calls compare favorably to results obtained by six popular peak calling methods, namely MACS2 [74, 75], DFilter [76], CisGenome [77], SPP [78], BCP [79] and MUSIC [80]. Furthermore, the enrichR normalization recapitulates the normalization factors estimated by NCIS [85] based on regression and those determined via an *in vitro* chromatin spike-in tradegy in ICeCHIP experiments [93]. Finally, enrichR is applied to improve chromatin segmentation resolution through a enrichR-chromHMM hybrid approach.

As a first assessment, I studied the coverage and enrichment calls for H3K4me3 and H3K36me3 ChIP-seq in the vicinity of the Glucose-6-Phosphate Isomerase gene (GPI, Fig. 4.2A) — a housekeeping gene that is highly expressed in all cell types [123]. GPI was also expressed in PHH as measured by RNA-seq and showed a characteristic chromatin signature of transcription, *i.e.* H3K4me3 and H3K36me3 in the promoter and the gene body, respectively. All tested methods identified these characteristic enrichments at the GPI locus. Moreover, the promoter of the WTIP gene was detected as H3K4me3-enriched by all methods. Together with the measured shallow coverage of RNA-seq reads along its gene body this indicated that WTIP is expressed suggesting a low but genuine H3K36me3 enrichment in its gene body. Interestingly, this minute H3K36me3 enrichment was exclusively recovered by enrichR.

Genome-wide enrichR called H3K4me3-enrichment in 142,451 500 base pairs (bp) regions in PHH, corresponding to 45,522 consecutive regions representing ∼3% of the mappable genome

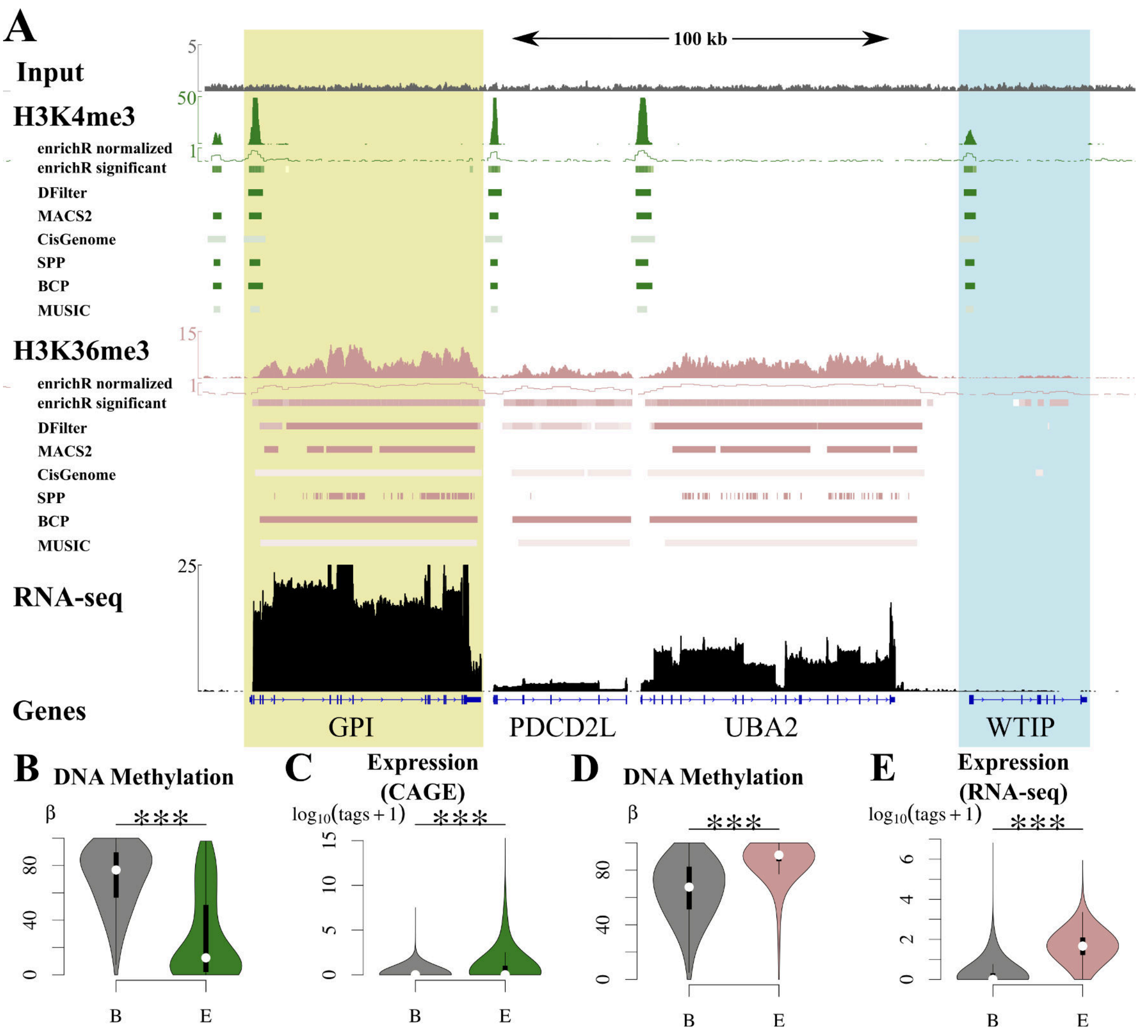

Fig. 4.2 – Enrichment Calling with enrichR on H3K4me3 and H3K36me3 ChIP-seq Data in Primary Human Hepatocytes. (A) Input (grey), H3K4me3 (green, high S/N), H3K36me3 (rose, lower S/N) and RNA-seq (black) barplots indicate coverage proximal to the human Glucose-6-Phosphate Isomerase (GPI, yellow overlay) locus on chromosome 19. Enrichment calls are indicated as colored boxes below respective tracks for enrichR, DFilter, MACS2, CisGenome's SeqPeak, SPP and BCP. The WTIP gene (blue overlay) had detectable H3K4me3 enrichment at its promoter and minute H3K36me3 is recovered solely by enrichR. (B-C) enrichR H3K4me3-enriched regions were DNA-hypomethylated (B) and expressed as measured by CAGE (C). (D-E) enrichR H3K36me3-enriched regions were DNA-hypermethylated (D) and expressed as measured by RNA-seq (E). Wilcoxon signed-rank Test: "***" ($P \leq 0.001$).

(71.2 Megabases (Mb)). The identified regions were characterized by low levels of DNA methylation (Fig. 4.2B), in line with the theory that H3K4me3 is repressing DNA methylation [104–106]. Furthermore, a higher density of CAGE-tags was observed in H3K4me3-enriched regions when compared to background regions (Fig. 4.2C) indicating that they serve as active transcriptional start sites (TSSs). In fact, enrichR H3K4me3-enriched regions showed a statistically significant overlap with annotated TSSs (odds-ratio=18.29, Fisher's signed exact test, P≤0.001, Table B.1). Together these observations support enrichR's identification of genuine H3K4me3-enriched regions.

For H3K36me3 enrichR identified 559,560 1 kilobase pair (kb) regions as enriched in PHH, corresponding to 85,293 consecutive regions representing 20% of the genome (599.6Mb). H3K36me3-enriched regions showed DNA hypermethylation (Fig. 4.2D), in line with the theory that H3-K36me3 recruits DNMT3B leading to *de novo* DNA methylation [108]. H3K36me3-enriched bins showed a significantly higher RNA-seq read coverage than background regions (Wilcoxon-signed-rank test P≤0.001, Fig. 4.2E) and a statistically significant overlap with annotated transcripts (odds-ratio = 17.06, Fisher's signed exact test, P≤0.001, Table B.2), in line with the reported association of H3K36me3 to transcriptional elongation [107]. These results support that enrichR also identifies genuine H3K36me3-enriched regions.

4.3.1 Systematic Comparison of Available Enrichment Callers

The enrichR H3K4me3 and H3K36me3 enrichment calls were compared genome-wide to MACS2, DFilter, CisGenome's SeqPeak, SPP, BCP and MUSIC results on two systematic levels (Methods Section 4.2.4):

(a) The **overlap of results** reported by all competitor tools with enrichR was quantified. Next, regions that are exclusive to a method were studied for auxiliary information like DNA methylation and expression.

(b) Because there was no genome-wide ChIP-seq benchmark set on-hand, I defined a **"tool-specific *bona-fide* benchmark"** for each method based on a consensus vote among the six remaining tools. This *bona-fide* benchmark set was used to evaluate every method for its enrichment classification accuracy and its robustness to an *in silico* reduced sequencing depth.

(a) Overlap of Results

All methods performed similarly for the *localized* H3K4me3 at a False Discovery Rate (FDR) of 5% (Fig. 4.3A), although in terms of covered bp DFilter (39.8Mb) and CisGenome (38.7Mb) called almost two-fold less enrichment than the other tools (mean=65.3Mb; Table 4.1). 13,364 regions that are only called by enrichR ("enrichR-exclusive") were distal to peaks called by competitors (median=7,137kb, Figure 4.3B). All H3K4me3-enriched regions shared among the tools ("tool-inclusive") were characterized by a significant DNA-hypomethylation when compared to background regions (Wilcoxon signed-rank test P≤0.001; Figure 4.3C). In fact, "tool-exclusive" regions are on average less methylated than the genomic average – with the exception of SPP and BCP. Up to DFilter and CisGenome, all "tool-exclusive" H3K4me3-enriched regions were

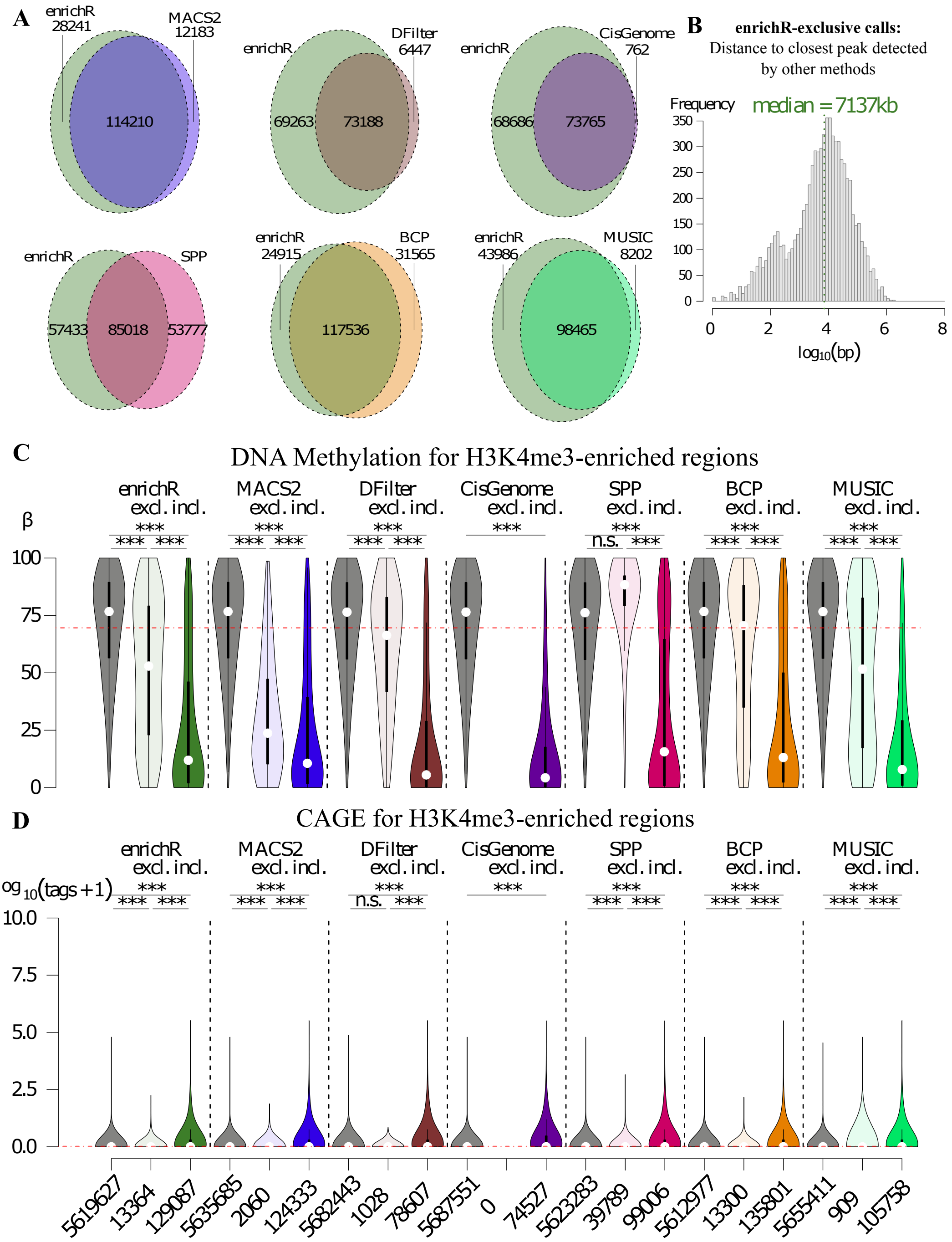

Fig. 4.3 – enrichR H3K4me3 Enrichment Calls Agree with Peaks Reported by Competitor Methods. (A) enrichR enrichment calls substantially overlap with peaks called by benchmark methods based on 500 genomic intervals. (B) Regions called exclusively by enrichR are distant to peaks called by other methods (median=7.137kb). (C-D) H3K4me3-enriched regions called exclusively by one (*"excl."*) or multiple (*"incl."*) methods are significantly lower DNA-methylated (C), except for SPP, and higher expressed (D), except for DFilter, as compared to background regions. There were no CisGenome-exclusive peaks. Red dashed line represents average genome-wide DNA-methylation or Expression. enrichR has the greatest amount of tool-exclusive regions (13,364) that were in concordance with auxiliary information. Wilcoxon signed-rank Test: "***" (P≤0.001) and "n.s." (P>0.05).

expressed significantly higher than background regions (Figure 4.3D). In summary, enrichR reported most exclusive regions that were also supported by DNA-hypomethylation and expression.

For the *delocalized* H3K36me3, enrichR identified up to >16-fold more enriched regions than its competitor methods (Table 4.1). When compared to enrichR (559.6Mb) far fewer regions were reported by MACS2 (407.7Mb), BCP (396.5Mb), MUSIC (402.3Mb) and especially DFilter (87.8Mb), SPP (25.1Mb) and CisGenome (36.4Mb). Almost all of these H3K36me3-enriched regions (MACS2: 399.1Mb; 97.9%, DFilter: 87.8Mb; 100%; CisGenome: 36.4Mb; 100%; SPP: 24.2Mb; 96.7%; BCP: 386.8Mb; 97.6%; MUSIC: 382.6Mb; 95.1%) were recovered by enrichR leading to few exclusive regions for competitor methods (Fig. 4.4A). Importantly, H3K36me3-enriched regions called exclusively by enrichR (93.6Mb; 16.7%) were characterized by a median distance of 2kb to peaks recovered by other methods (Fig. 4.4B). These regions also showed significantly higher DNA-methylation levels and transcriptional activity than background regions (Wilcoxon-signed-rank test $P \leq 0.001$, Fig. 4.4C-D). Taken together, many genuine H3K36me3-positive regions were only detected by enrichR as supported by information on DNA-methylation and expression – similar to results obtained for the *localized* H3K4me3.

(b) Evaluation by a *Bona-Fide* Benchmark

A systematic comparison of the tools' classification accuracy was performed in the following way. For each method a tool-specific validation set ("*bona-fide* benchmark") was computed as a consensus vote that comprised all regions *called by four out of six competitor methods* (Methods Section 4.2.4) In addition to precision ("positive predictive value") and recall ("true positive rate"), I use the F_2-score as a sensitivity-weighted score of the accuracy of a classification method to penalize especially false negative calls. The F_β-score is defined by:

$$F_\beta\text{-score} = (1+\beta^2)\frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot precision) + \text{recall}}. \tag{4.1}$$

The union of all tool-specific *bona-fide* benchmark sets was used to study the robustness of a method when the sequencing depth is reduced *in silico.*

Accuracy. For the *localized* H3K4me3, most methods performed well with a mean F_2-score of 0.79. At FDR=5%, CisGenome achieved highest precision (1.00), MACS2 had the highest recall (0.99) and the highest F_2-score (0.91) (Table 4.1A). enrichR achieved a high recall (0.97) and high F_2-score (0.87) with only slightly fewer regions called than BCP (74.5Mb; F_2-score=0.86). Lowest accuracy was observed for SPP (F_2-score=0.65) which also reported many invalid exclusive calls as described in the previous section. All other methods, except for SPP, performed equally well for H3K4me3 enrichment calling at different recall levels (Fig. 4.5A). The greatest area under the precision recall curve was observed for enrichR (AUC=0.97).

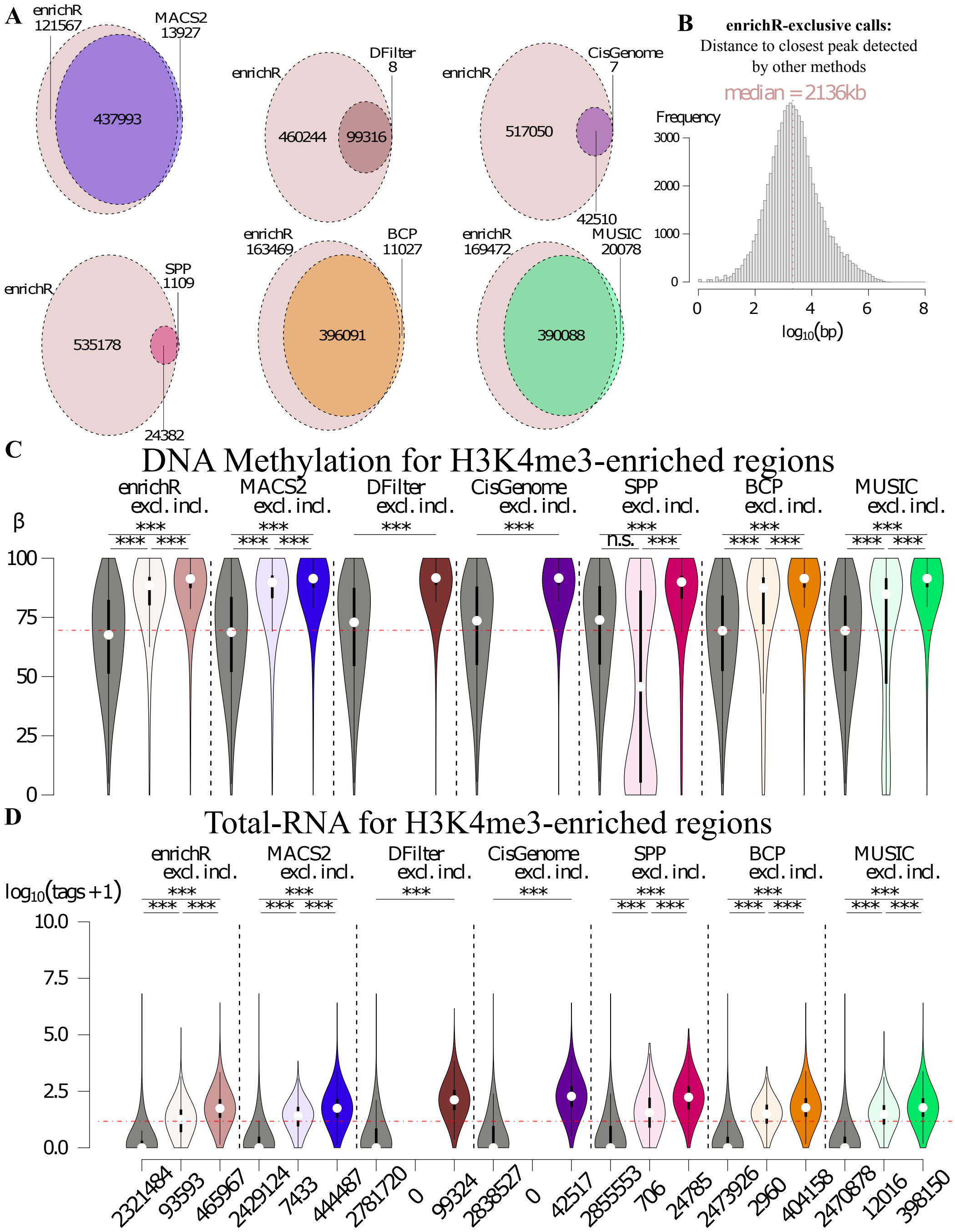

Fig. 4.4 – enrichR H3K36me3 Enrichment Calling Outperforms Competitor Methods. (A) enrichR enrichment calls substantially overlap with peaks called by benchmark methods based on 1,000bp genomic intervals. (B) Regions called exclusively by enrichR are distant to peaks called by other methods (median=2.136kb). (C-D) H3K36me3-enriched regions called exclusively by one (*"excl."*) or multiple (*"incl."*) methods are significantly more DNA-methylated (C) and higher expressed (D) as compared to background regions. There were no DFilter- and CisGenome-exclusive peaks. enrichR calls most exclusive regions (92,508) which are in concordance with auxiliary information, *i.e.* DNA-methylation and expression. Wilcoxon signed-rank Test: "***" (P≤0.001) and "n.s." (P>0.05).

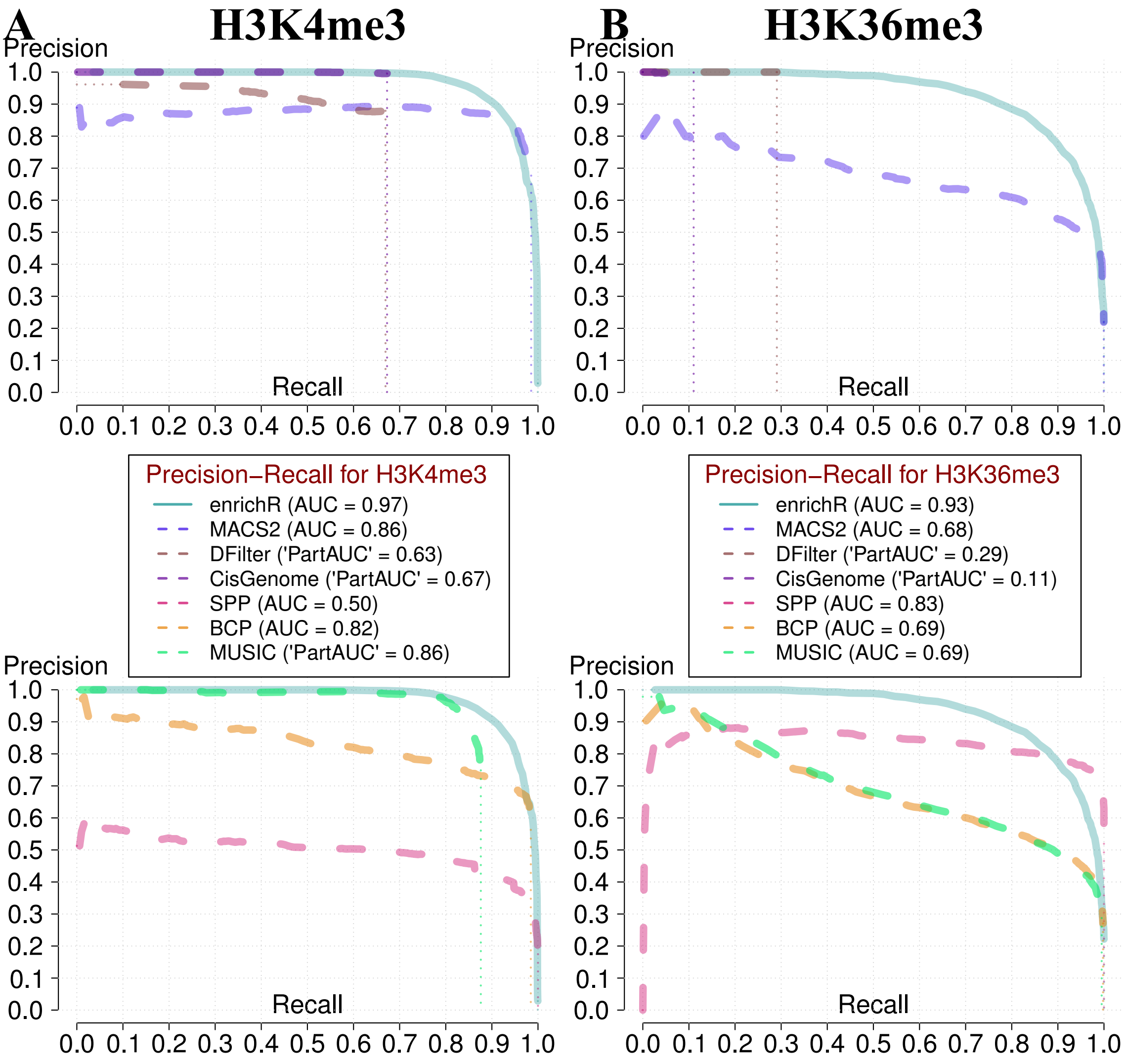

Fig. 4.5 – Precision-Recall-Curves Based on a *Bona-Fide* Benchmark Support the Superior Performance of enrichR. (A) All methods expect SPP have a precision≥0.82 at recall≤0.7 for H3K4me3. enrichR has greatest "PartAUC" with most regions reported. (B) For H3K36me3, enrichR reports the most bins enriched and has the greatest "PartAUC". Legends give number of H3K4me3- and H3K36me3-enriched regions at a FDR of 10%. Precision-Recall Curves and "PartAUC" were computed with respect to a tool-specific *bona-fide* benchmark (see Section 4.2.4).

For the *delocalized* H3K36me3, the performance of methods decreased ($\mu_{F_2\text{-score}}$=0.42) – probably accounted for by a diminished S/N (Table 4.1B). At FDR=10%, DFilter, CisGenome and SPP were most precise (≥0.87) with only a few regions called (≤87.8Mb), while enrichR, MACS2, BCP and MUSIC were most sensitive (≥0.99) with ≥4.5-fold more enriched regions reported. enrichR which called almost all regions of its six competitors combined (Fig. 4.4A) had an F_2-score=0.53 – its menial precision (0.19) is compensated by the best recall (1.00). Turning to precision at different recall levels, only enrichR could perpetuate a high precision throughout varying sensitivity levels (AUC=0.96) with SPP coming close (PartAUC=0.83; Fig. 4.5B). In fact, enrichR had the highest precision ($\geq$ 0.80) at high recall levels ($\geq$ 0.50) indicating that the *bona-fide* benchmark at FDR=10% misses most regions already identified by enrichR at FDR=10%. Taken together, the performances of tools in H3K36me3 enrichment calling differed extremely – a result supported by the analysis of the overlaps of enrichment calls in the previous section.

Given these observations the question arose whether the validity of the enrichment calls which were not represented by the *bona-fide* benchmark, *i.e.* "tool-specific" calls. To this extend, I defined a "unified *bona-fide* benchmark" of H3K36me3-enrichment (354,527 1kb regions) which represents the union of seven tool-specific *bona-fide* benchmark that were used in the previous paragraphs. For every method I catalogued detected regions that are not represented by this unified *bona-fide* benchmark for further investigation of their validity. The unified *bona-fide* benchmark exhibited a significantly higher fold change over Input than the enrichR-, MACS2-, SPP-, BCP- and MUSIC-specific regions (Wilcoxon-signed-rank test; P$\leq$0.001; Figure 4.6A). There were only 2 DFilter- and 3 CisGenome-specific regions. enrichR had the most tool-specific regions (205,064) which is equivalent the size of $\sim$57% of the benchmark and showed significantly higher fold changes as well as read coverages than background regions (Figure 4.6B). The

A **H3K4me3**

	Enrichment called		Scoring based on *bona-fide* benchmark					
	in Mb	in 500bp bins	Precision	Recall	$F_{0.5}$-score	F_1-score	F_2-score	Specificity
enrichR	71.23	142451	0.6010	**0.9730**	0.6500	0.7430	**0.8660**	**0.9900**
MACS2	63.71	126393	0.6830	**0.9880**	0.7280	**0.8080**	**0.9070**	**0.9930**
DFilter	39.84	79635	**0.8830**	0.6640	**0.8290**	**0.7580**	0.6990	**0.9980**
CisGenome	38.67	74527	**0.9950**	0.6830	**0.9110**	**0.8100**	0.7290	**1.0000**
SPP	69.40	138795	0.5040	0.6960	0.5330	0.5850	0.6470	**0.9880**
BCP	74.54	149101	0.5770	**0.9870**	0.6290	**0.7280**	**0.8640**	**0.9890**
MUSIC	53.45	106667	**0.7750**	**0.8740**	**0.7930**	**0.8210**	**0.8520**	**0.9960**

A **H3K36me3**

	Enrichment called		Scoring based on *bona-fide* benchmark					
	in Mb	in 1kb bins	Precision	Recall	$F_{0.5}$-score	F_1-score	F_2-score	Specificity
enrichR	559.56	559560	0.1859	**0.9997**	0.2220	0.3135	0.5330	**0.8360**
MACS2	407.65	451920	0.2301	**0.9996**	0.2720	0.3741	0.5990	**0.8747**
DFilter	87.78	99324	**0.9997**	0.2801	0.6604	0.4376	0.3272	**1.0000**
CisGenome	36.40	42517	**0.9994**	0.1199	0.4050	0.2140	0.1455	**1.0000**
SPP	25.09	25491	**0.8702**	0.0628	0.2436	0.1171	0.07710	**0.9987**
BCP	396.46	407118	0.2554	**0.9983**	0.3001	0.4068	0.6312	**0.8908**
MUSIC	402.25	410166	0.2535	**0.9910**	0.2979	0.4038	0.6265	**0.8897**

Table 4.1 – Enrichment Calling Statistics and Correctness for All Tools with Respect to the *Bona-Fide* Benchmark Set. (A) All tools perform well for enrichment calling in H3K4me3 ChIP-seq data at FDR=0.05, yet DFilter and CisGenome report two-fold less enrichment than other methods. CisGenome is most precise, MACS2 performs best for Recall and F_2-score. enrichR achieves a high recall and F_2-score. (B) Among all tools, enrichR reports most H3K36me3-enriched regions and achieves a high recall at FDR=0.10. All scoring values $\geq$0.75 are bold faced. Mb=Megabase; kb=kilobase.

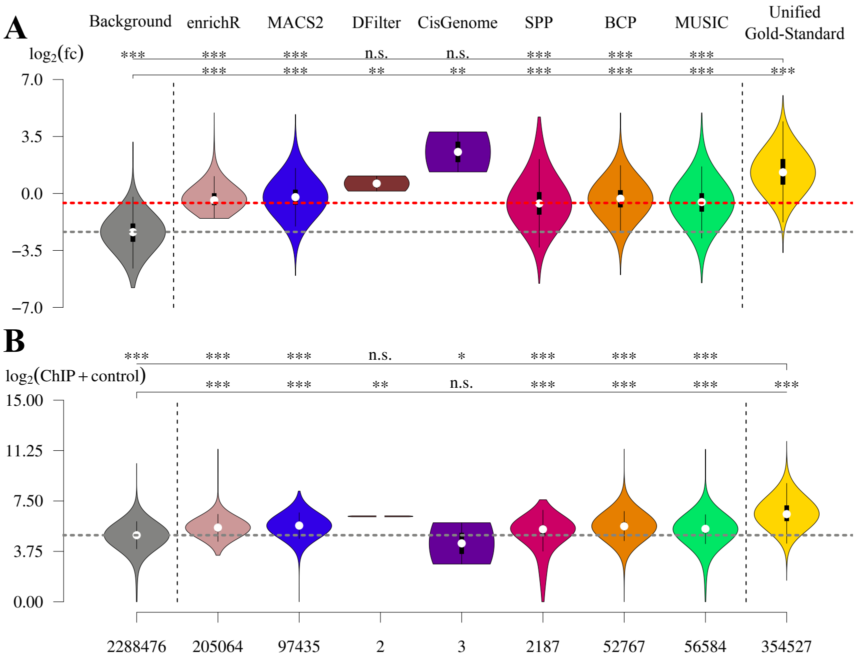

Fig. 4.6 – High Fold-Change and Read Coverage in Tool-Specific Regions Support Validity of enrichR Calls. (A) $\log_2$ fold-change (fc) of ChIP over Input and (B) $\log_2(\text{ChIP}+\text{Input})$ coverage compiled for "tool-specific" regions. Most tool-specific regions have a fc and read coverage significantly greater than background regions. 205,064 enrichR-specific regions have higher fc and read coverage than the genomic average and correspond to ∼57% the quantity of the benchmark. Red dashed line represents mean $\log_2(\text{fc})$. Grey dashed line represents mean $\log_2(\text{fc})$ (A) and mean $\log_2(\text{ChIP} + \text{Input})$. Tool-specific regions are defined to be not represented in a "unified *bona-fide* benchmark set". Wilcoxon signed-rank Test: "***" (P≤0.001), "**" (P≤0.01), "*" (P≤0.1) and "n.s." (P>0.05).

competitor methods contain many regions with read coverage and fold changes less than the genomic average. These observations support the validity of enrichR-specific regions. Furthermore, enrichR-specific regions were remote from unified benchmark regions (median=14Mb) and, yet, still overrepresented in annotated gene bodies (odds-ratio=13; Table B.1).

Robustness. Some ChIP-seq peak callers perform less well when the sequencing depth in the ChIP library is reduced [103]. To evaluate the robustness of enrichment calling methods, once again a unified *bona-fide* validation set was used to benchmark all tools on an *in silico* down sampled sequencing library (Methods Section 4.2.3). All methods, except CisGenome, recovered ≥70% of the H3K4me3 unified *bona-fide* benchmark (108,834 500bp regions; 54Mb) at a sequencing depth reduced by 80% (Fig. 4.7A). Only enrichR, MACS2 and BCP could achieve ≥90% precision at ∼50% of the original H3K4me3 sequencing depth. For H3K36me3, DFilter, CisGenome and SPP showed inconsistent performances (≤35% recovered) whereas enrichR, MACS2, BCP and MUSIC recovered ≥90% of the unified *bona-fide* benchmark (354,527 1kb regions; 355Mb) at a sequencining depth of 30% (Fig. 4.7B). Note, for sequencing depths far below 15%, enrichR

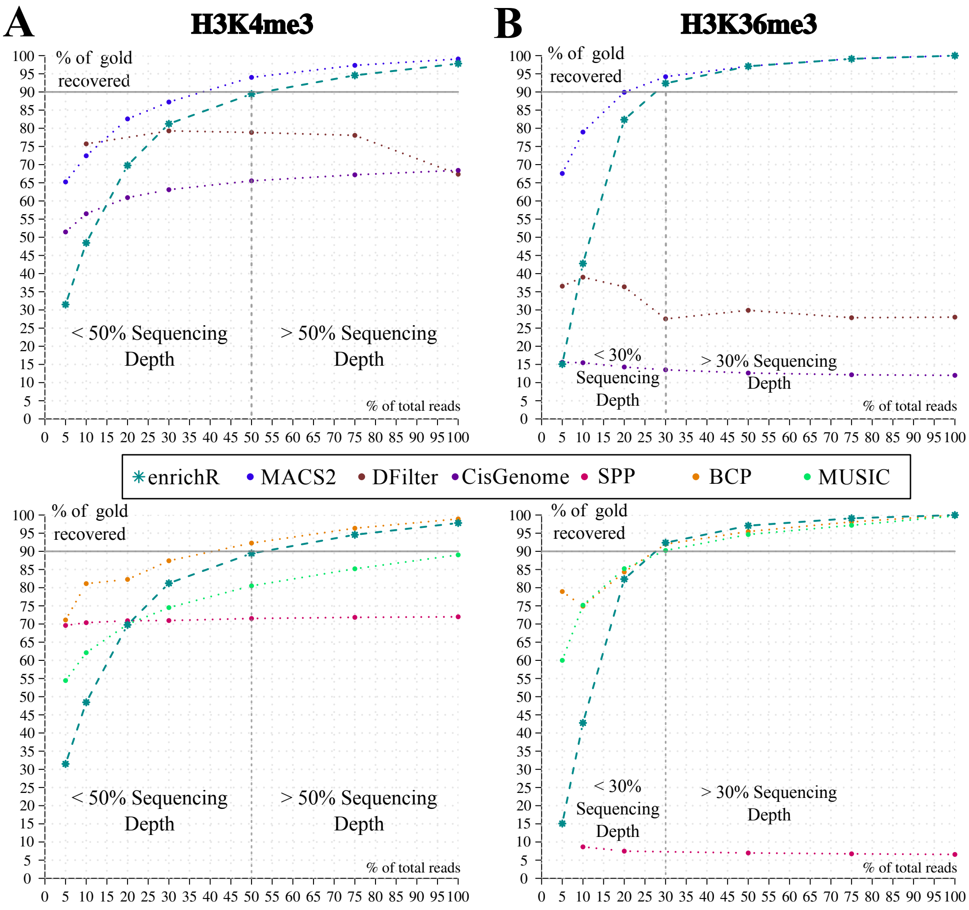

Fig. 4.7 – **Saturation Analysis Based on a Unified *Bona-Fide* Benchmark Set for enrichR, MACS2, DFilter, CisGenome, SPP, BCP and MUSIC.** Sequencing libraries of ChIP and Input were downsampled *in silico* (Methods Section 4.2.3). (A) For H3K4me3, all methods worked well. enrichR, MACS2 and BCP captured ≥90% of the unified benchmark set with 50% of the reads from the original library. (B) For H3K36me3, DFilter, CisGenome and SPP were inconsistent. enrichR, MACS2, BCP and MUSIC recovered ≥80% at 20% of the original sequencing depth.

filters rigorously for low power regions with the T Filter to avoid low power calls. In summary, enrichR, MACS2 and BCP were precise with respect to a consensus-vote inferred “gold-standard” in ChIP-seq libraries with a strongly reduced sequencing depth.

4.3.2 `enrichR` Normalization Corresponds to Published *In Silico* as well as *In Vitro* Normalization Methods

Here, I compare the normalization estimated by enrichR to two competing approaches: (i) the *in silico* estimation of c_B with NCIS [85]; and (ii) the *in vitro* spike-in-based estimation of the enrichment factor $\langle f \rangle$ in Internal Standard Calibrated ChIP (ICeChIP) [93]. The NCIS normalization is also based on *bona-fide* background regions. The method estimated a normalization factor

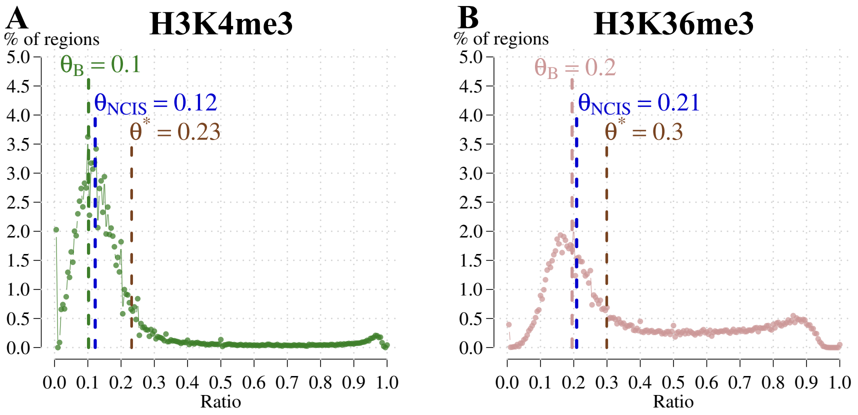

Fig. 4.8 – enrichR Normalization Improves on NCIS' Estimated Normalization Factor towards a Correct Background Estimation. (A-B) H3K4me3 (A) and H3K36me3 (B) ChIP over Input ratios are plotted. Genome-wide average enrichent (θ^*), enrichR background estimation (θ_B) and NCIS normalization (θ_{NCIS}) are given. NCIS correctly accounted for enrichment in the data in estimating the normalization factor.

that was >1.5-fold smaller than θ^* suggesting that NCIS also accounted for the effect of enrichment towards a correct background normalization. Interestingly, the NCIS estimates (H3K4me3: 0.14, H3K36me3: 0.263) were still ~1.35-fold greater than enrichR's estimate for both H3K4me3 (0.103) and H3K36me3 (0.195). A visual inspection indicated that NCIS over-estimated the ratio in the true background population (Fig. 4.8A). This over-estimation was more pronounced for H3K36me3 than for H3K4me3 (Fig. 4.8B). Despite different underlying models both normalizations account for the effect of enrichment on the overall read statistics, yet enrichR seemed more accurate.

In ICeChIP the ChIP-seq read coverage is transformed into a Histone Mark Density (HMD%) with the help of spiked-in modified nucleosomes reconstituted from recombinant and semi-synthetic histones on barcoded DNA. The ICeChIP normalization makes use of an assumed *putative* linear relationship between the amount of epitope present and corresponding ChIP-signal intensity. The HMD% per bp i is defined as

$$\text{HMD\%}_i = 100\% * \frac{\frac{\text{ChIP-coverage}}{\text{Input-coverage}}}{\langle \text{IP}_{\text{Ladder}} \rangle} \tag{4.2}$$

where $\langle \text{IP}_{\text{Ladder}} \rangle$ is the regression coefficient in the spike-in IP enrichment ladder. For comparison, I determined the average enrichment $\langle f \rangle$ with enrichR in four mouse embryonic stem cell (mESC) ICeChIP-seq data sets. When compared to $\langle f \rangle$, the $\langle \text{IP}_{\text{Ladder}} \rangle$ was greater for H3K4me3 (Table 4.2). However, for H3K36me3, H3K79me2, H3K27me3 and H3K9me3 $\langle \text{IP}_{\text{Ladder}} \rangle$ was smaller the enrichR's $\langle f \rangle$ suggesting that $\langle \text{IP}_{\text{Ladder}} \rangle$ and $\langle f \rangle$ are not directly comparable. A finding that is also true for the actual scaling factors $\langle \text{IP}_{\text{Ladder}} \rangle^{-1}$ and $(\log\langle f \rangle)^{-1}$. This dis-

Experiment	$\langle \mathrm{IP}_{\mathrm{Ladder}} \rangle$	$\langle f \rangle$	$\langle \mathrm{IP}_{\mathrm{Ladder}} \rangle^{-1}$	$(\log\langle f \rangle)^{-1}$	α_{ChIP}	α_{Input}
H3K4me3	29.03	18.77	0.04	0.34	19.85	103.78
H3K36me3	0.28	1.56	3.70	2.24	57.88	200.89
H3K79me2	0.15	1.64	7.69	2.03	50.15	204.17
H3K27me3	0.70	1.68	1.49	2.05	49.18	210.13
H3K9me3	1.36	1.68	0.73	1.93	51.19	211.01

Table 4.2 – ICeChIP's Enrichment Factor $\langle \mathrm{IP}_{\mathrm{Ladder}} \rangle$ is Not in Concordance with enrichR's $\langle f \rangle$. When compared, $\langle f \rangle$ is ~1.5-fold smaller in H3K4me3 and ≥1.2-fold greater in H3K36me3, H3K79me2, H3K27me3 and H3K9me3 where $\langle f \rangle$ is also quite similar for the latter histone modifications. The actual scaling factors are given by $\frac{1}{\langle \mathrm{IP}_{\mathrm{Ladder}} \rangle}$ and $\frac{1}{\log\langle f \rangle}$ where, again, a disparity between the two methods is observed.

crepancy may be related to the *wrong* assumption of a linear relationship between epitope and ChIP-signal intensity in the determination of $\langle \mathrm{IP}_{\mathrm{Ladder}} \rangle$. Thus, this assumption in ICeChIP needs more in-depth investigation, *e.g.* by extending the spike-in ladder over a greater range of epitope quantities.

To further analyze the disparity between enrichR's *in silico* normalization and ICeChIP's *in vitro* spike-in based normalization, I studied the enrichR enrichment and ICeChIP-reported HMD% directly. Strikingly, a high correlation (≥0.81) was observed for all five ICeChIP experiments which indicated that enrichR's e and ICeChIP's HMD% are in concordance (Fig. 4.9A). A simple linear model could explain the relation in all data sets very well – yet, to a lesser extent for H3K4me3 where the relation seemed non-linear. Note that the HMD% calculation tends to inflate ChIP/Input fold changes in low count regions whereas the enrichR approach penalizes the fold changes in those regions by adding model-derived pseudo counts α_{ChIP} and α_{Input} (see Methods Section 4.2.2). To analyze this potential discrepancy further, I studied the studentized residuals of the linear model fit in relation to the raw ChIP and Input read counts (Fig. 4.9B). Throughout all experiments, the variances of residuals were greatest in regions with ChIP and/or Input read counts smaller than enrichR's inferred pseudo counts α_{ChIP} and α_{Input}. This observation confirmed that HMD% inflates fold changes of regions with low statistical power. In a final assessment, ICeChIP's $\langle \mathrm{IP}_{\mathrm{Ladder}} \rangle$ and enrichR's $\langle f \rangle$ in relation to the fragment coverage per bp in ChIP and Input. Both enrichment factor estimates represented the maximal enrichment observed in H3K4me3 and H3K9me3 well (Fig.4.9C). However, only $\langle f \rangle$ estimated the maximal enrichment correctly in H3K36me3, H3K79me2 and H3K27me3 whereas ICeChIP's $\langle \mathrm{IP}_{\mathrm{Ladder}} \rangle$ under-estimated the maximal enrichment to a large degree. This under-estimation of the maximal enrichment leads to a further inflation of many HMD% estimates far beyond 100% and, in consequence, complicates the interpretation of ICeChIP's HMD% for these histone modifications. Taken together, the enrichR normalization compares favorably to previously published *in silico* as well as *in vitro* normalization methods in terms of its correct background and foreground enrichment estimation, respectively.

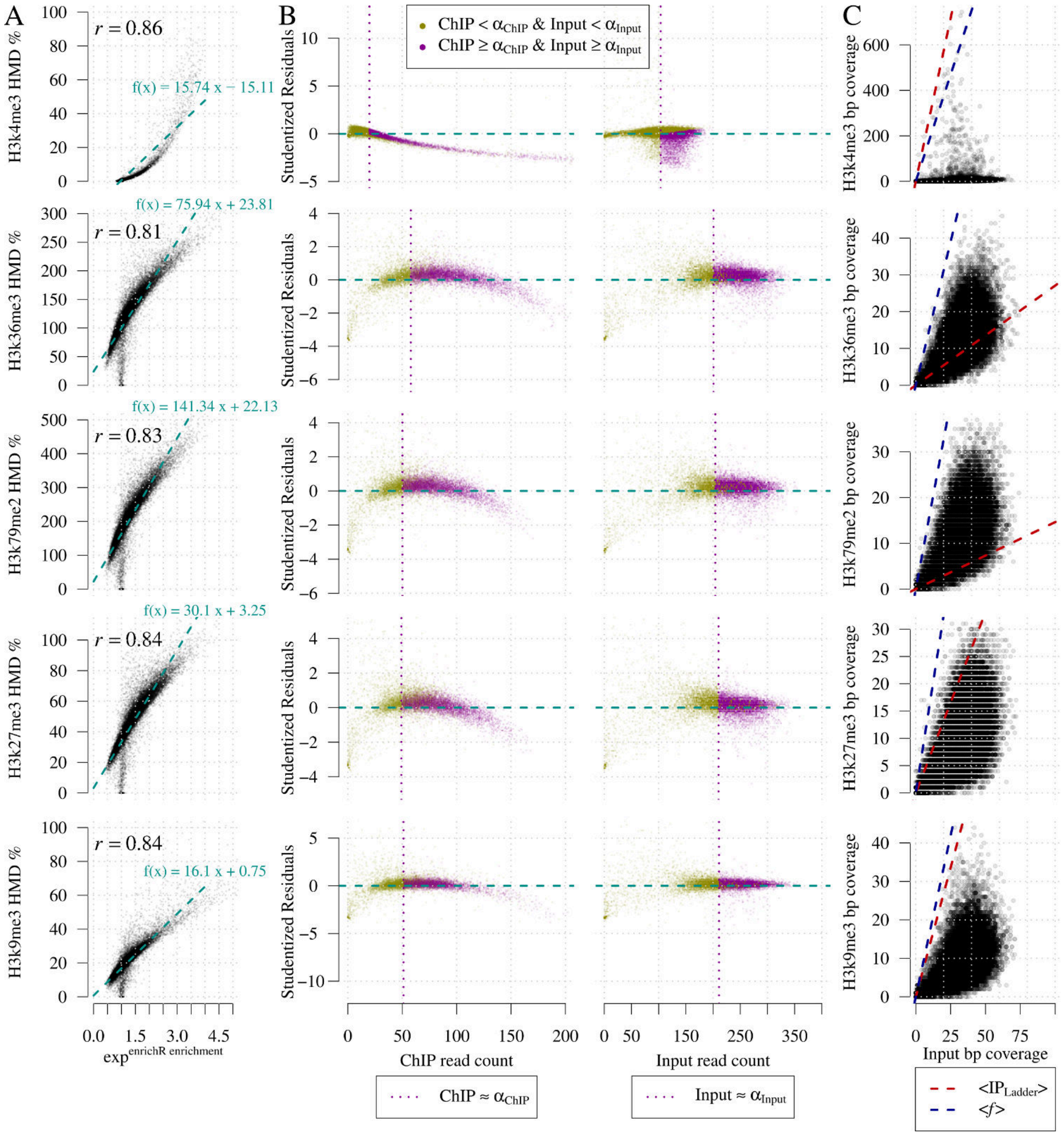

Fig. 4.9 – enrichR Enrichment e and ICeChIP's HMD% are Strongly Correlated but HMD% Incorrectly Inflates Low Count Fold Changes. (A) HMD% and enrichR enrichment positively correlate and this relation can be well explained by a simple linear model for H3K4me3, H3K36me3, H3K79me2, H3K27me3 and H3K9me3 (regression line in green). (B) Spread of model residuals are smallest for read counts greater than enrichR-estimated pseudo counts, *i.e.* α_{ChIP} and α_{Input} (marked in purple). An observation indicative for the HMD% to inflate the fold change in low power regions. (C) ChIP fragment coverage per bp plotted against Input coverage per bp. ICeChIP's $\langle \text{IP}_{\text{Ladder}} \rangle$ (red) and enrichR's $\langle f \rangle$ (blue) are indicated by dashed lines. $\langle \text{IP}_{\text{Ladder}} \rangle$ and $\langle f \rangle$ correctly estimate the putative maximal enrichment in H3K4me3 and H3K9me3. However, $\langle \text{IP}_{\text{Ladder}} \rangle$ largely under-estimates the putative maximal enrichment in H3K79me2, H3K36me3 and H3K27me3 which results in HMD% up to 5-fold greater than 100% (see also (A)).

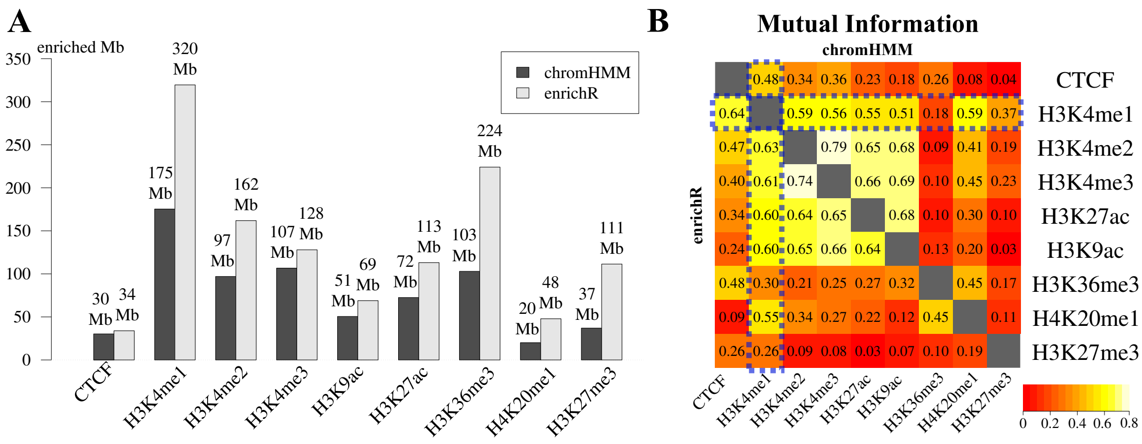

Fig. 4.10 – chromHMM Binarization Input: enrichR Improved on the Mutual Information between Histone Modifications Associated to Active Transcription in GM12878 ChIP-seq Data. (A) enrichR called up to 3-fold more enrichment in ENCODE GM12878 ChIP-seq data. (B) Covariances of enrichment $(0, 1)$-matrices showed a high mutual information (MI) of H3K4me1/2/3, H3K27ac and H3K9ac in the chromHMM (upper triangular matrix) and enrichR binarization (lower triangular matrix). In the enrichR binarization, the MI of H3K4me1 to euchromatic histone modifications (H3K4me2/3, H3K27ac, H3K9ac, H3K36me3) increased but its MI decreased for repressive marks (H4K20me1, H3K27me3).

4.3.3 Improved Chromatin Segmentation with an `enrichR`-chromHMM Hybrid Approach

Given a set of ChIP-seq experiments, the task of chromatin segmentation aims to assign most of the genome to meaningful chromatin states, *i.e.* reoccurring patterns of coinciding enrichment in the ChIP-seq data. I refer to this task as “resolving the epigenome into chromatin states”. The chromHMM method [36] uses a Hidden Markov Model to segment the genome in distinct epigenetic states based on enrichment calls in a set of ChIP-seq experiments. Hither to shown was the correctness of the enrichment calling and the normalization capabilities of enrichR. In this last section, a previously developed technique, namely the chromHMM chromatin segmentation, is augmented by sensitive enrichment calling performed by enrichR. To this extent, I developed an enrichR-chromHMM hybrid approach that, first, calls enrichment with enrichR (**binarization**) and; second, computes a chromatin segmentation with chromHMM’s Hidden Markov Model based on these enrichment calls (**segmentation**). I applied this enrichR-chromHMM hybrid as well as the conventional chromHMM approach on published ENCODE ChIP-seq Data for CTCF, H3K4me1/2/3, H3K27ac, H3K9ac, H3K36me3, H4K20me1 and H3K27me3 in the lymphoblastoid cell line GM12878 (Methods Section 4.2.6).

On the level of **binarization**, enrichR identified more ChIP-seq enrichment than the chromHMM binarization and improved on the mutual information (MI) between certain histone modifications associated to active transcription. enrichR could increase the number of enriched regions by up to 3-fold (*e.g.* H3K27me3) when compared to the conventional chromHMM binarization approach based on a Poisson background model (Fig. 4.10A). Nevertheless, “more” enrichment calls do not imply “more meaningful“ enrichment calls. To study the validity of this gain in enrichment calls, the MI (see [89]) of the $(0, 1)$-matrix was computed where 0 and 1 indicate background and enriched, respectively. A high MI of H3K4me1/2/3, H3K27ac and H3K9ac which is indicative for active promoters was already apparent in the chromHMM enrichment calls (Fig. 4.10B). However, enrichR enrichment calling increased the MI between modifications associated to active chromatin, *e.g.* H3K4me1 and CTCF, and decreased the MI between those marks and repressive modifications, *e.g.* H3K4me3 and H3K27me3. Thus, enrichR enrichment calling is a promising augmentation to the conventional chromHMM binarization approach.

On the level of **segmentation**, the enrichR binarization improved the chromatin segmentation on two levels: (i) by identification of a previously undetected poised enhancer state and; (ii) by dissecting a large unresolved state into a elongation-associated state as well as into a lamina-associated and a lamina-distal heterochromatic state. When compared to the conventional chromHMM approach, the enrichR-chromHMM hybrid evoked many similar chromatin states stratified by a hierarchical clustering of a combined emission matrix (Fig. 4.11A). The majority of chromatin states covered only a small fraction ($\leq$3.9%) of the genome in both approaches (Fig. 4.11B). Nevertheless, the enrichR binarization in the hybrid approach led to the identification of a chromatin state with low prevalence ($\sim$70Mb; “enrichR | state 14a” highlighted in green) that is characterized by H3K4me1/2 as well as H3K27me3 enrichment. Given that this state is found predominantly proximal to genes and it is characterized by H3K4me1 and H3K27me3 alike suggest that it marks *putative* poised enhancers (see [124] for review). The conventional chromHMM approach reported a large state ($\sim$2.2Gb; “chromHMM | state 15”) that is associated neither to CTCF nor to any chromatin modification. On the contrary, the enrichR binarization led to a reduction of this state by almost 2-fold ($\sim$1.2Gb; “enrichR | state 15b”; Fig. 4.11B). These states are associated to the cell lamina and most likely represent stably silenced regions in telomeric and/or centromeric regions, *i.e.* constitutive heterochromatin. Those genomic loci are hard to assay in ChIP-seq due to repeats in their DNA sequences. Interestingly, the enrichR binarization led to the dissection of the “chromHMM | state 15” into three chromatin states (Fig. 4.11D): the larger aforementioned heterochromatic state (“enrichR | state 15b”), a $\sim$580Mb state marked by H3K27me3 (“enrichR | state 14b”, marked in yellow) and a $\sim$456Mb state characterized by H3K4me1, H3K36me3 and H3K27me3 (“enrichR | state 15a”, marked in pink). The presence of H3K27me3 in “enrichR | state 14b” and its association to the lamina suggest that this state is, in fact, representing stably silenced heterochromatin. This H3K27me3 enrichment was not detectable with the conventional chromHMM binarization. As suggested by the presence of H3K36me3, “enrichR | state 15a” may be associated to active transcription because this state is found enriched in gene bodies distant to the cell lamina (Fig. 4.11C). It remains further investigation if this state describes genes with allele-specific or sample-heterogeneous expres-

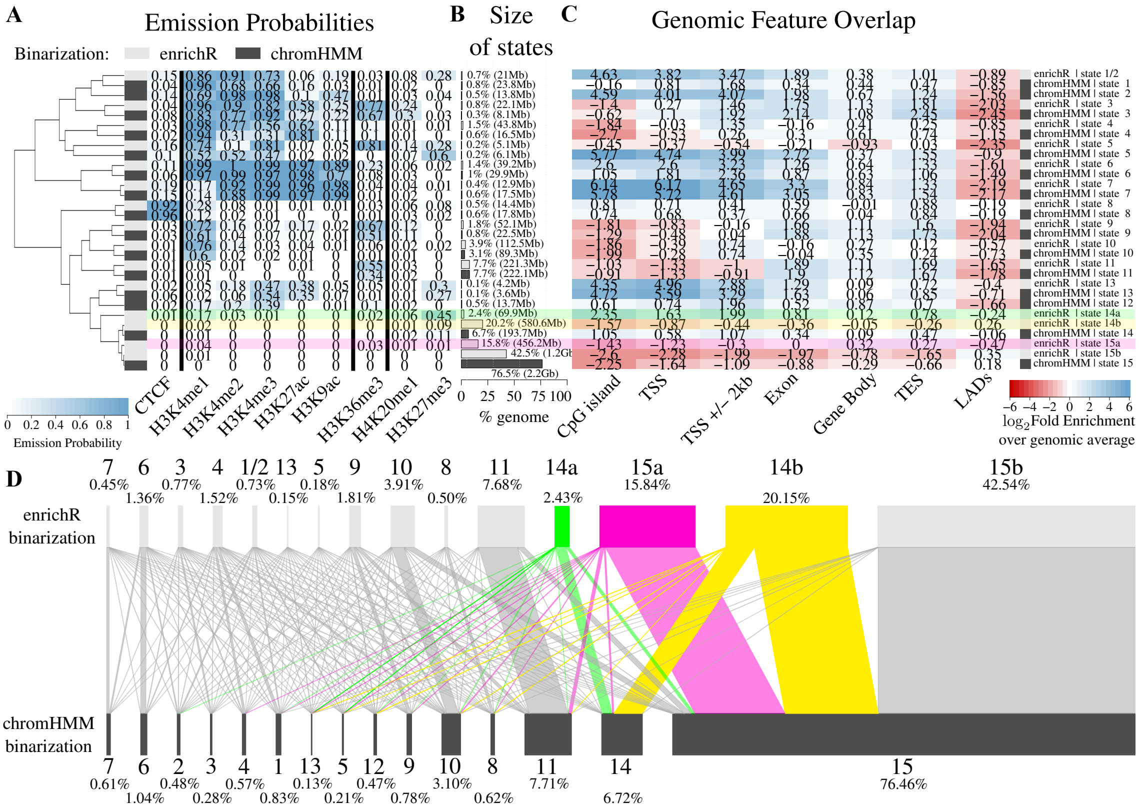

Fig. 4.11 – chromHMM Segmentation Output: An enrichR-chromHMM Hybrid Approach Increased Resolution of Chromatin Segmentation in GM12878 Cells. (A) A hierarchical clustering on the combined emission probabilities reported by the conventional chromHMM approach (dark grey) and the enrichR-chromHMM hybrid (light grey) indicates a good agreement between both approaches but also highlights differences. For example, enrichR state 14a (green overlay) is characterized by H3K4me1/2 and H3K27me3 simultaneously and does not cluster with a chromHMM state. The large enrichR state 15a (pink overlay) is characterized by H3K36me3 and H3K4me1 and not found in the conventional chromHMM approach. (B) Most states in both approaches are of low prevalence. The enrichR-chromHMM hybrid resolves the majority of the genome whereas the conventional approach leaves >75% unresolved. (C) $\log_2$ fold enrichments for genomic features highlights again the consistency of the hybrid approach. "enrichR | state 14b" (yellow) is enriched at lamina associated domains (LADs) and "enrichR | state 15a" (pink) describes a putative gene body-associated state distant from the cell lamina. (D) A bipartite graph of segmentations reported by both approaches shows coherent results between the two. It also highlights how the enrichR-chromHMM hybrid approach dissects the large unresolved state reported by the conventional approach into three states.

sion patterns. Taken together, the enrichR-chromHMM hybrid approach increased the mutual information of reported transcription-associated histone modifications in the binary input matrix. The more sensitive enrichment calling by enrichR enabled the hybrid approach to detected a *putative* poised enhancer state and to dissect a previously unresolved state into two meaningful chromatin states.

4.4 Discussion

In summary, the application of a simple two-component implementation of the “normR” framework is represented is represented by “enrichR”. By modeling foreground and background jointly, normalization and enrichment calling are performed simultaneously. The implicit modeling of the effect of enrichment on the overall read statistics results in an adequate normalization factor that increases the sensitivity in detecting shallow differences in ChIP enrichment while, at the same time, maintaining high specificity. The enrichR method can readily be used as a self-contained software package in the extensive analysis of ChIP-seq data in epigenetic studies.

The enrichR method facilitates enrichment calling in high and low signal-to-noise ratio (S/N) ChIP-seq data alike. The ChIP-seq enrichment calling for the *localized* histone modification like H3K4me3 is successfully achieved by enrichR and by most of its competitors. Auxiliary information such as DNA-hypomethylation and transcriptional activity support the validity of those enrichment calls. As opposed to H3K4me3 which has a high S/N, *delocalized* histone modifications like H3K36me3 impede enrichment calling because of a low S/N in the ChIP-seq data. Herein, enrichR outperforms other approaches in terms of robustness to a reduced sequencing depth and accuracy of classification as scored under a novel *bona-fide* ChIP-seq benchmark, *i.e.* a trustworthy validation set, derived from a set of seven enrichment calling approaches. Furthermore, enrichR recovers many regions with promiscuous H3K36me3 enrichment that other methods miss. Once again, auxiliary information like DNA-hypermethylation and substantial fold enrichment over Input could confirm the genuineness of these calls. The superior performance of enrichR is attributed to the sensitive normalization technique which accounts not only for varying sequencing depth but specifically addresses the effect of ChIP enrichment on the overall read statistics. In the future, enrichR can be used to detect regions with low ChIP-seq enrichment like chromatin modifications in heterogeneous samples or low affinity protein binding sites wherein the signal level does not pass the detection level of precedent peak calling methods. Revising canonical transcription factor binding sites might be considered if they are based on peaks called by those in-sensitive methods.

Aside from a correct enrichment calling, enrichR improves on current *in silico* as well as *in vitro* ChIP-seq normalization methods. The application of enrichR to a recently reported spike-in controlled ICeChIP-seq data set revealed a peculiar disparity in the estimation of a correct enrichment factor to infer the histone modification density. The assumption of the *putative* linear relationship between the quantity of the epitope spike-in and the ChIP-seq signal intensity is incorrect. In consequence, the initial spike-in derived normalization factor $\langle \mathrm{IP}_{\mathrm{Ladder}} \rangle$ underestimates the maximal enrichment in three out of five experiments. Strikingly, enrichR’s enrichment factor $\langle f \rangle$ correctly estimated the maximal enrichment in all experiments.

As a proof of concept, a chromatin segmentation based on enrichR enrichment calls improves the resolution of the segmentation result. Whereas the conventional segmentation approach by chromHMM leaves ∼75% of the epigenome unresolved, an enrichR-chromHMM hybrid approach resolves the majority of the genome into previously undetected meaningful states. For example, a detected *putative* poised enhancer state is a promising candidate for future investigation. Moreover, the dissection of the large previously unresolved heterochromatin state into a Lamina-associated and a Lamina-dissociated heterochromatic states supports the theory of distinct repressive mechanisms of these forms of chromatin organization (for review see [37]). Thus, I envision how enrichR augments today's epigenetic analyses ranging from clustering to visualization.

In the next two chapters I am going to present two more sophisticated applications of the normR framework in ChIP-seq data analysis. A simple augmentation to the enrichR model allows for the identification of distinct H3K27me3 and H3K9me3 enrichment regimes that are ultimately linked to formation of facultative and consecutive heterochromatin, respectively. Furthermore, I will study the effects induced by the immortalization of primary hepatocytes with a normR incarnation that compares to ChIP-seq tracks without the need of an Input experiment.

Chapter 5

regimeR – Regime Enrichment Calling in ChIP-seq Data

This chapter introduces "regimeR" – an expansion of the enrichR model that dissects canonical signals in ChIP-seq data into distinct regimes of different enrichment levels. The regimeR method is described via an application to low S/N ChIP-seq data of the heterochromatic histone modifications H3K27me3 and H3K9me3 in HepG2 cells. The regimeR-based analysis identified *peak*, *i.e.* high, enrichment regions to be associated to higher levels of methyltransferase binding for those marks and, also, to be embedded within regions *broad*, *i.e.* low, enrichment. These results and distinct sequence features suggest that the heterochromatin *peak* regions resemble nucleation sites for gene repression in silenced regions of facultative (H3K27me3) and consecutive (H3K9me3) heterochromatin.

5.1 Introduction

In the previous chapter, enrichR was shown to confidently identify genomic loci in ChIP-seq data that harbor a statistically significant ChIP enrichment given a background model which is accurately fit using the normR framework. A comparison to previously developed approaches for enrichment calling confirmed that this background normalization aided the identification of ChIP-seq signals with low signal over background. Now the question arises whether those sites of low ChIP enrichment are qualitatively different from the "canonical" protein binding sites with high enrichment – apart from the difference in their average enrichment level. To the best of my knowledge, there is no precedent methodology reported in the scientific literature that solves the principled discrimination of ChIP-seq enrichment regimes.

ChIP-seq signal intensities are predictive for quantitative outputs like gene expression [34] which is indicative of the quantitative information inherent to ChIP-seq data. It would be favorable to correlate the normalized ChIP-seq signal intensity to the prevalence of a protein binding event in the sample cell population. In fact, the acquisition of very homogeneous cell populations is the exception rather than the rule and, if not taken into account, the sample heterogeneity leads to spurious results in downstream analyses of protein binding sites [93]. A conventional in-sensitive enrichment calling approach is doomed to exclusively identify genomic loci bound in the majority of sample population which harbor a predominantly high ChIP enrichment. In a very heterogeneous sample population, there may exist only a few of these regions and, thus, it is an asset to also identify regions of weak protein binding. A qualitative **separation of the signal-regions into low (*broad*) and high (*peak*) enrichment regions** aids the differentiation of genomic loci that are bound by the protein in only a small sub-population of cells from those that are bound in most cells, respectively. This analysis of heterogeneous protein binding can not be performed using existing methods.

The qualitative differences of broad and peak protein binding events can be stratified, for example, by a significantly different co-occupancy with a known interacting protein or by distinctive sequence features of the underlying DNA at those sites. For instance, genomic loci harboring a certain histone modification should be associated with the presence of an enzyme catalyzing this modification, *e.g.* H3K27me3 is deposited by EZH2 [125–128]. To this extent, H3K27me3-enriched regions should coincide with EZH2 binding – specifically, H3K27me3-*broad* regions should show lower EZH2 binding than H3K27me3-*peak* regions. In addition, *broad* and *peak* regions can also be studied on the basis of determining genomic features like the CpG content [129, 130] or conservation [131]. Thus, by compiling multiple auxiliary data a biological relevance can be assigned to the two protein binding regimes of broad and peak signals.

To study the prevalence of protein binding events via ChIP-seq, I used another implementation of the normR framework called **"regimeR"** which classifies significantly enriched regions into multiple enrichment regimes. As a proof of principle, I applied regimeR with two enrichment components to two heterochromatic histone modifications H3K27me3 and H3K9me3 in the HepG2 hepatocarcinoma cell line to study the preservation of facultative and constitutive heterochromatin, respectively. The low S/N nature of these marks generally impedes enrichment calling with conventional approaches (see Section 3.2.2) but, as shown before, the normR framework provides high sensitivity in this setting (see Chapter 4). In fact, regimeR successfully dissected the ChIP-seq enrichment therein into two distinct enrichment regimes of low and high signal over background. The regimeR-based study found that H3K27me3 *peak* regions are specifically embedded within regions of *broad* H3K27me3 enrichment and that these peaks, in comparison to their flanking *broad* regions, were associated to CpG-dense genomic loci preferentially bound by the H3K27 methyl transferase EZH2. Similarly, H3K9me3 *peak* regions were also flanked by *broad* H3K9me3 domains. The H3K9me3 *peak* regions were coincident with repetitive elements and the binding of ZNF274 – a transcription factor that recruits the H3K9 methyltransferase SETDB1 [132]. Taken together, these distinct regions of high H3K27me3 and H3K9me3 ChIP-seq enrichment resemble nucleation sites of broad facultative and constitutive heterochromatin, respectively. Thus, the regimeR approach enables for an unprecedented stratification of sample heterogeneity in ChIP-seq experiments to study differences of protein binding with low and high propensity in the sample population.

5.2 Methods

This section details the processing and the quality of the sequencing data used. Next, the normR framework is adapted to the tasks of ChIP-seq regime enrichment identification which is referred to as “regimeR”. In the last section, the steps that facilitate the analysis of DNA sequence features are outlined.

5.2.1 Data Sets

ChIP-seq Data. Paired end reads from Input, H3K27me3 and H3K9me3 ChIP-seq for HepG2 cells and Primary Human Hepatocytes (PHH) were processed and quantified as described in Section 4.2.1. EZH2 (`GSM1003576`), ZNF274 (`GSM935350`) ChIP-seq alignments and the respective control alignment (`GSM733780`) were downloaded from the UCSC ENCODE DCC repository [43] under `hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/`. Enrichment over background estimated with `enrichR`() in 1kb bins for reads with Mapping Quality $\geq$30 shifted by 100bp (see Chapter 4).

5.2.2 The `normR` Methods: `regimeR`

The normR framework (see Chapter 3) was adapted to calling enrichment regimes in ChIP-seq Data, referred to as **“regimeR”**. Now, $m = 3$ mixture components were used, *i.e.* background B and two foreground components F_1 (*broad* enrichment) and F_2 (*peak* enrichment), to normalize and call enrichment regimes over Input. The number of free parameters is now increased to five in comparison to enrichR (see Chapter 4), namely $\theta = \{\theta_B, \theta_{F_1}, \theta_{F_2}\}$ and $\pi = \{\pi_B, \pi_{F_1}, \pi_{F_2}\}$. Note, π_{F_2} is simply $(1 - \pi_B - \pi_{F_1})$. Given the normR model in Equation (3.2) and the normR likelihood function defined by Equation (3.3) the following “regimeR” likelihood function can be derived:

$$\mathcal{L} = P(\pi, \theta | s_i, r_i) = \prod_i \binom{s_i + r_i}{s_i} \sum_{k \in \{B, F_1, F_2\}} \pi_k \cdot \theta_k^{s_i} \cdot (1 - \theta_k)^{r_i}$$

where s_i (r_i) corresponds to the number of reads in the ChIP (Input) for non-overlapping, fixed size regions $i = 1, \ldots, n$. Parameters are fitted by running the EM algorithm as described for enrichR in Section 4.2.2. Identically to enrichment calling with enrichR, the background model B is set to the mixture component k with smallest θ_k. To identify significantly enriched regimes, regions significantly different from background are recovered as described before in Section 4.2.2. Next every bin j that is significantly enriched is assigned to one of the two enrichment regimes by *Maximum A Posteriori*

$$z_j = \underset{k \in \{F_1, F_2\}}{\arg\max} \left(\frac{P(\pi_k, \theta_k | s_j, r_j)}{P(\pi, \theta | s_j, r_j)} \right),$$

where z_j indicates the assignment of a significantly enriched region j to either F_1 or F_2.

The regularized ChIP-seq enrichment e^* is computed as reported for enrichR (Section 4.2.2). To account for the maximal ChIP enrichment, e^* is normalized with the $\log$ of the average enrichment factor for F_2

$$\langle f_2 \rangle = \frac{\theta_{F_2}}{1 - \theta_{F_2}} \cdot \frac{1 - \theta_B}{\theta_B}$$

to obtain a regularized and normalized enrichment e:

$$e_i = \frac{e_i^*}{\log \langle f_2 \rangle}.$$

These routines were implemented in the normR R package [4] as the `enrichR`()-function (see also R code snippet below).

These routines were implemented in the normR R package [4] as the `regimeR`()-function (see R code below). Note that, in principle, an arbitrary number of components representing background plus a fixed number of enrichment regimes is possible in `regimeR`().

The same read counting configuration as before was used (see Section 4.2.2). Read counts obtained from H3K27me3 and H3K9me3 ChIP-seq experiments in HepG2 cells were modeled in regimeR with 3 components: (i) background (no enrichment); (ii) *broad* regions (low enrichment) and; *peak* regions (high enrichment) with `regimeR`() in R:

```
regimes <- regimeR(
  treatment = "ChIP.bam",
  control = "Input.bam",
  genome = genome,
  models = 3,
  countConfig = countConfig,
  procs = 8
)
```

Bins with q-value≤0.1 (FDR=10%) were called enriched and assigned to an enrichment component by *Maximum A Posteriori* in `regimeR`(). Results were exported to bed using normR's `exportR`() function:

```
exportR(
  x = regimes,
  filename = "regimes.bed",
  type = "bed",
  fdr = 0.1
)
```

5.2.3 Validation of regimeR Calls via Sequence Features

CpG Odds Ratio. The CpG odds ratio for a genomic region i was calculated as the ratio of the number of observed CpGs "#CpG" therein over the number of expected CpGs under *i.i.d.*, *i.e.* $(\#\mathrm{C}_i \cdot \#\mathrm{G}_i)$:

$$\text{CpG-odds}_i = \frac{\#\mathrm{CpG}_i}{\#\mathrm{C}_i \cdot \#\mathrm{G}_i}.$$

Conservation. PhastCons100way [133, 134] and PhyloP100way scores were downloaded from UCSC. The maximum value in a 1,000bp window was cataloged.

Repetitive Elements. RepeatMasker [135] annotations for hg19 were downloaded from `repeatmasker.org/species/hg.html`. Repetitive elements flagged as "simple/tandem repeats" or "low complexity regions" were filtered out and only repeats with "score"≥1,000 were considered.

KRAB-ZNF Motif Scanning. To identify potential binding sites for Krüppel-associated box zinc-finger transcription factors that correlate with H3K9me3 peaks, I searched footprintDB [31] "DNA Binding Motifs" for the term "ZNF" and out of 31 motifs retained all motifs annotated as "KRAB box" (5 motifs: ZNF263, ZNF274, ZNF306, ZNF354C, ZNF713). I used the MEME SUITE's [136] FIMO routine [137] to identify genome-wide occurrences of these motifs. Reported p-values give the probability of a random DNA sequence of the same length as the respective motif to match the occurrence with the same or a better score, *i.e.* sum of "used" entries in position-dependent scoring matrix of the motif.. I retained only motifs with FDR≤0.01 – only ZNF274 and ZNF263 passed this threshold:

```
awk -F"\t" \
  ’OFS="\t" { if ($8 <= .01) {print $2,$3,$4,$1"(q="$8")",(30*$6),$5}}’ \
  FIMO_0All.IDs.txt | sort-bed - > FIMO_0All.IDs.qsig.bed
```

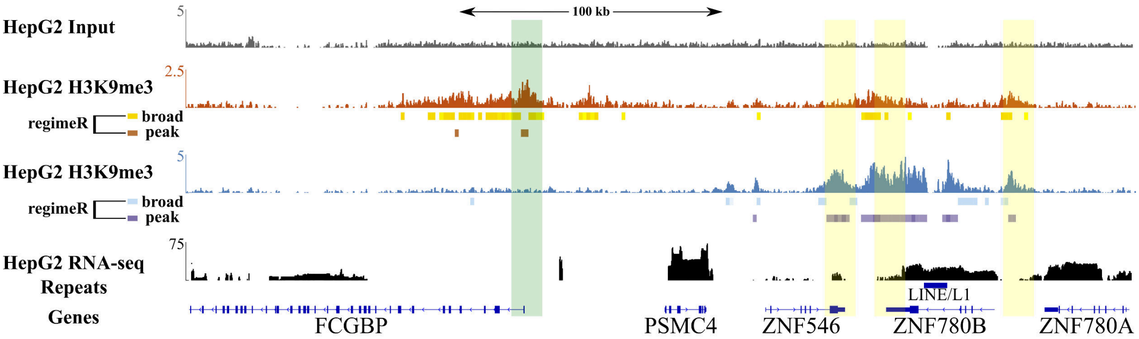

Fig. 5.1 – regimeR Identifies H3K27me3 and H3K9me3 Enrichment Regimes at the FCGBP Gene Locus. Input (grey), H3K27me3 (orange), H3K9me3 (blue) and RNA-seq (black) coverage around a ZNF cluster on chromosome 19 in HepG2 cells. Individual regimeR-computed regimes are displayed as boxes below respective tracks. Repressed promoters of the FCGBP are marked by a H3K27me3 peak within a *broad* H3K27me3 domain (green overlay). The 3'-ends of ZNF genes are marked with high H3K9me3 enrichment (yellow overlay).

5.3 Results - Distinct Heterochromatic Enrichment Regimes

Hither to discussed was the applicability of normR to a well-studied problem: the discrimination of enrichment against background. Here, a problem is studied for which I did not find to the best of my knowledge a precedent in the literature: the discrimination of low enrichment from high enrichment. This problem can easily be addressed by increasing the number of foreground components in enrichR from one single component to multiple components (see Methods Section 5.2.2) – an approach referred to as **"regimeR"**. In the case of two foreground components, regimeR discriminates a *peak* regime (high enrichment) and a *broad* regime (low enrichment) over the background. Here, I applied regimeR to H3K27me3 and H3K9me3 ChIP-seq data from the hepatocarcinoma cell line HepG2 and studied distinctive features of the two ChIP enrichment regimes therein.

Fig. 5.1 depicts a representative region of Human chromosome 19 harboring active and repressed genes. Therein, regimeR segmented the ChIP-seq enrichment into *broad* and *peak* regions. For example, a H3K27me3 peak was identified at the most upstream promoter of the "Fc Fragment Of IgG Binding Protein" gene (FCGBP). This promoter is repressed as confirmed by RNA-seq read coverage. For H3K9me3, regimeR identified three peaks that are flanked by *broad* H3K9me3 enrichment at the 3'-ends of zinc finger genes ZNF546 and ZNF780A/B which are reported to be bound by SETDB1, the catalyst of H3K9me3 [132, 138, 139]. This potentially contradictory role of the repressive mark H3K9me3 at the 3'-ends of expressed genes has been described previously [140].

5.3.1 H3K27me3 *Peaks* Coincide with CpG Islands Bound by EZH2

For H3K27me3, regimeR called 42.4% (1.2Gb) of the HepG2 epigenome H3K27me3-enriched (1,221,850 1kb regions) and subdivided this into 940,753 *broad* (77%, 941Mb; μ_{ChIP}=12.03; θ_{F_1}=0.46) and 281,097 *peak* regions (23%, 281Mb; μ_{ChIP}=29.62; θ_{F_2}=0.68; Fig. 5.2A). H3K27me3-*peak* regions were characterized by a higher CpG odds ratio, *i.e.* CpG-content corrected for GC content, than, both *broad* or background regions (Fig. 5.2B). In conjunction with an elevated conservation (Fig. 5.3A) and a statistically significant overlap with annotated TSSs (Fisher's signed exact test; $P \leq 0.001$; odds ratio=1.98; Table 5.1A) these observations reaffirm that the TSSs targeted for *peak* H3K27me3 levels in HepG2 cells are CpG island promoters [129]. Moreover, H3K27me3-*peak* regions had a significantly higher level of EZH2 ChIP-seq enrichment (Wilcoxon signed-rank test; $P \leq 0.001$, Fig. 5.2C) which is the major H3K27 methyltransferase [125–128]. Together these observations suggest that H3K27me3-*broad* and -*peak* regions show distinct characteristics with respect to CpG content, localization and EZH2 enrichment.

5.3.2 H3K9me3 *Peaks* are Found within Repeats Bound by ZNF274

For H3K9me3, 14.7% (424Mb) of the HepG2 epigenome got classified into 202,390 *broad* (47.8%, 202Mb; μ_{ChIP}=11.27; θ_{F_1}=0.39) and 221,741 *peak* regions (52.2%, 222Mb; μ_{ChIP}=23.75; θ_{F_2}= 0.70; Fig. 5.2D). H3K9me3 covered ~3-fold less of the genome than H3K27me3, yet, with a higher fraction of *peak* regions. Both H3K9me3-*broad* and −*peak* regions showed a statistically significant overlap with repetitive DNA elements over background (Wilcoxon-signed-rank test; $P \leq 0.001$; Fig. 5.2E and Fig. 5.3B), which is a reported feature of constitutive heterochromatin marked by H3K9me3 [141]. H3K9me3-*peak* regions were significantly more enriched for ZNF274 than back-

A **H3K27me3**

		1.5kb Promoter			**Genes**		
		+	-	**Odds**, *P-value*	+	-	**Odds**, *P-value*
*b*road	enriched	51796	1464895	**0.637**	635970	880721	**0.3781**
	background	71750	1292603	*P<2.2e-16*	895445	468908	*P<2.2e-16*
*p*eak	enriched	21335	263211	**1.978**	144364	140182	**0.898**
	background	102211	2494287	*P<2.2e-16*	1387051	1209447	*P=2.932e-163*

B **H3K9me3**

		1.5kb Promoter			**Genes**		
		+	-	**Odds**, *P-value*	+	-	**Odds**, *P-value*
*b*road	enriched	17705	385127	**1.03**	189828	213004	**0.755**
	background	105841	2372371	*P=0.0003163*	1341587	1136625	*P<2.2e-16*
*p*eak	enriched	6488	227776	**0.6156**	83147	151117	**0.4553**
	background	117058	2529722	*P<2.2e-16*	1448268	1198512	*P<2.2e-16*

Table 5.1 – Transcriptional Start Site and Gene Overlap of H3K27me3- and H3K9me3-*Broad* and -*Peak* Regions in HepG2 Cells. (A) H3K27me3-*peak* regions are over-represented at transcriptional start sites (TSS). (B) Both H3K9me3 regimes are not over-represented at genic features. A "1.5kb Promoter" is defined as 750bp down- and upstream of the TSS; "+" = overlapping; "-" = non-overlapping; P-values are obtained from Fisher's exact test.

ground and H3K9me3-*broad* regions (Wilcoxon-signed-rank test, P≤0.001; Fig. 5.2F), in line with the theory that ZNF274 recruits the H3K9 methyltransferase SETDB1 [132]. Thus, the identified H3K9me3-*peak* regions may coincide with nucleation sites for heterochromatin assembly at repetitive elements.

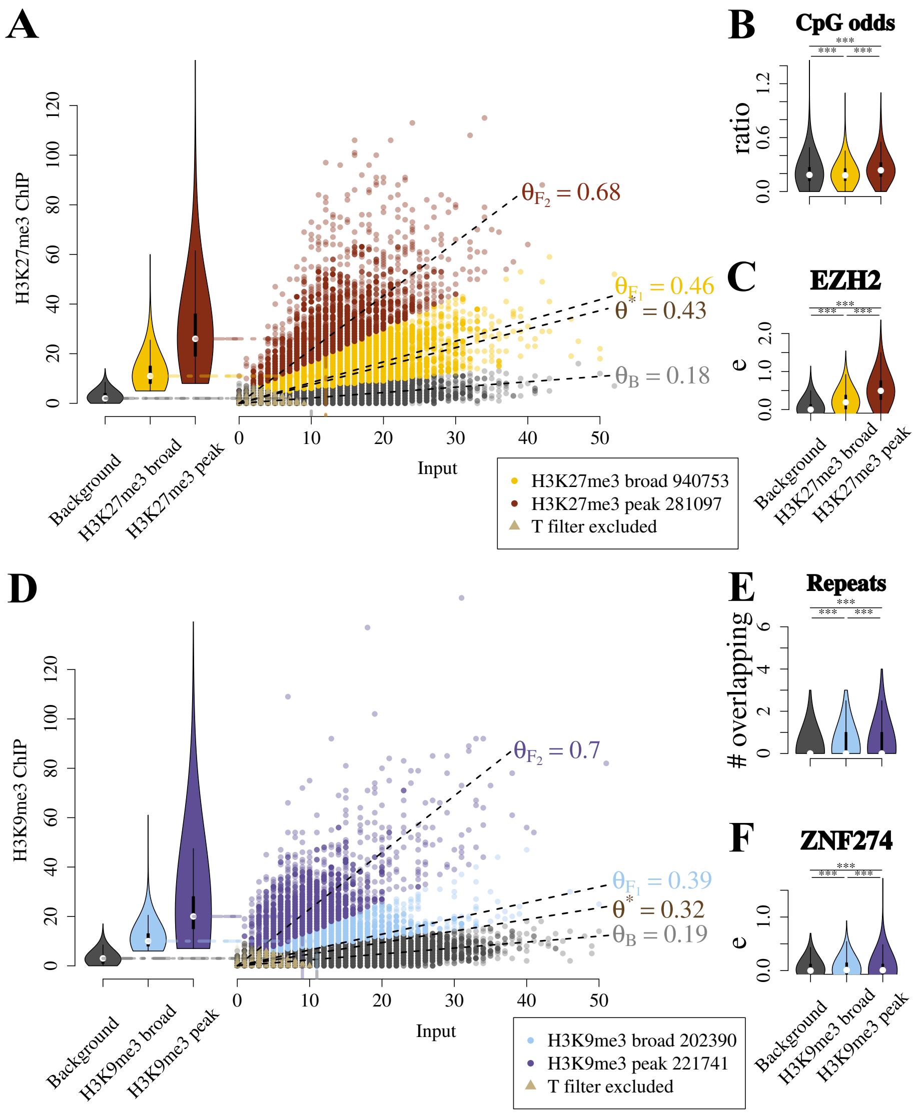

Fig. 5.2 – H3K27me3 and H3K9me3 Enrichment Regimes Coincide with Distinctive Sequence Features and Protein Binding. (A) regimeR identifies *broad* and *peak* H3K27me3 enrichment with distinctive levels of ChIP intensity (left). (B-C) H3K27me3-peaks have significantly greater CpG odds (B) and EZH2 ChIP-seq enrichR enrichment e (C) as compared to background and *broad* regions. (D) H3K9me3 regimes identified by regimeR (right) also have distinctive levels of ChIP signals (left). (E-F) H3K9me3-peaks are significantly enriched for repeats (E) and have a higher ZNF274 ChIP-seq enrichment e (F) as compared to both background and *broad* regions. Wilcoxon signed-rank Test: "***" (P≤0.001). For repeats see also Fig. 5.3.

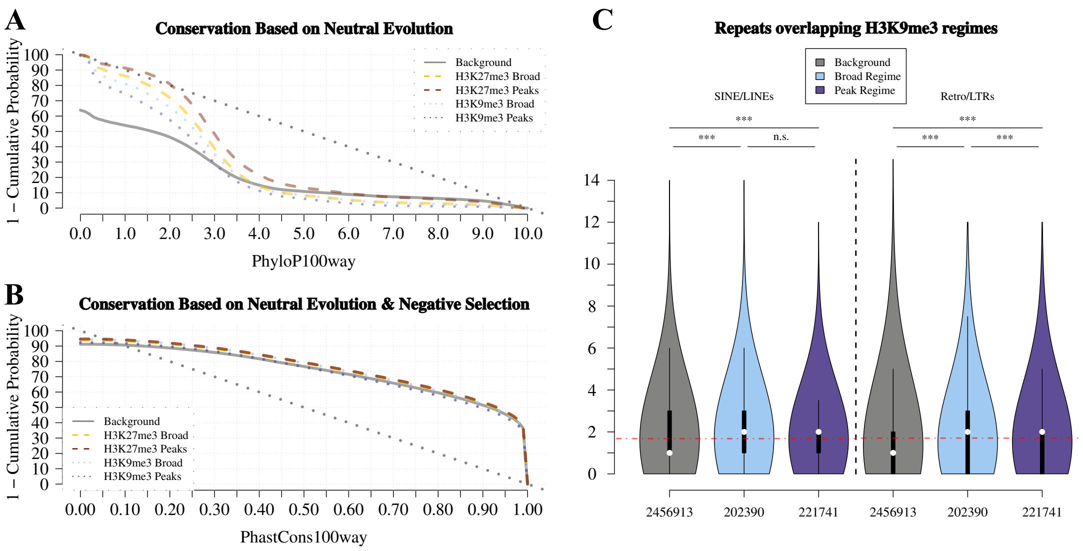

Fig. 5.3 – H3K27me3 Heterochromatin is Punctually More Conserved Than Background Regions and Enriched for Long Term Repeats (LTRs) in HepG2 Cells. (A-B) Complementary cumulative empirical density plots of regional PhyloP100way (A, punctual conservation) and PhastCons100way (B, broad conservation) scores reveals that H3K27me3 regimes are marginally more conserved than H3K9me3 regimes and background-regions in the vertebrate lineage. (C) The numbers of SINE/LINEs and Retro-elements/LTRs in H3K9me3 regimes are significantly greater than in background regions. Red dashed line represents genomic average of repeat overlap. "Retro/LTRs" = Retro elements and long term repeats; Wilcoxon signed-rank Test: "***" (P$\leq$0.001) and "n.s." (P>0.05).

5.3.3 Heterochromatic *Peaks* Resemble Nucleation Sites for Heterochromatin Embedded within Regions of *Broad* Enrichment

The observation that *peak* regions coincided with significantly higher levels of proteins associated to their catalysis than *broad*- and background-regions indicates that they may correspond to *putative* nucleation sites for heterochromatin assembly. In line with this idea I found the vast majority of H3K27me3-*peak* regions were embedded in H3K27me3 *broad* domains (82.8%). Also most H3K9me3-*peak* regions are either embedded in an H3K9me3-*broad* domain (43.4%) or at the border of a *broad* domain (35.1%). In addition, the DNA sequences of H3K27me3-*peak* and -*broad* regions are more conserved than background according to different measures of conservation (Fig. 5.3A, B) with the tendency of H3K27me3-*peaks* to be hyper-conserved. This finding is in line the theory that hyper-conserved domains underlie polycomb silencing [142]. On the contrary, H3K9me3 peaks were less conserved than *broad* regions and are predominantly found in repeat regions and retroelements (Fig. 5.3C) further supporting the aforementioned theory that repetitive elements recruit the H3K9 methyltransferase SETDB1 [141, 143]. Taken together, these findings strongly support the idea that the H3K27me3 and H3K9me3 peaks identified by regimeR are nucleation sites for facultative and consecutive heterochromatin, respectively.

5.3.4 H3K27me3 and H3K9me3 do Overlap by a Minority within and between Tissues

Next, the overlap of H3K27me3 and H3K9me3 regimes within and between the HepG2 and primary human hepatocytes (PHH) epigenome was analyzed. In ∼203Mb of the HepG2 epigenome H3K27me3 and H3K9me3 coincided (Fig. 5.4A) wherein *broad* regions tended to overlap more (∼80Mb; 8.5% of H3K27me3-*broad*, 39% of H3K9me3-*broad*) than *peak* regions (∼14Mb; 5.2% of H3K27me3-*peak*, 6.6% of H3K9me3-*peak*). Once more, the small overlap of H3K27me3 and H3K9me3 peaks supports the distinct nature of those two *putative* heterochromatic nucleation sites supporting the theory of their mutual exclusivity [144]. Surprisingly, PHH which are the tissue-of-origin of the HepG2 hepatocarcinoma cell line showed an overall increase in heterochromatin with the tendency towards more *peak* regions in H3K27me3 (∼613Mb) and more *broad* regions in H3K9me3 (∼900Mb; Fig. 5.4B). When compared, ∼41% (385Mb) of HepG2 H3K27me3-*broad*, ∼38.6% (108Mb) of HepG2 H3K27me3-*peak* and ∼49% (100Mb) of HepG2 H3K9me3-*broad* regions were coincident with regions of the same type in PHH (Fig. 5.4C). However, only 41Mb (18.4% of HepG2 H3K9me3-*peak*) shared a H3K9me3 peak. Taken together, heterochromatic regimes overlap only by a minority between HepG2 cells and PHH – especially H3K9me3-*peak* regions are diverse.

To investigate the diversity of H3K9me3 *peak* regions between the two studied cell types, I scanned the genome for occurrences of Krüppel-associated box zinc-finger transcription factor (KRAB-ZNF) binding sites (see Methods Section 5.2.3). A region on chromosome 19 confirmed the co-occurrence of potential KRAB-ZNF binding sites and H3K9me3-*peak* regions in both cell types (Fig. 5.5), especially for ZNF274 [132]. However, there existed ZNF274 motif occurrences

A **HepG2 cells**

		ZNF263			ZNF274		
		+	-	**Odds**, *P-value*	+	-	**Odds**, *P-value*
broad	enriched	1496	200894	**1.12**	78	202312	**1.33**
	background	17712	2660942	*P=0.00004*	777	2677877	*P=0.02*
peak	enriched	1049	220692	**0.69**	507	221234	**17.51**
	background	18159	2641144	*P<2.2e-16*	348	2658955	*P<2.2e-16*

B **Primary Human Hepatocytes**

		ZNF263			ZNF274		
		+	-	**Odds**, *P-value*	+	-	**Odds**, *P-value*
broad	enriched	5889	893880	**0.97**	68	899701	**0.19**
	background	13319	1967956	*P=0.09*	787	1980488	*P<2.2e-16*
peak	enriched	691	126177	**0.81**	679	126189	**84.16**
	background	18517	2735659	*P=1.9e-8*	176	2754000	*P<2.2e-16*

Table 5.2 – Overlap of ZNF263 and ZNF274 Motif Occurences with H3K9me3 *Broad* and *Peak* Regions in HepG2 Cells and Primary Human Hepatocytes. (A-B) H3K9me3-*peak* regions overlap significantly with ZNF274 motif occurrences for HepG2 cells (A; odds=17.51) and primary human hepatocytes (B; odds=84.16). "+" = overlapping; "-" = non-overlapping; P-values are obtained from Fisher's exact test.

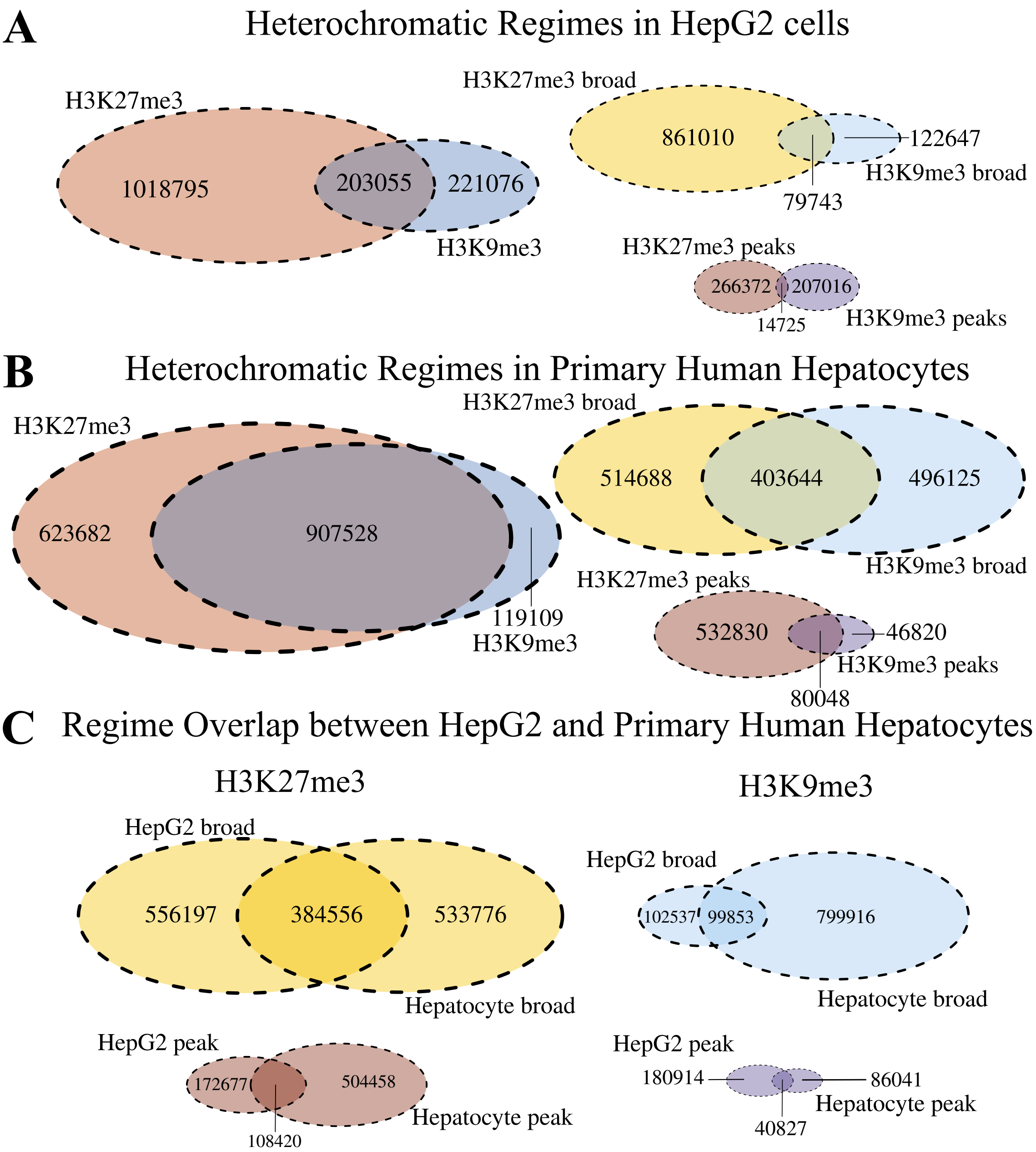

Fig. 5.4 – H3K27me3 and H3K9me3 Heterochromatic Regimes Predominantly Do Not Overlap within HepG2 Cells and are Predominantly Cell-Type Specific. (A) HepG2 H3K27me3 and H3K9me3 heterochromatin is shared for 203Mb but heterochromatic peaks overlap by a minority (∼6%). (B) PHH H3K27me3 and H3K9me3 heterochromatin is shared in 908Mb and 63% of H3K9me3 peaks overlap with a H3K27me3 peak. (B) On average ∼43% of HepG2 H3K27me3-*broad* and -*peak* and H3K9me3-*broad* regions overlap with same regimes in hepatocytes. Only ∼18% of HepG2 H3K9me3 peaks are also H3K9me3 peaks. Venn diagrams are scaled to the size of regions therein.

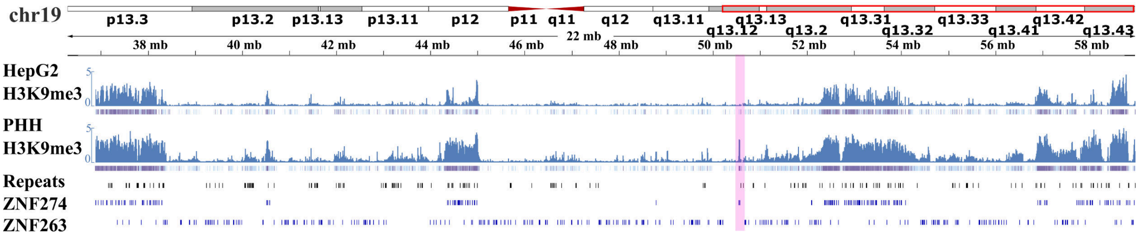

Fig. 5.5 – H3K9me3 Heterochromatin Correlates with KRAB-ZNF274 Motif Ocurrences. A genome browser shot of a 22Mb region on human chromosome 19 illustrates the strong overlap of constitutive heterochromatin *peak* regions with ZNF274 and ZNF263 motif occurrences (darkblue boxes) in HepG2 cells and primary human hepatocytes. However, also differences are apparent (pink overlay).

which were marked by heterochromatic peaks only in the primary tissue, *i.e.* PHH. Genome-wide, ZNF263 motifs are not over-represented in heterochromatin (Table 5.2) which supports the theory that ZNF263 has both repressing and activating roles [145]. On the other hand, ZNF274 motif occurrences are over-represented in H3K9me3-*peaks* of HepG2 cells (odds=17.51) and PHH (odds=84.16). Taken together, these observations highlight the differences in constitutive heterochromatin marked by H3K9me3 between cultured and primary cells.

5.4 Discussion

normR can be used to facilitate the discrimination of *peak*- and *broad*-regions against background in a single principled analysis, referred to as "regimeR". An analysis of H3K27me3 and H3K9me3 in HepG2 cells revealed that there exist distinct regions of *broad* (low) and *peak* (high) enrichment in HepG2 cells. For the first time, one principled approach was able to detect these distinct enrichment regimes in the low S/N ChIP-seq data. These findings suggest that those enrichment regimes are ultimately linked to the protein localization propensity in a heterogeneous sample and, furthermore, are indicative of the modes of action in the preservation of heterochromatin.

The histone modification H3K27me3 marks the facultative heterochromatin which can be dissected into two regimes of ChIP enrichment with regimeR. Specifically, conserved H3K27me3-*peak* regions were found within H3K27me3-*broad* domains at CpG-dense regions bound by EZH2, supporting the idea of CpG-enriched polycomb recruitment sites [142]. As opposed to canonical polycomb response elements in Drosophila [125, 128], a working model of polycomb recruitment for the establishment of facultative gene silencing is still lacking for mammals. The regimeR-based analysis of H3K27me3 ChIP-seq data constitutes an adequate approach to investigate on possible mechanisms of polycomb silencing in both Drosophila and mammals alike.

On the other hand, the constitutive heterochromatin is mutually exclusive to facultative chromatin which is predominantly marked by H3K9me3 [144]. Therein, regimeR can also identify two distinct regimes of enrichment. Similarly to H3K27me3-*peak* regions, the identified

H3K9me3-*peak* regions are embedded within or flanked by regions of *broad* enrichment. Those H3K9me3 peaks coincide with repetitive elements like retrotransposons which have recently been shown to provide gene regulatory potential [143]. H3K9me3-*peak* regions are bound by Krüppel-associated box zinc-finger transcription factor 274 (ZNF274) – a recruitment protein for the H3K9 methyltransferase SETDB1 [132]. I anticipate that regimeR can readily be used to interrogate on the stability of constitutive heterochromatin across conditions, *e.g.* in aging and environmental stress (see [141] for review).

My findings implicate a novel mode of action in the preservation of heterochromatin: High propensity heterochromatic regions with high (*peak*) ChIP-seq enrichment signal methyltranferase recruitment to establish the stably silenced heterochromatin. As a consequence of these recruitment signals, surrounding regions may be silenced with low propensity. In line with this idea, a low (*broad*) ChIP-seq enrichment of the heterochromatic modifications is observed around these nucleation sites. The herein identified preclusion of H3K27me3- and H3K9me3-*peak* regions hints at distinct modes of action of nucleation sites for facultative and constitutive heterochromatin, respectively. Especially, the low overlap of H3K9me3 *peak* regions between HepG2 cells and their cell type of origin, *i.e.* primary human hepatocytes, is an interesting subject for further investigation of the dynamics of constitutive heterochromatin in relation to environmental stimuli [141]. In particular, malignancies can induce substantial disruptions of the heterochromatin [139] and the immortalization of hepatocytes is an appropriate model to study such effects [146,147]. A systematic comparison of the differences in facultative heterochromatin between HepG2 cells and primary human hepatocytes is given in Chapter 6.

In the future, regimeR will prove useful in studies of heterogeneity in cellular epigenetic markings to identify regions of weak protein binding. For instance, regimeR has recently been used to model histone modification asymmetries [5]. Furthermore, I anticipate that the enhanced decomposition of ChIP-seq signals is beneficial to predict gene regulation based low affinity transcription factor binding [148, 149].

Chapter 6

diffR – Conditional Difference Calling in ChIP-seq Data

This chapters presents "diffR" – an implementation of the normR framework that enables for the direct comparison of NGS experiments. diffR fits a mixture model with three components representing background, control- and treatment-differential enrichment. The fitted background component enables the detection of conditional changes in read coverage by means of a two-sided statistical test. The usage of diffR is illustrated by a comparison of H3K27me3 and H3K4me3 ChIP-seq data between hepatocarcinoma cell line HepG2 and its cell type-of-origin, *i.e.* primary human hepatocytes (PHH). The diffR-based analysis revealed a disruption in the facultative heterochromatin of HepG2 cells and HepG2-specific H3K4me3 enrichment at promoters of transcription- and cell cycle-associated genes. A systematic comparison to one enrichR-based approach (see Chapter 4) and also three competitor methods for ChIP-seq difference calling [90–92] shows that diffR is very sensitive in the detection of genomic loci that change their association with the ChIP-seq target. I anticipate that diffR is well-suited to identify conditional differences in a variety of NGS-based approaches without a control experiment, *e.g.* DNaseI-seq and ATAC-seq.

6.1 Introduction

ChIP-seq enrichment calling is essential to identify genomic loci bound by a protein of interest. In Chapter 4 enrichR successfully achieved robust and sensitive ChIP-seq enrichment calling. While calling ChIP-seq enrichment over control is essential, another common task is the identification of differential ChIP-seq enrichment between two conditions. A conditional difference manifests itself in either **mutually exclusive** (*i.e.* condition-specific presence or absence of signal) or **quantitatively different** (*i.e.* differential intensity of signal) ChIP-seq enrichment. Similar to enrichment calling, the discrimination of signal against background is essential for the identification of regions that are differentially bound in two samples, *e.g.* healthy and diseased tissues. A difference caller should recover those regions, even if conditional control experiments are not available.

The binding of proteins to the DNA is naturally very dynamic between different conditions or cell-types (see Section 1.1.3). For instance, environmental stimuli like stress conditions affect the DNA binding of transcription factors like the glucocorticoid receptor to induce or repress certain transcriptional programs [150], *i.e.* transactivation/-repression. In addition, the patterns of chromatin modifications along the epigenome are greatly cell type-specific [44, 151] as manifested in the high correlation of their ChIP-seq signals to gene expression [34] and mutational processes [152]. An accurate quantification of these differential protein bindings/epigenetic alterations is needed to systematically compare samples on a molecular level. Critically, an even higher resolution can be achieved by also recovering genomic loci wherein a change in the strength of protein binding is observed.

Some approaches, *e.g.* MACS2 [74], aim to identify condition-specific exclusive enrichment, whereas other recently developed methods [90–92] also allow for the stratification of quantitatively different enrichment to the identification of differential ChIP-seq enrichment. The latter, more recent tools, employ a three-state Hidden Markov Model to identify, in addition to mutually exclusive enrichment, also condition-specific changes of signal within regions of concurrent ChIP enrichment. To this extent, a computationally intensive training is needed to learn a hidden state representation of the ChIP-seq data. This concept leads to the putatively meaningful interpolation of the ChIP-seq signal based on the ChIP-seq read coverage in adjacent genomic loci. However, this "smoothing" may abstract the original signal substantially and, in consequence, can lead to spurious results. To my knowledge, a methodical comparison of these competing methods is still lacking in the literature.

Here, I present an implementation of the normR framework (see Chapter 3) referred to as **"diffR"** to simultaneously call mutually exclusive and quantitatively different enrichment in two ChIP-seq samples. diffR employs a three-component mixture model derived from the normR framework to perform the joint normalization and identification of differential enrichment. As a proof of concept, I use diffR to call differences between PHH and the hepatocarcinoma HepG2 cell line in ChIP-seq data for the heterochromatic H3K27me3 and euchromatic H3K4me3 histone modifications. The analysis recovers many epigenetic alterations between the two cell types close to genes whose dis-regulation is known to be a hallmark of the immortalization of cancer cells, *e.g.* E2F2 [153, 154]. A methodical comparison based on a trustful validation set reveals a superior performance of diffR over previously developed approaches. In addition, diffR is shown to precisely recover annotated genomic amplifications in the HepG2 cell line from the Input tracks in PHH and HepG2 cells. Once more, the versatility of the normR framework is illustrated to open up new possibilities in the analysis of NGS read count data with the normR framework.

6.2 Methods

Firstly, this section provides details on the processing and the quality of the sequencing data used. Secondly, the normR framework is adapted to facilitate the tasks of ChIP-seq difference calling between two conditions which is referred to as "diffR". Next, diffR is used to call differential H3K4me3 and H3K27me3 ChIP-seq enrichment between PHH and HepG2 cells. Thirdly, regions that are called differential by diffR are overlapped with transcription start sites and genes which are then subjected to Gene Ontology over-representation analysis for functional assessment. Finally, diffR difference calls are validated by the comparison (i) to mutually exclusive enrichment calls obtained from two individual enrichR calls, referred to as "enrichR-compare", and (ii) to results of other tools for calling differential ChIP-seq enrichment [90–92] via the previously introduced consensus-vote strategy to obtain a trustful validation set (see Section 4.2.4).

6.2.1 Data Sets

ChIP-seq Data. Paired end reads from Input, H3K4me3 and H3K27me3 ChIP-seq for HepG2 cells and PHH were processed and quantified as described in Section 4.2.1. Peaks for HepG2 Polymerase II and CTCF were downloaded from UCSC under accessions wgEncodeEH001792 and wgEncodeEH000080, respectively.

HepG2 Genotyping. HepG2 genotype information for hg19 was generated by ENCODE/HudsonAlpha (Gene Expression Omnibus [121] under accession: `GSM999286`). A bed file containing annotated amplifications and deletions was downloaded from the UCSC ENCODE DCC repository (`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibGenotype/wgEncodeHaibGenotypeHepg2RegionsRep1.bedLogR.gz`).

6.2.2 The `normR` Methods: `diffR`

The normR framework (see Chapter 3) was adapted to call differential ChIP-seq enrichment between two conditions, referred to as **"diffR"**. Essentially, diffR uses $m = 3$ mixture components, *i.e.* background B and two foreground components F_1 (*control* enriched) and F_2 (*treatment* enriched). The number of free parameters is now increased to five, similar to regimeR (see Section 5.2.2). Given the normR model given by Equation (3.2) and the normR likelihood function defined by Equation (3.3) the following "diffR" likelihood function can be derived:

$$\mathcal{L} = P(\pi, \theta | r_i, s_i) = \prod_i \binom{s_i + r_i}{s_i} \sum_{k \in \{B, F_1, F_2\}} \pi_k \cdot \theta_k^{s_i} \cdot (1 - \theta_k)^{r_i}$$

where r_i (s_i) corresponds to the number of reads in the control (treatment) ChIP-seq for non-overlapping, fixed size bins $i = 1, \ldots, n$. Again, parameters are fit by running the EM algorithm as described in Section 4.2.2. In contrast to enrichR and regimeR, the mixture component k with θ_k closest to $\theta^* = \frac{\sum s_i}{\sum s_i + r_i}$ is used as background B for a two-sided binomial test. To achieve a robust multiple testing correction, the T method (Section 2.1.2) uses the maximal T threshold obtained from the P-values of the two-sided binomial test for either (r, s) or the label-switched (s, r). Next, P-values are transformed to q-values [56].

To identify regions with significant differences between conditions, regions with a q-value $\leq\alpha$ (user-chosen significance level) are recovered. Next, every *significant* bin j is assigned to one of the two conditions by *Maximum A Posteriori*

$$z_j = \underset{k \in \{F_1, F_2\}}{\arg\max} \left(\frac{P(\pi_k, \theta_k | s_j, r_j)}{P(\pi, \theta | s_j, r_j)} \right),$$

where z_j indicates the assignment of a significantly enriched region j to either F_1 (control) or F_2 (treatment).

The regularized differential ChIP-seq signal e^* is computed as described in Section. 4.2.2. In diffR, two average enrichment factors are estimated, namely the average enrichment factor $\langle f_1 \rangle$ for control

$$\langle f_1 \rangle = \frac{\theta_{F_1}}{1 - \theta_{F_1}} \cdot \frac{1 - \theta_B}{\theta_B}$$

and the average enrichment factor $\langle f_2 \rangle$ for treatment

$$\langle f_2 \rangle = \frac{\theta_{F_2}}{2 - \theta_{F_2}} \cdot \frac{2 - \theta_B}{\theta_B}.$$

Dependent on the algebraic sign of e_i^* the regularized and normalized enrichment e_i is

$$e_i = \begin{cases} e_i^* \cdot (\log\langle f_1 \rangle)^{-1}, & e_i^* \leq 0 \\ e_i^* \cdot (\log\langle f_1 \rangle)^{-1}, & e_i^* > 0 \end{cases}.$$

These routines were implemented in the normR R package [4] as the diffR()-function (see R code snippet below).

Read counting was configured as previously described in Section 4.2.2. Read counts obtained from H3K4me3 (H3K27me3) ChIP-seq experiments in PHH and HepG2 cells were modeled with the diffR()-function in 500bp (1,000bp) bins:

```
diffs <- diffR(
  treatment = "ChIP1.bam",
  control = "ChIP2.bam",
  genome = genome,
  countConfig = countConfig,
  procs = 24
)
```

Bins with q-value≤0.05 (0.1) for H3K4me3 (H3K27me3) were called differentially enriched and assigned to control or treatment by *Maximum A Posteriori* internally in `diffR`. Results were exported to bed and bigWig using normR's `exportR`()-function:

```
for (filetype in c("bed", "bigWig")) {
  exportR(
    x = diffs,
    filename = paste0("differences.", filetype),
    type = filetype
    fdr = 0.05 #0.1
  )
}
```

6.2.3 Gene Ontology Analysis

Transcription start site (TSS) and gene annotations were compiled as previously described (see Section 4.2.1). After overlapping differentially enriched regions with genic features, topGO [155] was used on the gene ontology "Biological Process" (BP) with algorithms "classic" (algorithm="classic") and "elim" (algorithm="elim") for statistics "fisher" (statistic="fisher") and "ks" (statistic="ks") for GENCODE gene IDs mapped to Ensembl gene IDs. The "ks" statistic allows for supplying a score for each entity. Here, the diffR calculated "q-value" was used as score. I retained only the top 1,000 (n=1000) GO terms which were ordered by "elim" algorithm and ranked by "classic" algorithm calculated P-values:

```
require(topGO)
#get GO annotated Ensembl Genes
go2ensembl <- annFUN.org(ontology, mapping="org.Hs.eg.db", ID="ensembl")
#get GENCODE genes and filter these for the ones in gene universe
gencode <- loadDb("data/gencode.v19.annotation.transcriptDb.sqlite")
gene.universe <- intersect(
  unique(GenomicFeatures::genes(gencode)[["genes"]]),
  unique(unlist(go2ensembl))
)
#set diffR pvalue as score for differentially modified TSSs
idx <- gene.universe %in% diffTSSs
```

```
allGenes <- 1-as.integer(idx)
names(allGenes) <- gene.universe
allGenes[idx] <- pvals[diffTSSs %in% gene.universe]
goData <- new("topGOdata",
  description="diffR differential TSS histone marking study (scored)",
  ontology="BP",
  allGenes=allGenes, geneSel=function(p) { return(p <= 0.05) },
  annot=annFUN.GO2genes, GO2genes=go2ensembl, #GO mapping for ensembl IDs
  nodeSize=10
)
#testing
resultFisher <- runTest(goData, algorithm="classic", statistic="fisher")
resultKS <- runTest(goData, algorithm="classic", statistic="ks")
resultKS.elim <- runTest(goData, algorithm="elim", statistic="ks")
#compile results
resDf <- GenTable(goData,
  classicFisher = resultFisher,
  classicKS = resultKS,
  elimKS = resultKS.elim,
  orderBy ="elimKS",
  ranksOf = "classicFisher",
  topNodes=1000
)
```

6.2.4 Comparison of ChIP-seq Difference Callers

The performance of diffR was evaluated on two levels: (i) via a comparison to mutually exclusive enrichment calls obtained from an approach referred to as **"enrichR-compare"**; and (ii) via a comparison to results from three competitor methods [90–92] based on a trustful validation set generated by overlapping competitor methods calls.

Mutually Exclusive Enrichment with "enrichR-compare"

To evaluate diffR results, enrichR-compare represents an alternative approach to detect *mutually exclusive enrichment*. Firstly, enriched regions in HepG2 (PHH) conditions were called with enrichR on HepG2 (PHH) ChIP-seq over HepG2 (PHH) Input for H3K4me3 (500bp bins) and H3K27me3 (1,000bp). For a fair comparison to diffR, I considered only significant regions with a posterior of $\geq$0.50. Secondly, I called a bin *"both enriched"* or *"HepG2-exclusive"* (*"PHH-exclusive"*) if it was enriched in both conditions or exclusively in HepG2 (PHH), respectively. Finally, I determined accuracy of the diffR classification by comparing it to the enrichR-compare classification.

Enrichment Calling in Third-Party Tools

Similarly to the approach described in Section 4.2.4, I downloaded a set of competitor methods to evaluate their performance in relation to diffR. I called conditional differences in ChIP enrichment with ChIPDiff [90], histoneHMM [91] (v1.6) and ODIN [92] (v0.4) in the following way:

Firstly, duplicated fragments were removed, only keeping reads with a mapping quality higher than 20 by using samtools [119] (`v0.1.19-44428cd`) to allow for a fair comparison with diffR:

```
samtools view -F 1024 -q 20 in.bam > out.bam
```

Secondly, conditional differences for H3K4me3 and H3K27me3 between PHH and HepG2 cells were called. ODIN was run with the following command incorporating condition specific Input alignments and the `hs37d5` genome sequence:

```
rgt-ODIN -m -v --input-1=Input1.bam --input-2=Input2.bam ChIP1.bam ChIP2.bam \
  hs37d5.fa hs37d5_chromSizes
```

For histoneHMM, enriched regions were called prior to differential enrichment detection as suggested in the tutorial (`http://histonehmm.molgen.mpg.de/v1.6/histoneHMM.pdf`):

```
./histoneHMM_call_regions.R -b 500 -c hs37d5_chromSizes_Autosomes \
  -o ChIP1_histoneHMM -t 20 ChIP1.bam
./histoneHMM_call_regions.R -b 500 -c hs37d5_chromSizes_Autosomes \
  -o ChIP2_histoneHMM -t 20 ChIP2.bam
./histoneHMM_call_differential.R --sample1 ChIP1.bam --sample2 ChIP2.bam \
  ChIP1.txt ChIP2.txt
```

For ChIPDiff which works on single end read alignments only, I considered first reads only in a properly mapped pair (`-f 66`) for a fair comparison to difference callers that work with paired end data:

```
samtools view -b -f 66 ChIP.bam > ChIP_SE.bam
bedtools bamtobed -I ChIP_SE.bam > ChIP_SE.bed
./ChIPDiff ChIP1_SE.bed ChIP2_SE.bed hs37d5_chromSizes
```

Thirdly, to compare called peaks using above methods to diffR differential regions, overlaps of peaks with 500bp (1,000bp) bins were calculated in R for H3K4me3 (H3K27me3) if a differential region at FDR 5% (10%) overlapped a window by at least 250bp. As opposed to enrichR, the matrix now contains two columns for each tool because a region can be differential in either control or treatment:

```
binsize = 500; fdr = 0.05;
```

```
gr <- tileGenome(genome.gr, width = binsize)
ov <- matrix(0, nrow = length(gr), ncol = 6)
colnames(ov) <- c(
  "diffR_ctrl",
  "diffR_treat",
  "ODIN_ctrl",
  "ODIN_treat",
  "ChIPDiff_ctrl",
  "ChIPDiff_treat",
  "histoneHMM_ctrl",
  "histoneHMM_treat"
)
for (method in colnames(ov)) {
    peaks.sig <- peaks[[meth]][which(peaks[[meth]][["lqval"]] >= -log10(fdr))]
    ov[,method][countOverlaps(gr, peaks.sig, minoverlap = 250)> 0 )] <- 1
}
```

Fourthly, a “tool-condition-specific *bona-fide* benchmark” for the comparison of all tools was defined as follows: A bin is differentially enriched for one condition under the gold standard if at least two out of three other methods (including diffR) called this bin differentially enriched for this condition. The following R code was used:

```
gs <- lapply(colnames(ov), function(method) {
    col.idx <- which(colnames(ov) != method &
      grep(strsplit(method, "_")[[1]][2], colnames(ov))
    which(apply(ov[,col.idx], 1, sum) >= 2)
})
names(gs) <- colnames(ov)
```

Finally, precision and recall were computed under these tool-condition-specific *bona-fide* benchmark sets for all peaks reported by a tool:

```
getPrecRecall <- function(ov, gs) {
  mp <- which(ov == 1)
  tp <- sum(mp %in% gs)
  fn <- sum(!(gs %in% mp))
  fp <- sum(!(mp %in% gs))tn <- dim(ov)[1] - tp - fn - fp
  specificity <- tn/(fp+tn)
  precision <- tp/length(mp)
  recall <- tp/(tp + fn)
  f2 <- fscore(precision, recall, 2)
  return(c(
    "precision"=precision,
    "recall"=recall,
    "f2"=f2,
    "specificity"=specificity
  ))
}
stats <- mapply(getPrecRecall, as.list(ov), gs)
```

6.3 Results

An important task in the analysis of ChIP-seq data is represented by the identification of epigenetic alterations between conditions. The normR framework can address this problem by calling differential ChIP-seq enrichment between two conditions with a three-component mixture model, referred to as **"diffR"**. The diffR approach fits a robust background model which allows for a reliable identification of conditional differences by means of a two-sided binomial test. As a proof of principle, diffR was applied to H3K27me3 (low S/N) and H3K4me3 (high S/N) ChIP-seq data from PHH and the hepatocarcinoma cell line HepG2. The results of diffR were systematically compared to those of four competitor approaches in two ways: (i) those obtained by calling mutually exclusive enrichment with enrichR (see Chapter 4) on the two conditions separately, referred to as "enrichR-compare"; and (ii) those obtained by three previously developed methods, namely ChIPDiff [90], histoneHMM [91] and ODIN [92]. Furthermore, diffR was used to identify HepG2-associated Copy Number Variations (CNVs) by calling conditional differences on the Input sequencing experiments of PHH and HepG2.

6.3.1 Difference Calling in HepG2 Cells and Primary Human Hepatocytes

Visual inspection of a 50kb region on chromosome 19 confirmed that, therein, H3K27me3 heterochromatic domains were mostly exclusive to HepG2 cells and that most H3K4me3 peaks were common despite detectable differences in signal intensity (Fig. 6.1). The majority of the cell type-exclusive enrichment of H3K27me3 was called by most methods, *e.g.* the HepG2-specific heterochromatin downstream of E2F Transcription Factor 2 (E2F2). The fidelity of the borders of this mutually exclusive enrichment varied strongly between the methods. For H3K4me3, the quantitative differences between HepG2 cells and PHH were recovered mainly by diffR, ChIPDiff and ODIN. Strikingly, the detected HepG2-specific gain of H3K4me3 signal at the E2F2 promoter coincided with an increased E2F2 gene expression as revealed by RNA-seq – an observation that underpinned the positive correlation of the promoter's H3K4me3 ChIP-seq signal and its

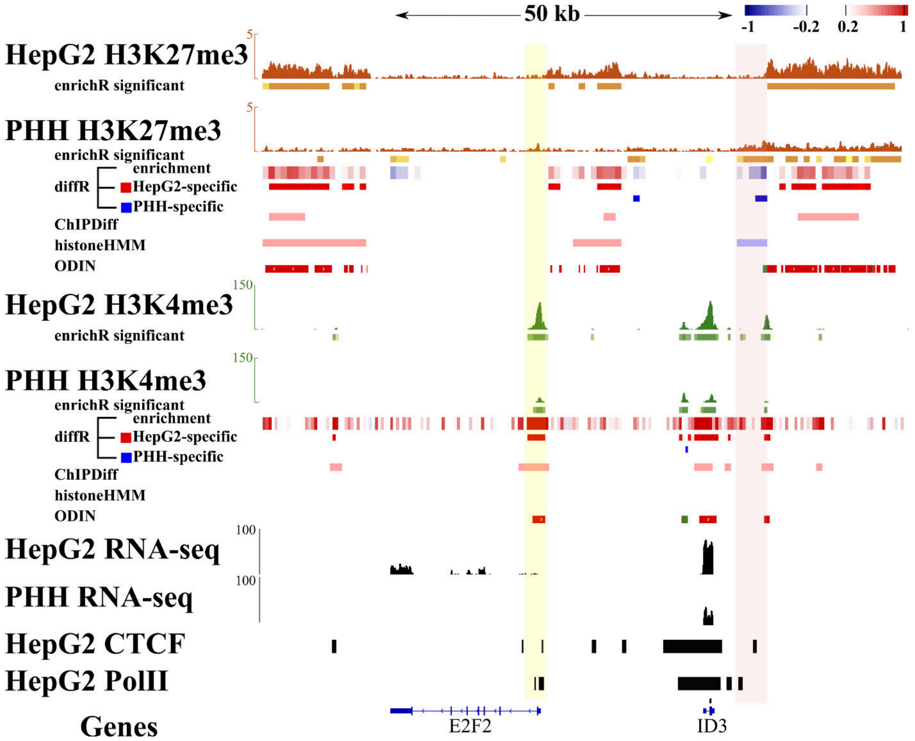

Fig. 6.1 – diffR Difference Calling Uncovers Mutually Exclusive Enrichment and Quantitive Differences in H3K27me3 and H3K4me3 ChIP-seq Data alike at the E2F2 Locus. Prim ary Human Hepatocytes (PHH) and HepG2 Input-seq (grey), H3K27me3 (orange) and H3K4me3 (green) ChIP-seq with RNA-seq (black) coverage around E2F Transcription Factor 2 promoter (E2F2) locus on human chromosome 19. A region ∼40kb upstream of E2F2 (pink overlay) shows a PHH-exclusive enrichment for H3K27me3 that is recovered by diffR, histoneHMM and enrichR-compare. The PHH-exclusive change in heterochromatin is accompanied by a quantitative difference in H3K4me3 which is detected by diffR, ChIPDiff and ODIN as well as peaks for CTCF and polymerase 2 in HepG2 cells. The E2F2 promoter (yellow overlay) is detected HepG2-differential for H3K4me3 by diffR, ChIPDiff and ODIN which is also supported by an increased RNA-seq coverage along the E2F2 gene body and a polymerase 2 peak. Calls of differential enrichment are displayed as red (HepG2 conditional) or blue (PHH conditional) boxes for diffR, ChIPDiff, histoneHMM and ODIN. enrichR enriched regions displayed as boxes below respective tracks.

activity [34]. The HepG2-specific expression of this crucial regulator of the cell cycle [147] indicated that the uncovered difference in H3K4me3 at the E2F2 promoter is in fact genuine and that its induction may reflect the increased proliferation potential in HepG2 cells in comparison to PHH [153, 154]. Downstream of the E2F2 locus, diffR and histoneHMM identified a PHH-specific H3K27me3 domain which was accompanied by an emerging H3K4me3 peak in HepG2 cells which was only detected by diffR, ChIPDiff and ODIN. It can be speculated that the E2F2 induction in hepatocarcinoma cells is related to the HepG2-specific activation of an enhancer at that locus – an idea that is supported by the reported presence of RNA polymerase 2 and CTCF

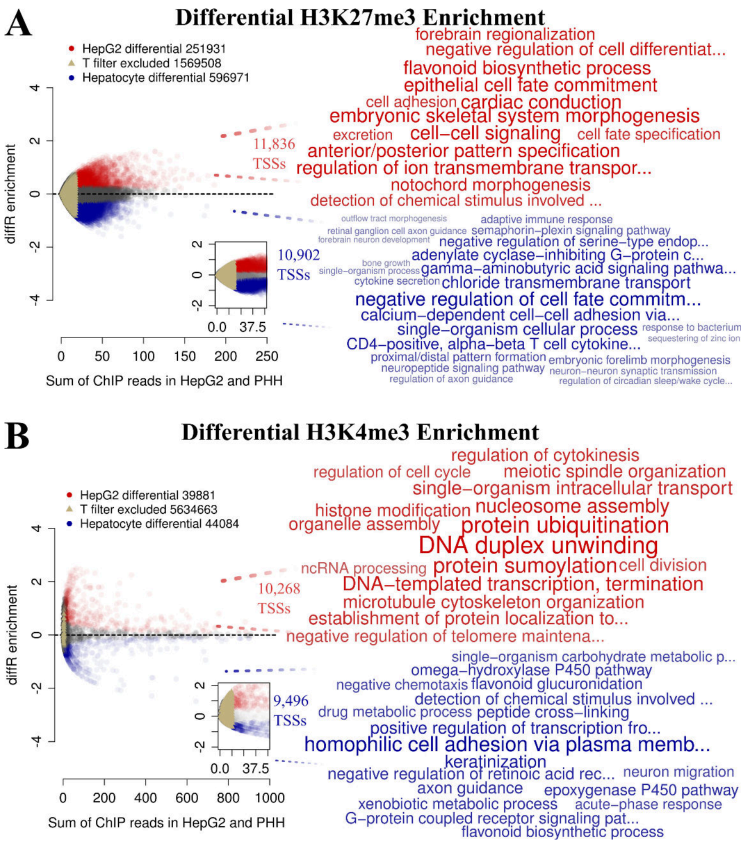

Fig. 6.2 – diffR Detects Functional Epigenetic Alterations between HepG2 Cells and Primary Human Hepatocytes. (A-B) diffR fold enrichment plotted against sum of H3K27me3 (A) and H3K4me3 (B) ChIP-seq read counts in HepG2 cells and primary human hepatocytes (PHH). diffR detects differential enrichment for HepG2 cells (red) and PHH (blue) in low and high count regions even for low diffR differential enrichment values. T method is stringently filtering low count regions because of diffR's two-sided binomial test (buff triangles, see also main text). (A) A wordcloud (right) depicts how HepG2-differential H3K27me3 regions (red) overlap/repress 11,836 TSSs that drive genes in morphogenesis and signaling and how PHH-differential H3K27me3 regions (blue) overlap / repress 10,902 TSSs that drive genes in cell fate commitment and adhesion. (B) A wordcloud (right) depicts how HepG2-differential H3K4me3 regions (red) overlap 10,268 active TSSs that drive genes in transcription and cell cycle and how PHH-differential H3K4me3 regions (blue) overlap 9,496 active TSSs that drive genes in keratinization and the P450 pathway. "TSSs"=Transcriptional Start Sites. Wordclouds represent significantly enriched ($P \leq 0.05$) Gene Ontology terms with their fontsize based on "$-\log_{10}$ q-value"of the hypergeometric test.

in that region for HepG2 cells ([43], Fig. 6.1). In summary, diffR detected mutually exclusive and quantitatively differential enrichment in low S/N H3K27me3 and high S/N H3K4me3 ChIP-seq data and these calls were also detected by the competitor methods at different accuracy levels.

Genome-wide, diffR reported in total 848,902 1kb regions (849Mb) as differentially H3K27-me3-enriched (Fig. 6.2A). Out of these 251,931 regions (252Mb) were HepG2-differential and repressed 11,836 TSSs of genes regulating morphogenesis and cell-cell signaling. 596,971 PHH-differential regions (597Mb) repressed 10,902 TSSs of genes functioning in cell fate commitment and T-cell development which represent pathways not functioning in liver cells. Together, this suggests an overall decrease in heterochromatin in HepG2 cells in comparison to PHH. For H3K4me3, diffR recovered 83,965 500bp regions (42Mb) as being differentially enriched between HepG2 and PHH (Fig. 6.2B). In contrast to the prevalence of H3K27me3 in PHH, H3K4me3-differential regions were similar in numbers between the two cell types. 39,881 regions (20Mb) were HepG2-differential and overlapped 10,268 TSSs that drove genes mainly related to the transcription and cell cycle. 44,084 PHH-differential H3K4me3 regions (22Mb) overlapped 9,496 TSSs of genes that were associated with liver function (*e.g.* P450 pathway) and tissue characteristics (*e.g.* keratinization or cell adhesion) that are absent in a cell line like HepG2. Taken together, diffR uncovered many hetero- and euchromatic alterations between HepG2 cells and PHH around genes that regulate diverse functions related to the biology of these cell types.

6.3.2 Comparison of ChIP-seq Difference Callers

A systematic assessment of the valdity of diffR results was achieved on two levels (see Methods Section 6.2.4):

(a) diffR results were compared to another normR approach that calls conditional differences by calling individual ChIP-seq enrichment over Input for each condition and then identifies mutually exclusive enrichment by overlapping enriched regions of samples, referred to as **"enrichR-compare"**.

(b) Three competitor tools [90–92] were used to call conditional differences. A trustful validation set for each method based on a consensus vote among the remaining tools ("tool-condition-specific *bona-fide* benchmark", Methods 6.2.4). The *bona-fide* benchmark was used to assess every method for its enrichment classification accuracy.

(a) enrichR-compare

On the first level, enrichR-compare was applied to call enrichment in ChIP-seq over Input for HepG2 cells and PHH individually to yield the following classification (Methods 6.2.4):

"No enrichment": enrichR did not detect enrichment in neither HepG2 cells nor PHH.

"PHH-exclusive": enrichR detected enrichment in PHH but not in HepG2 cells.

"HepG2-exclusive": enrichR detected enrichment in HepG2 cells but not in PHH.

"both enriched": enrichR detected enrichment in, both, HepG2 cells and PHH.

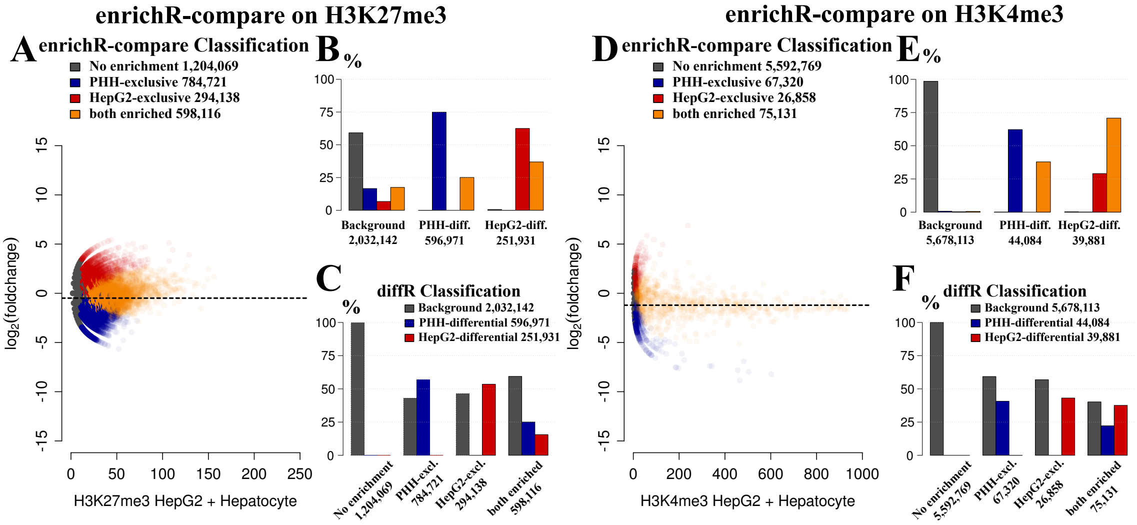

Fig. 6.3 – diffR Recovers Mutually Exclusive Enrichment that is Detected by enrichR-compare. (A,D) $\log_2$ H3K27me3 (A) / H3K4me3 (D) fold changes between HepG2 cells and primary human hepatocytes (PHH) plotted against the sum of ChIP-seq counts. enrichR-compare classifies regions into “both enriched” (small fold change, high ChIP-seq counts) and “no enrichment” (low ChIP-seq counts) as well as “HepG2-exclusive” and “PHH-exclusive” concordant with high absolute fold changes and elevated ChIP-seq count levels that pass the T filter. (B,E) diffR recovered “HepG2-” and “PHH-differential” H3K27me3 (B) / H3K4me3 (E) regions that are majorly classified as “HepG2-” and “PHH-exlusive”, respectively. (C,F) diffR detects H3K27me3 (C) / H3K4me3 (F) differential enrichment in “both enriched” regions but misses some cell type-exclusive regions.

This enrichR-compare classification was used to benchmark results obtained from diffR.

Genome-wide, for H3K27me3, enrichR-compare revealed that many enriched regions were common in HepG2 and PHH (“both enriched”: 598,116; 598Mb) and 294,138 (294Mb) were “HepG2-exclusive” (Fig. 6.3A). Similarly to diffR, enrichR-compare detected many “PHH-exclusive” regions (784,721; 784Mb) for H3K27me3 suggesting once more a disruption of the hepatocyte heterochromatin in hepatocarcinoma cells. When compared to enrichR-compare, the majority of differential regions detected by diffR were classified as cell type-exclusive (*i.e.* “HepG2-exclusive” and “PHH-exclusive”) by enrichR-compare (Fig. 6.3B). This result reaffirmed that diffR is precise in detecting mutually exclusive enrichment. A substantial fraction of diffR differential regions are classified as “both enriched” by enrichR-compare – representing most likely regions of quantitatively different ChIP-seq signal intensity between PHH and HepG2 cells (Fig. 6.3B) which, in consequence, lead to a reduced sensitivity (Table 6.1). Sensitivity was also reduced because 44% of the H3K27me3 cell type-exclusive regions were not called by diffR (Fig. 6.3C, see also below).

For H3K4me3, the enrichR-compare analysis revealed that most enriched 500bp regions were “both enriched” (75,131; 38Mb) while 26,858 (14Mb) were “HepG2-” and 67,320 (34Mb) “PHH-exclusive” (Fig. 6.3D). Similarly to H3K27me3, diffR detected, both, the cell type-exclusive enriched regions as well as the quantitative differences in the level of enrichment in “both enriched” regions (Fig. 6.3E). Again, a reduced sensitivity of diffR was observed – also because diffR did not recover all cell type-exclusive regions detected by enrichR-compare (Fig. 6.3F). This first assessment revealed a disagreement of diffR with enrichR-compare to recover some cell type-exclusive regions between HepG2 cells and PHH. Though, the comparison could confirm that diffR is very precise in terms of detecting mutually exclusive and also quantitatively different enrichment which could not be detected by enrichR-compare.

Next, two properties of diffR were studied which could lead to a discrepancy in sensitivity when compared to enrichR-compare: (i) diffR’s detection of quantitative differences in enrichR-compare’s “both enriched” regions; and (ii) diffR’s shortcoming to to classify some enrichR-

A H3K27me3

		enrichR-compare	**diffR Classification Performance**			
		1kb bins called	True Positives	False Positives	False Negatives	Recall
Unfiltered	HepG2	294138	157397	94534	136741	0.535
	PHH	784721	447176	149795	337545	0.570
	combined	1078859	604573	244329	474286	0.553
No Low Counts	HepG2	174088	147236	94027	26852	**0.846**
	PHH	478152	415990	149783	62162	**0.870**
	combined	652240	563226	243810	89014	**0.858**
No diffR CNVs	HepG2	187527	81573	33531	105954	0.435
	PHH	582610	342493	122745	240117	0.588
	combined	770137	424066	156276	346071	0.512
No ENCODE CNVs	HepG2	214932	105316	57689	109616	0.490
	PHH	605590	353522	123113	252068	0.584
	combined	820522	458838	180802	361684	0.537

B H3K4me3

		enrichR-compare	**diffR Classification Performance**			
		1kb bins called	True Positives	False Positives	False Negatives	Recall
Unfiltered	HepG2	26858	11577	28304	15281	0.431
	PHH	67320	27400	16684	39920	0.407
	combined	94178	38977	44988	55201	0.419
No Low Counts	HepG2	10681	10362	28283	319	**0.970**
	PHH	36518	26601	16684	9917	0.728
	combined	47199	36963	44967	10236	**0.849**
No diffR CNVs	HepG2	13574	5268	12646	8306	0.388
	PHH	46540	19437	12814	27103	0.418
	combined	60114	24705	25460	35409	0.403
No ENCODE CNVs	HepG2	20309	8590	19738	11719	0.423
	PHH	52332	21523	13960	30809	0.411
	combined	72641	30113	33698	42528	0.417

Table 6.1 – Consistency between diffR and enrichR-compare can be Increased by Removing Low Power and CNV Regions. (A-B) Performance of diffR with respect to enrichR-compare calls as ground truth on differential H3K27me3 (A) or H3K4me3 (B) enrichment calls in HepG2 cells and Primary Human Hepatocytes (PHH) for different filters. Filtering of low power regions reduced the number of false negatives and achieved a boost in the consistency as measured by recall (sensitivity). Filtering out *in silico* or *in vitro* determined CNVs improved the recall only marginally. Recall values greater than 0.75 set in bold font.

compare “cell type-exclusive” regions. Firstly, I studied the impact of detecting quantitative differences. The number of diffR’s false positives is influenced by its ability to detect conditional differences in ChIP-seq signals in regions that are classified as “both enriched” by enrichR-compare. As a measure for *true* quantitative differences between two ChIP-seq experiments, I determined the conditional $\log_2$ fold changes between HepG2 cells and PHH. The fold changes in diffR’s false positive regions were equal (H3K27me3) or greater (H3K4me3) than those in cell type-exclusive regions (Fig. 6.4A). This observation suggested that those cell type-differential regions detected by diffR, in fact, correspond to regions of differential signal intensity in the two conditions.

Secondly, to study diffR’s shortcoming to classify some cell type-exclusive regions, I investigated the test framework that is used in both methods. By design, diffR uses a two-sided binomial test to recover regions significantly different from the fitted background model (see Methods Section 6.2.2) whereas enrichR-compare depends on enrichR which uses a one-sided test (Methods Section 4.2.2). Both methods use the T method described in section 2.1.2 to filter out low power (*i.e.* low count) regions. This filtering, in fact, depends on the chosen statistical test. Because the two-sided test is more strict under a certain significance level α, a higher T threshold is obtained in the T method used by diffR (T-thresholds: diffR 14 (H3K4me3), 19 (H3K27me3), enrichR-compare 8 (H3K4me3), 11 (H3K27me3)). Indeed, most of the discrepancies between diffR and enrichR-compare were attributed to a more strict T threshold to eliminate low power regions in the two-sided binomial test (Fig. 6.4B): By applying the T thresholds of diffR to enrichR, enrichR-compare called on average $\sim$1.7-fold less cell type-exclusive regions (H3K27me3: 652,240; H3K4me3: 47,199). The sensitivity of diffR in recovering the cell type-exclusive enrichment could be increased substantially, *e.g.* only 2.99% false negatives for H3K4me3 in HepG2 cells (Table 6.1). Taken together, the low sensitivity of diffR to recover enrichR-compare’s results is attributed, on the one side, to diffR’s calls that represent true conditional differences of signal intensity (diffR’s false positives) and, on the other side, to differences in the statistical testing framework used (diffR’s false negatives).

In addition to discrepancies described above, some differences between diffR and enrichR-compare can be attributed to Copy Number Variations (CNVs) in HepG2 cells which are prevalent in immortalized cell types [156, 157]. To alleviate this problem, diffR was applied to HepG2 and PHH Input tracks with 20kb and 50kb windows. An initial visual investigation confirmed that diffR detected previously reported CNVs [43] on the short arm of human chromosome 14 in HepG2 cells (Fig. 6.5) Given the reasonable assumption that there are no CNVs in PHH, diffR recovered genome-wide 91% of 6,487 windows (odds-ratio=112.7) that overlapped 80 annotated large amplifications in HepG2 (13% of genome; median(length) = 163kb). Nevertheless, diffR did not detect 88% of 249 windows (odds-ratio=40.8) which overlapped 170 annotated very short heterozygous and homozygous deletions (6% of genome; median(length) = 9kb). The consistency of diffR and enrichR-compare calls could be improved by filtering out diffR called or experimentally validated CNV regions (Fig. 6.4C, D; Table 6.1). This approach reduced the number of diffR false

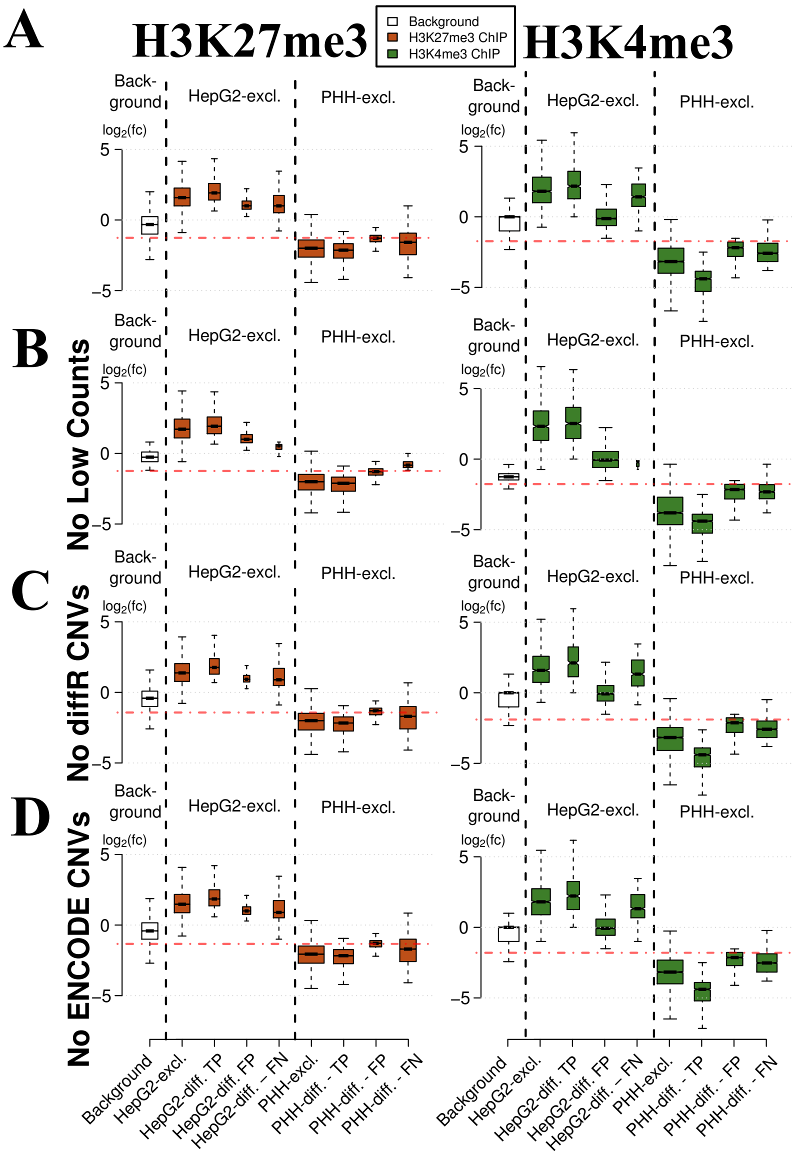

Fig. 6.4 – Discrepancies between diffR and enrichR-compare. (A) Conditional $\log_2$ fold change between HepG2 cells and Primary Human Hepatocytes (PHH) for H3K27me3 (left) and H3K4me3 (right) in background regions (no enrichment), "HepG2-exclusive" and "PHH-exclusive" regions. Based on each "cell type-exclusive" by enrichR-compare, boxes show diffR's true positives (TP), false positives (FP) and false negatives (FN). Conditional fold change in FP regions are equal or greater than the ones in TP regions suggesting genuine differences in ChIP-seq signal intensity. (B) Same as (A) with diffR's T Filter applied to enrichR-compare. The number of FN decreases substantially improving consistency between diffR and enrichR-compare (C) Same as (A) with diffR HepG2 CNVs removed. Most groups do not change but the number of FP in HepG2-differential calls is reduced. (D) Same as (A) with ENCODE HepG2 CNVs. See also Table 6.1 and Supplementary Fig. A.5. Whiskers extend up to 1.5-times the interquartile range. Width of boxes are proportional to the square-root of the number of regions in the groups.

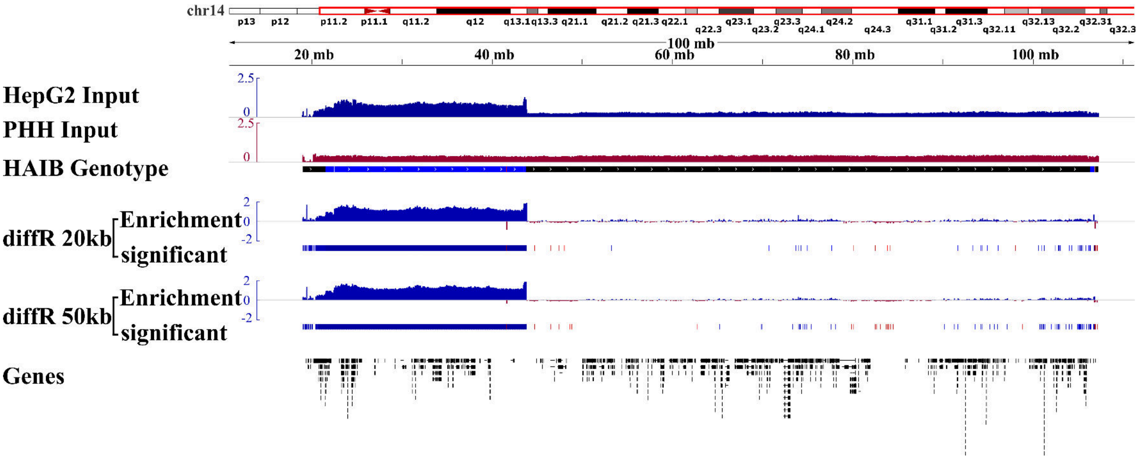

Fig. 6.5 – The Application of diffR on Input for HepG2 Cells and Primary Human Hepatocytes (PHH) Identifies Copy Number Variations (CNVs). Input-seq for HepG2 cells (blue) and PHH (red) indicate CNV presence in HepG2 cells on Human chromosome 14. The genotype of the HepG2 cell line (HAIB Genotype track [43]) encompasses amplified (blue) and deleted regions (red) that deviate from the reference genotype (black). diffR on HepG2 and PHH Input identified amplifications and deletions alike, assuming no CNVs in PHH. Performance values greater than 0.75 set in bold font.

positive calls and improves sensitivity (Table 6.1). In summary, diffR's agreement with enrichR-compare could be improved by filtering out *in silico* or *in vitro* inferred CNVs, yet not to the extent that was achieved by filtering for low count regions.

(b) Evaluation by a *Bona-Fide* Benchmark

On the second level, diffR results were compared systematically to those obtained from ChIPDiff [90], ODIN [92] and histoneHMM [91]. After calling differentially enriched regions with each tool, a trustworthy validation set ("*bona-fide* benchmark") was defined to assess their performance (see Methods section 6.2.4). For H3K27me3 at FDR=0.10, ChIPDiff was most precise ($\mu_{\text{Precision}}$=0.976) and diffR had the highest recall (μ_{Recall}=0.777), together with the best F$_{1.5}$-score ($\mu_{\text{F}_{1.5}\text{-score}}$=0.616; Table 6.2A). For H3K4me3 at FDR=0.05, histoneHMM was the most precise tool ($\mu_{\text{Precision}}$=0.585) and diffR had the highest recall (μ_{Recall}=0.797), together with the best F$_{1.5}$-score ($\mu_{\text{F}_{1.5}\text{-score}}$=0.539; Table 6.2B). These results showed that diffR slightly sacrifices precision for a substantial advance in sensitivity when compared to its competitors.

Next, I investigated the genuineness of the calls that are not represented by the *bona-fide* benchmark, *i.e.* "tool-specific" calls. A unified *bona-fide* benchmark revealed that the most tool-specific regions were called by ODIN (701.7Mb) and histoneHMM (689.1Mb) for H3K27me3 and by diffR (28.9Mb) and ODIN (25.4Mb) for H3K4me3 (Table 6.2). Turning to conditional fold changes for H3K27me3, only diffR- and histoneHMM-specific regions had fold changes comparable to those in the *bona-fide* benchmark (Fig. 6.6A). However, histoneHMM-specific regions had an average read coverage ≤18 in a 1,000bp window (Fig. 6.6B) indicating that these calls may be spurious. For H3K4me3, diffR-, ChIPDiff- and histoneHMM-specific regions had absolute fold changes greater or equal than the *bona-fide* benchmark (Fig. 6.6C). Among those diffR-specific regions had the highest read coverage (Fig. 6.6D) suggesting that these are valid calls. In conclusion, diffR identified conditional differences for, both, H3K27me3 and H3K4me3 which were supported by a good classifier performance, a high absolute fold change as well as a read coverage that is large enough to eliminate low power (*i.e.* low count) regions.

A H3K27me3

Method		**# regions**	**Scoring based on *bona-fide* benchmark**					
			Precision	Recall	$F_{0.5}$-score	F_1-score	$F_{1.5}$-score	# tool-specific
	diffR	251931	0.666	0.722	0.677	0.693	0.710	84069
HepG2	ChIPDiff	126747	**0.976**	0.401	**0.758**	0.568	0.455	3076
	ODIN	661239	0.264	**0.969**	0.309	0.415	0.632	486431
	histoneHMM	229209	0.627	0.570	0.615	0.597	0.581	85548
	diffR	596971	0.315	0.834	0.360	0.457	0.627	408805
PHH	ChIPDiff	9990	**0.975**	0.017	0.078	0.033	0.021	246
	ODIN	402129	0.465	0.419	0.455	0.441	0.428	215255
	histoneHMM	791961	0.238	0.653	0.273	0.349	0.484	603510
	diffR	848902	0.419	**0.777**	0.462	0.545	0.616	492874
combined	ChIPDiff	136737	**0.976**	0.150	0.464	0.260	0.202	3322
	ODIN	1063368	0.340	0.578	0.371	0.428	0.475	701686
	histoneHMM	1021170	0.325	0.614	0.359	0.425	0.482	689058

B H3K4me3

Method		**# regions**	**Scoring based on *bona-fide* benchmark**					
			Precision	Recall	$F_{0.5}$-score	F_1-score	$F_{1.5}$-score	# tool-specific
	diffR	39881	0.182	**0.865**	0.217	0.301	0.495	32603
HepG2	ChIPDiff	25193	0.290	0.290	0.290	0.290	0.290	17893
	ODIN	40464	0.139	0.419	0.161	0.209	0.299	34833
	histoneHMM	4065	0.713	0.095	0.311	0.168	0.115	1165
	diffR	44084	0.428	**0.774**	0.470	0.551	0.666	25212
PHH	ChIPDiff	33656	0.562	0.677	0.582	0.614	0.651	14738
	ODIN	24854	0.357	0.303	0.345	0.328	0.313	15984
	histoneHMM	26597	0.566	0.490	0.549	0.525	0.503	11547
	diffR	83965	0.311	**0.797**	0.355	0.448	0.539	57815
combined	ChIPDiff	58849	0.446	0.493	0.454	0.468	0.478	32631
	ODIN	65318	0.222	0.340	0.239	0.268	0.292	50817
	histoneHMM	30662	0.585	0.294	0.488	0.391	0.347	12712

Table 6.2 – Difference Calling Performance for diffR, ChIPDiff, ODIN and histoneHMM with Respect to a *Bona-Fide* Benchmark Set. (A) Performance for H3K27me3 ChIP-seq difference calling between HepG2 cells and Primary Human Hepatocytes (PHH) in 1,000bp bins. ODIN and histoneHMM call most regions, ChIPDiff calls least regions but is very precise and diffR has the highest recall and F_1-score. Most tool-specific bins are called by ODIN and histoneHMM. (B) Same as (A) for H3K4me3 in 500bp bins. diffR calls most tool-specific regions and has the highest recall and F_1-score.

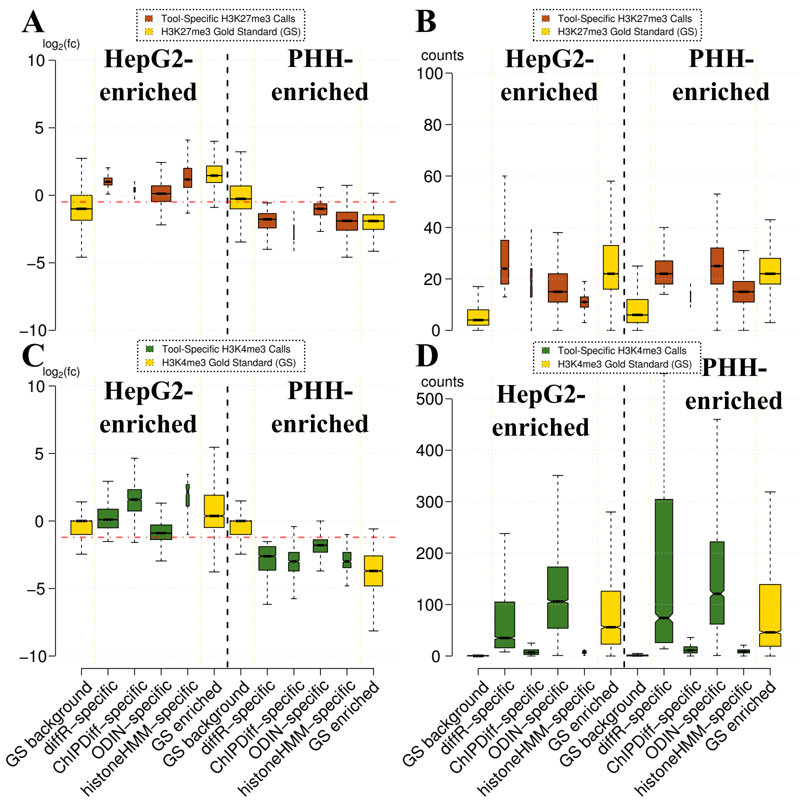

Fig. 6.6 – Fold-Changes and Read Coverage for diffR-, ChIPDiff-, ODIN- and histoneHMM-Specific Regions with Respect to the *Bona-Fide* Benchmark Set. (A) Conditional $\log_2$ fold changes for H3K27me3 in HepG2 cells and Primary Human Hepatocytes (PHH) in background (GS background), tool-specific and benchmark (GS enriched) regions for HepG2- (left) and PHH-differential (right). Fold changes in diffR- and histoneHMM-specific regions are equal to fold changes in GS enriched regions. (B) ChIP-seq read coverage in GS background, tool-specific and GS enriched regions for HepG2- (left) and PHH-differential (right). Only diffR-specific regions have a read coverage that is comparable to GS enriched regions. (C) Same as (A) for H3K4me3. diffR-, ChIPDiff- and histoneHMM-specific regions have fold changes that are comparable to GS enriched regions. Read coverage in diffR- and ODIN-specific regions is greater or equal to coverage in GS enriched regions.

6.4 Discussion

In this chapter I presented an implementation of the normR framework for the direct comparison of two ChIP-seq experiments, referred to as “diffR”. The diffR approach fits a mixture model with three components to simultaneously model background and two condition-specific foreground components. Herein, a robust background is estimated without the need for an Input experiment and this background estimation enables for the statistical inference of mutually exclusive and conditionally different enrichment alike. While mutually exclusive enrichment can also be identified by a qualitative integration of individual enrichment calls, the detection of conditionally different signal levels substantially increases the resolution in the comparison of two conditions. The latter scenario benefits from the diffR-estimated conditional enrichment e which directly relates to the difference in ChIP-seq signals to the conditional prevalence of the protein binding.

In a proof-of-principle study on H3K27me3 and H3K4me3 ChIP-seq data, a comparison of hepatocarcinoma cell line HepG2 and primary human hepatocytes revealed that the H3K27me3 heterochromatin covers less of the genome in HepG2 cells and that the H3K4me3 enrichment differs mostly on quantitative levels. diffR uncovered implications of the cancer-associated disruption of the hepatocyte heterochromatin in hepatocarcinoma cells (see [141] for review). Specifically, downstream of the crucial cell cycle regulator E2F2 [147] diffR detected a HepG2-specific shortening of a H3K27me3 domain accompanied by a HepG2-specific gain of H3K4me3 suggesting a potential *cis*-regulatory effect on E2F2. Detected quantitative differences in H3K4me3 levels correlate to expression level changes and are coincident with promoters of genes associated to functions like cell division for HepG2 cells and the P450 pathway for hepatocytes. This report on epigenetic alterations between cancer and primary cells could in the future be integrated with DNA methylation and mutational signatures [152] to understand the effects of immortalization on an epigenetic level.

When compared to other previously applied approaches, diffR is very precise and sensitive in detecting mutually exclusive enrichment and conditionally different signal intensities. Strikingly, even without an Input experiment, diffR recovers the majority of cell type-exclusive regions detected by enrichR when an identical T threshold is used prior to FDR correction. Furthermore, diffR's accuracy can be marginally increased by the incorporation of CNV information – either measured experimentally or by applying diffR directly on the Input tracks of the two conditions. A systematic benchmark based on a *bona-fide* benchmark set showed that diffR is much more sensitive than three competitor tools in both low S/N H3K27me3 and high S/N H3K4me3 ChIP-seq data. In the future, an experimentally validated gold-standard of conditional ChIP-seq enrichment has to be generated to approve the herein described validation set.

Recently, I used diffR to model enrichment in reChIP-seq data over a primary ChIP-seq experiment to detect H3K4me3-H3K27me3 bivalently modified nucleosomes genome-wide [7]. In the future, it is conceivable to also call mutually enriched regions in the two conditions with diffR by integrating the two foreground enrichment factors $\langle f_1 \rangle$ and $\langle f_2 \rangle$ into one multinomial component. The subtle influence of CNVs on diffR's performance could be diminished by an approach that jointly models conditional ChIP-seq tracks together with the respective Input tracks. Hereafter, diffR can detect conditional differences in other NGS experiments apart from ChIP-seq, *e.g.* STARR-seq [158] or ATAC-seq [159] – especially if there is no actual control experiment defined (as for the latter).

Part III

Conclusion

Conclusion

In this thesis I presented an extendable methodology called "normR" that enables for the extensive analysis of NGS read count data. By modeling foreground(s) and background jointly, normalization and difference calling are performed simultaneously using a intuitive binomial mixture model and robust statistics (Chapter 3). The implicit modeling of the effect of signal-regions on the overall read statistics results in an adequate normalization factor that increases the sensitivity in detecting regions with only minute signal over background. In this work I used the normR framework to analyze ChIP-seq read count data from the hepatocarcinoma cell line HepG2 and primary human hepatocytes (PHH) under three scenarios (Fig.6.7): (i) the identification of enriched genomic loci in low and high signal-to-noise ratio settings with `enrichR`; (ii) the dissection of ChIP enrichment in two distinct enrichment regimes with `regimeR`; and (iii) the stratification of conditional differences in ChIP enrichment between HepG2 cells and PHH with `diffR`.

A simple two-component implementation of the normR framework called `enrichR` was shown (in Chapter 4) to achieve a more sensitive enrichment calling than six competitor methods [74–80]. Due to the lack of a experimentally validated and comprehensive gold-standard for ChIP-seq, I introduce a *bona-fide* validation set based on a consensus vote among peak callers to systematically assess their performance of a cohort of peak callers. In this regard, the validity of thousands of enrichR-exclusive enrichment calls could be confirmed by auxiliary information like expression and DNA methylation. Yet, a comprehensive ChIP-seq gold-standard is needed to assuredly assess all available enrichment callers. enrichR's background normalization factor improves on current *in silico* and *in vitro* ChIP-seq normalization methods [85, 93]. Strikingly, my analysis of ICeChIP-seq data [93] revealed that the assumption of a linear relationship between the epitope spike-in and the ChIP-seq signal intensity may be incorrect. In a proof-of-principle study, I show how enrichR-based enrichment calls can vastly improve the chromatin segmentation with chromHMM [36] by resolving the majority of the epigenome and detecting a novel state characterized by histone modification patterns of poised enhancers. A possible extension to enrichR represents the incorporation of replicates to account also for biological variability within the sample condition.

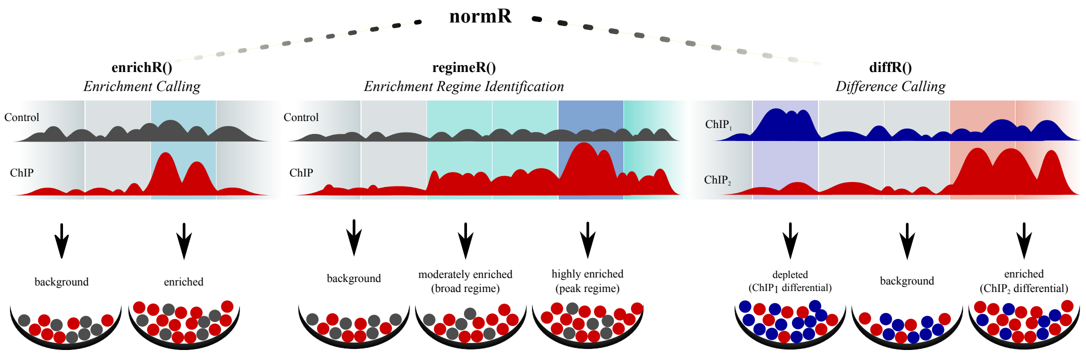

Fig. 6.7 – The normR Approach: Normalization and Difference Calling in NGS Data.

The identification of two distinct enrichment regimes in facultative and constitutive heterochromatin was achieved with a three-component implementation of the normR framework referred to as `regimeR` (Chapter 5) For the first time, one principled approach was able to dissect heterogeneous enrichment into *peak* and *broad* enrichment in, both, H3K27me3 and H3K9me3 ChIP-seq data. Apart from distinctive genetic and epigenetic features associated to the two regimes, my findings suggest a novel mode of action in the preservation of heterochromatin where high propensity heterochromatic regions function as nucleation sites for large heterochromatic domains. The disparities in heterochromatin between hepatocarcinoma cells and their cell type-of-origin, *i.e.* PHH, begs the question how cancer and immortalization affect the stability of heterochromatin [139]. In the future, an automated determination of the number of enrichment components by means of a Dirichlet Process or a β-binomial foreground component (Section 3.2.4) will be adjuvant in studying epigenomic heterogeneity in conjunction with recently reported single cell ChIP- seq data [160].

The direct comparison of two ChIP-seq experiments was enabled in Chapter 6 by `diffR` which also uses a three-component implementation of the normR framework to model conditional differences. In an exemplary study, I could show that HepG2 cells have a diminished H3K27me3 decoration when compared to PHH and that most differences in H3K4me3 are of quantitative nature. Recently, I used diffR to identify co-localizing histone modifications in a novel reChIP-seq data set [7] where the background estimation is complicated by the presence of enrichment in the control ChIP-seq experiment. The diffR framework does not require a control and I anticipate it would be beneficial to identify conditional differences in assays where no technical control experiment is defined, *e.g.* ATAC-seq [159]. In the future, diffR could be extended to account for biases, *e.g.* CNVs, by the incorporation of condition-specific control experiments.

Taken together normR proved as a versatile and sensitive framework for the analysis of ChIP-seq data. In principle, it is readily available for the analysis of conditional differences in other NGS-based experiments, *e.g.* STARR-seq [158] or even RNA-seq.

Bibliography

[1] Alessandro Mammana and Johannes Helmuth. Bamsignals: Extract Read Count Signals from Bam Files. `http://bioconductor.org/packages/bamsignals`, 2016. R Bioconductor package.

[2] Johannes Helmuth, Na Li, Laura Arrigoni, Kathrin Gianmoena, Cristina Cadenas, et al. normR: Regime Enrichment Calling for ChIP-seq Data. *bioRxiv*, page http://dx.doi.org/10.1101/082263, October 2016. `http://biorxiv.org/content/early/2016/10/25/082263`.

[3] Johannes Helmuth and Ho-Ryun Chung. Auswertung von Histonmodifikations-ChIP-Seq-Datensätzen. *BIOspektrum*, 22(6):568–570, 2016. `http://link.springer.com/article/10.1007/s12268-016-0728-6`.

[4] Johannes Helmuth and Ho-Ryun Chung. normR: Normalization and Difference Calling in ChIP-seq Data. `http://bioconductor.org/packages/normr/`, 2016. R Bioconductor package.

[5] Johannes Helmuth, Na Li, Sarah Kinkley, and Ho-Ryun Chung. The Tale of Two Tails. In *FEBS JOURNAL*, volume 282, pages 37–38. WILEY-BLACKWELL 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2015. `http://scholar.google.com/scholar?cluster=8482313655783029739&hl=en&oi=scholarr`.

[6] Jun Yang, M.-Hossein Moeinzadeh, Heiner Kuhl, Johannes Helmuth, Peng Xiao, et al. The haplotype-resolved genome sequence of hexaploid Ipomoea batatas reveals its evolutionary history. *bioRxiv*, page 064428, 2016. `http://biorxiv.org/content/early/2016/07/18/064428.abstract`.

[7] Sarah Kinkley and Johannes Helmuth, Julia K. Polansky, Ilona Dunkel, Gilles Gasparoni, Sebastian Fröhler, et al. reChIP-seq reveals widespread bivalency of H3K4me3 and H3K27me3 in CD4(+) memory T cells. *Nat Commun*, 7:12514, August 2016. 00000.

[8] James D. Watson, Francis HC Crick, and others. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953. `http://www.nature.com/physics/looking-back/crick/`.

[9] Zephyris. English: The structure of DNA showing with detail showing the structure of the four bases, adenine, cytosine, guanine and thymine, and the location of the major and minor groove. `https://commons.wikimedia.org/wiki/File:DNA_Structure%2BKey%2BLabelled.pn_NoBB.png`, April 2011. 00000.

[10] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001. `http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html`.

[11] Yair Field, Evan A. Boyle, Natalie Telis, Ziyue Gao, Kyle J. Gaulton, et al. Detection of human adaptation during the past 2000 years. *Science*, page aag0776, October 2016. `http://science.sciencemag.org/content/early/2016/10/12/science.aag0776`.

[12] R. D. Kornberg. Chromatin structure: A repeating unit of histones and DNA. *Science*, 184(4139):868–871, May 1974. 01972.

[13] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389(6648):251–260, September 1997. 00045.

[14] Original uploader was Richard Wheeler at en.wikipedia. Čeština: Obrázek zobrazuje vyšší úrovně skládání molekuly DNA až do Chromatinu. `https://commons.wikimedia.org/wiki/File:Chromatin_Structures.png`, November 2006. 00000.

[15] V. A. Spencer and J. R. Davie. Role of covalent modifications of histones in regulating gene expression. *Gene*, 240(1):1–12, November 1999. 00330.

[16] X. Wang, C. He, S. C. Moore, and J. Ausio. Effects of histone acetylation on the solubility and folding of the chromatin fiber. *J. Biol. Chem.*, 276(16):12764–12768, April 2001. 00089.

[17] Andrew J. Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell Res.*, 21(3):381–395, March 2011. 01522.

[18] T. Jenuwein and C. D. Allis. Translating the histone code. *Science*, 293(5532):1074–1080, August 2001. 08096.

[19] B. D. Strahl and C. D. Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–45, January 2000. 06941.

[20] Michael Weber and Dirk Schübeler. Genomic patterns of DNA methylation: Targets and function of an epigenetic mark. *Curr. Opin. Cell Biol.*, 19(3):273–280, June 2007. 00287.

[21] Walfred W. C. Tang, Sabine Dietmann, Naoko Irie, Harry G. Leitch, Vasileios I. Floros, et al. A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development. *Cell*, 161(6):1453–1467, June 2015. 00100.

[22] Scott Berry, Matthew Hartley, Tjelvar S. G. Olsson, Caroline Dean, and Martin Howard. Local chromatin environment of a Polycomb target gene instructs its own epigenetic inheritance. *Elife*, 4, May 2015. 00020.

[23] Juan M. Vaquerizas and Maria-Elena Torres-Padilla. Developmental biology: Panoramic views of the early epigenome. *Nature*, 537(7621):494–496, September 2016. `http://www.nature.com/nature/journal/v537/n7621/full/nature19468.html`.

[24] Johannes Bohacek and Isabelle M. Mansuy. Molecular insights into transgenerational non-genetic inheritance of acquired behaviours. *Nat. Rev. Genet.*, 16(11):641–652, November 2015. 00027.

[25] Francis Crick and others. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970. `http://cs.brynmawr.edu/Courses/cs380/fall2012/CrickCentralDogma1970.pdf`.

[26] Jack D. Keene. RNA regulons: Coordination of post-transcriptional events. *Nat Rev Genet*, 8(7):533–543, July 2007. `http://www.nature.com/nrg/journal/v8/n7/full/nrg2111.html`.

[27] Bradley M. Lunde, Claire Moore, and Gabriele Varani. RNA-binding proteins: Modular design for efficient function. *Nat. Rev. Mol. Cell Biol.*, 8(6):479–490, June 2007. 00556.

[28] Roger D. Kornberg. The molecular basis of eukaryotic transcription. *Proc. Natl. Acad. Sci. U.S.A.*, 104(32):12955–12961, August 2007. 00292.

[29] Tim R. Mercer, Marcel E. Dinger, and John S. Mattick. Long non-coding RNAs: Insights into functions. *Nat Rev Genet*, 10(3):155–159, March 2009. `http://www.nature.com/nrg/journal/v10/n3/abs/nrg2521.html`.

[30] Michael Petrascheck, Dominik Escher, Tokameh Mahmoudi, C. Peter Verrijzer, Walter Schaffner, et al. DNA looping induced by a transcriptional enhancer in vivo. *Nucleic Acids Res.*, 33(12):3743–3750, 2005. 00062.

[31] Alvaro Sebastian and Bruno Contreras-Moreira. footprintDB: A database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*, 30(2):258–265, January 2014. 00017.

[32] Helge G. Roider, Aditi Kanhere, Thomas Manke, and Martin Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, January 2007. 00150.

[33] Philipp A. Steffen and Leonie Ringrose. What are memories made of? How Polycomb and Trithorax proteins mediate epigenetic memory. *Nat Rev Mol Cell Biol*, 15(5):340–356, May 2014. `http://www.nature.com/nrm/journal/v15/n5/full/nrm3789.html`.

[34] Rosa Karlić, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahovicek, and Martin Vingron. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 107(7):2926–2931, February 2010. 00332.

[35] Bing Li, Michael Carey, and Jerry L. Workman. The role of chromatin during transcription. *Cell*, 128(4):707–719, February 2007. 02496.

[36] Jason Ernst and Manolis Kellis. ChromHMM: Automating chromatin-state discovery and characterization. *Nat. Methods*, 9(3):215–216, February 2012. 00414.

[37] Adriana Gonzalez-Sandoval and Susan M. Gasser. On TADs and LADs: Spatial Control Over Gene Expression. *Trends Genet.*, 32(8):485–495, August 2016. 00007.

[38] R. K. Saiki, D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839):487–491, January 1988. 19332.

[39] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, January 2009. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/.

[40] Toshiyuki Shiraki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *PNAS*, 100(26):15776–15781, December 2003. http://www.pnas.org/content/100/26/15776.

[41] Hazuki Takahashi, Timo Lassmann, Mitsuyoshi Murata, and Piero Carninci. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc*, 7(3):542–561, February 2012. 00078.

[42] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, et al. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, May 2007. 04388.

[43] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012. 04636.

[44] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, February 2015. 00630.

[45] Morgane Thomas-Chollier, Elodie Darbo, Carl Herrmann, Matthieu Defrance, Denis Thieffry, et al. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nat Protoc*, 7(8):1551–1568, July 2012. 00042.

[46] Alessandro Mammana and Ho-Ryun Chung. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biol.*, 16:151, July 2015. 00018.

[47] Goran Kungulovski, Ina Kycia, Raluca Tamas, Renata Z. Jurkowska, Srikanth Kudithipudi, et al. Application of histone modification-specific interaction domains as an alternative to antibodies. *Genome Res.*, 24(11):1842–1853, November 2014. 00011.

[48] John C. Marioni, Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18(9):1509–1517, January 2008. http://genome.cshlp.org/content/18/9/1509.

[49] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010. http://bioinformatics.oxfordjournals.org/content/26/1/139.

[50] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010. http://dx.doi.org/10.1186/gb-2010-11-10-r106.

[51] George Casella and Roger L. Berger. *Statistical Inference*. Cengage Learning, Australia ; Pacific Grove, CA, second edition, June 2001. 07886.

[52] William S. Noble. How does multiple testing correction work? *Nat. Biotechnol.*, 27(12):1135–1137, December 2009. 00181.

[53] Carlo E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Libreria internazionale Seeber, 1936. 00000.

[54] Yoav Benjamini and Yosef 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. http://www.jstor.org/stable/2346101.

[55] Stan Pounds and Cheng Cheng. Improving false discovery rate estimation. *Bioinformatics*, 20(11):1737–1745, July 2004. http://bioinformatics.oxfordjournals.org/content/20/11/1737.

[56] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, August 2002. http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00346/abstract.

[57] Martin Krzywinski and Naomi Altman. Points of significance: Comparing samples—part II. *Nat Meth*, 11(4):355–356, April 2014. http://www.nature.com/nmeth/journal/v11/n4/full/nmeth.2900.html.

[58] Stan Pounds and Stephan W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, July 2003. 00343.

[59] David B. Allison, Gary L. Gadbury, Moonseong Heo, José R. Fernández, Cheol-Koo Lee, et al. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39(1):1–20, March 2002. https://www.sciencedirect.com/science/article/pii/S0167947301000469.

[60] R. A. Fisher. On the Interpretation of $x2$ from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922. http://www.jstor.org/stable/2340521.

[61] Isaac Dialsingh, Stefanie R. Austin, and Naomi S. Altman. Estimating the proportion of true null hypotheses when the statistics are discrete. *Bioinformatics*, 31(14):2303–2309, July 2015. http://bioinformatics.oxfordjournals.org/content/31/14/2303.

[62] R. E. Tarone. A Modified Bonferroni Method for Discrete Data. *Biometrics*, 46(2):515–522, 1990. http://www.jstor.org/stable/2531456.

[63] Isaac Dialsingh. *False Discovery Rates When the Statistics Are Discrete*. Ph.D. thesis, December 2011. https://etda.libraries.psu.edu/catalog/12650.

[64] Tim Bancroft, Chuanlong Du, and Dan Nettleton. Estimation of False Discovery Rate Using Sequential Permutation p -Values: Sequential Permutation p -Values. *Biometrics*, 69(1):1–7, March 2013. http://doi.wiley.com/10.1111/j.1541-0420.2012.01825.x.

[65] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009. http://link.springer.com/10.1007/978-0-387-84858-7.

[66] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15(12), 2014. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4302049/.

[67] R. A. Fisher. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 222(594-604):309–368, January 1922. http://rsta.royalsocietypublishing.org/content/222/594-604/309.

[68] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, first edition, 2007. 00058.

[69] Devinderjit Sivia and John Skilling. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, U.S.A., Oxford; New York, second edition, July 2006. 02181.

[70] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. http://www.jstor.org/stable/2984875.

[71] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Disentangling Gaussians. *Communications of the ACM*, 55(2):113, February 2012. http://dl.acm.org/citation.cfm?doid=2076450.2076474.

[72] Clifford A. Meyer and X. Shirley Liu. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.*, 15(11):709–721, November 2014. 00050.

[73] Melanie Schirmer, Umer Z. Ijaz, Rosalinda D'Amore, Neil Hall, William T. Sloan, et al. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.*, 43(6):e37, March 2015. 00087.

[74] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-9-r137.

[75] Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. Identifying ChIP-seq enrichment using MACS. *Nat. Protocols*, 7(9):1728–1740, September 2012. http://www.nature.com/nprot/journal/v7/n9/full/nprot.2012.101.html.

[76] Vibhor Kumar, Masafumi Muratani, Nirmala Arul Rayan, Petra Kraus, Thomas Lufkin, et al. Uniform, optimal signal processing of mapped deep-sequencing data. *Nat Biotech*, 31(7):615–622, July 2013. http://www.nature.com/nbt/journal/v31/n7/full/nbt.2596.html.

[77] Hongkai Ji, Hui Jiang, Wenxiu Ma, David S. Johnson, Richard M. Myers, et al. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, 26(11):1293–1300, November 2008. 00632.

[78] Peter V. Kharchenko, Michael Y. Tolstorukov, and Peter J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, 26(12):1351–1359, December 2008. 00485.

[79] Haipeng Xing, Yifan Mo, Will Liao, and Michael Q. Zhang. Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. *PLoS Comput. Biol.*, 8(7):e1002613, 2012. 00023.

[80] Arif Harmanci, Joel Rozowsky, and Mark Gerstein. MUSIC: Identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol.*, 15(10):474, 2014. 00004.

[81] R. Core Team and others. R: A language and environment for statistical computing. 2013. `http://cran.fiocruz.br/web/packages/dplR/vignettes/timeseries-dplR.pdf`.

[82] Wolfgang Huber, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, 12(2):115–121, February 2015. 00210.

[83] Vinsensius B. Vega, Edwin Cheung, Nallasivam Palanisamy, and Wing-Kin Sung. Inherent signals in sequencing-based Chromatin-ImmunoPrecipitation control libraries. *PLoS ONE*, 4(4):e5241, 2009. 00037.

[84] Christoffer Flensburg, Sarah A. Kinkel, Andrew Keniry, Marnie E. Blewitt, and Alicia Oshlack. A comparison of control samples for ChIP-seq of histone modifications. *Front. Genet*, 5:329, 2014. `http://journal.frontiersin.org/article/10.3389/fgene.2014.00329/abstract`.

[85] Kun Liang and Sündüz Keleş. Normalization of ChIP-seq data with control. *BMC Bioinformatics*, 13:199, August 2012. 00040.

[86] Aaron Diaz, Kiyoub Park, Daniel A. Lim, and Jun S. Song. Normalization, bias correction, and peak calling for ChIP-seq. *Stat Appl Genet Mol Biol*, 11(3):Article 9, March 2012. 00031.

[87] David Sims, Ian Sudbery, Nicholas E. Ilott, Andreas Heger, and Chris P. Ponting. Sequencing depth and coverage: Key considerations in genomic analyses. *Nat. Rev. Genet.*, 15(2):121–132, February 2014. 00279.

[88] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates Inc, Hillsdale, N.J, second edition, August 1988. 01307.

[89] E. T. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press, Cambridge, UK ; New York, NY, first edition, June 2003. 05119.

[90] Han Xu, Chia-Lin Wei, Feng Lin, and Wing-Kin Sung. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics*, 24(20):2344–2349, October 2008. 00127.

[91] Matthias Heinig, Maria Colomé-Tatché, Aaron Taudt, Carola Rintisch, Sebastian Schafer, et al. histoneHMM: Differential analysis of histone modifications with broad genomic footprints. *BMC Bioinformatics*, 16(1):60, February 2015. `http://www.biomedcentral.com/1471-2105/16/60/abstract`.

[92] Manuel Allhoff, Kristin Seré, Heike Chauvistré, Qiong Lin, Martin Zenke, et al. Detecting differential peaks in ChIP-seq signals with ODIN. *Bioinformatics*, 30(24):3467–3475, December 2014. 00008.

[93] Adrian T. Grzybowski, Zhonglei Chen, and Alexander J. Ruthenburg. Calibrating ChIP-Seq with Nucleosomal Internal Standards to Measure Histone Modification Density Genome Wide. *Molecular Cell*, 58(5):886–899, June 2015. `http://www.cell.com/molecular-cell/abstract/S1097-2765(15)00304-4`.

[94] Alejandra Medina-Rivera, Matthieu Defrance, Olivier Sand, Carl Herrmann, Jaime A. Castro-Mondragon, et al. RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res.*, 43(W1):W50–56, July 2015. 00044.

[95] Nathaniel D. Heintzman, Rhona K. Stuart, Gary Hon, Yutao Fu, Christina W. Ching, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, 39(3):311–318, March 2007. 01863.

[96] Nathaniel D. Heintzman, Gary C. Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, May 2009. 01363.

[97] Alexander M. Tsankov, Hongcang Gu, Veronika Akopian, Michael J. Ziller, Julie Donaghey, et al. Transcription factor binding dynamics during human ES cell differentiation. *Nature*, 518(7539):344–349, February 2015. 00055.

[98] Stefan Bonn, Robert P. Zinzen, Charles Girardot, E. Hilary Gustafson, Alexis Perez-Gonzalez, et al. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.*, 44(2):148–156, January 2012. 00230.

[99] Reuben Thomas, Sean Thomas, Alisha K. Holloway, and Katherine S. Pollard. Features that define the best ChIP-seq peak calling algorithms. *Brief Bioinform*, page bbw035, May 2016. `http://bib.oxfordjournals.org/content/early/2016/05/10/bib.bbw035`.

[100] Francesco Strino, Fabio Parisi, and Yuval Kluger. VDA, a method of choosing a better algorithm with fewer validations. *PLoS ONE*, 6(10):e26074, 2011. 00004.

[101] Mariann Micsinai, Fabio Parisi, Francesco Strino, Patrik Asp, Brian D. Dynlacht, et al. Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic Acids Res.*, 40(9):e70, May 2012. 00049.

[102] Fabio Parisi, Francesco Strino, Boaz Nadler, and Yuval Kluger. Ranking and combining multiple predictors without labeled data. *Proc. Natl. Acad. Sci. U.S.A.*, 111(4):1253–1258, January 2014. 00028.

[103] Leonid Teytelman, Deborah M. Thurtle, Jasper Rine, and Alexander van Oudenaarden. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 110(46):18602–18607, November 2013. 00131.

[104] Steen K. T. Ooi, Chen Qiu, Emily Bernstein, Keqin Li, Da Jia, et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, 448(7154):714–717, August 2007. 00968.

[105] Hannah K. Long, David Sims, Andreas Heger, Neil P. Blackledge, Claudia Kutter, et al. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *Elife*, 2:e00348, February 2013. 00083.

[106] Jia-Lei Hu, Bo O. Zhou, Run-Rui Zhang, Kang-Ling Zhang, Jin-Qiu Zhou, et al. The N-terminus of histone H3 is required for de novo DNA methylation in chromatin. *Proc. Natl. Acad. Sci. U.S.A.*, 106(52):22187–22192, December 2009. 00079.

[107] AeRi Kim, Christine M. Kiefer, and Ann Dean. Distinctive signatures of histone methylation in transcribed coding and noncoding human beta-globin sequences. *Mol. Cell. Biol.*, 27(4):1271–1279, February 2007. 00000.

[108] Tuncay Baubec, Daniele F. Colombo, Christiane Wirbelauer, Juliane Schmidt, Lukas Burger, et al. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature*, 520(7546):243–247, April 2015. 00095.

[109] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009. 09829.

[110] Fidel Ramírez, Friederike Dündar, Sarah Diehl, Björn A. Grüning, and Thomas Manke. deepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, 42(Web Server issue):W187–191, July 2014. 00100.

[111] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedländer, Peter A. C. 't Hoen, Jean Monlong, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, September 2013. 00517.

[112] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, et al. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14(4):R36, April 2013. 02286.

[113] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, January 2015. 01374.

[114] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.*, 22(9):1760–1774, September 2012. 01041.

[115] FANTOM Consortium and the RIKEN PMI and CLST (DGT), Alistair R. R. Forrest, Hideya Kawaji, Michael Rehli, J. Kenneth Baillie, et al. A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470, March 2014. 00000.

[116] Yaping Liu, Kimberly D. Siegmund, Peter W. Laird, and Benjamin P. Berman. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol.*, 13(7):R61, July 2012. 00072.

[117] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, March 2012. 04976.

[118] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, 9(8):e1003118, 2013. 00295.

[119] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009. 10010.

[120] Aaron R. Quinlan. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics*, 47:11.12.1–34, September 2014. 00108.

[121] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucl. Acids Res.*, 30(1):207–210, January 2002. `http://nar.oxfordjournals.org/content/30/1/207`.

[122] Michael Lawrence, Robert Gentleman, and Vincent Carey. Rtracklayer: An R package for interfacing with genome browsers. *Bioinformatics*, 25(14):1841–1842, July 2009. `https://academic.oup.com/bioinformatics/article/25/14/1841/225816/rtracklayer-an-R-package-for-interfacing-with`.

[123] Eli Eisenberg and Erez Y. Levanon. Human housekeeping genes, revisited. *Trends Genet.*, 29(10):569–574, October 2013. 00185.

[124] Sven Heinz, Casey E. Romanoski, Christopher Benner, and Christopher K. Glass. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol*, 16(3):144–154, March 2015. `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517609/`.

[125] Birgit Czermin, Raffaella Melfi, Donna McCabe, Volker Seitz, Axel Imhof, et al. Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell*, 111(2):185–196, October 2002. 01140.

[126] Jürg Müller, Craig M. Hart, Nicole J. Francis, Marcus L. Vargas, Aditya Sengupta, et al. Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. *Cell*, 111(2):197–208, October 2002. 01101.

[127] Andrei Kuzmichev, Kenichi Nishioka, Hediye Erdjument-Bromage, Paul Tempst, and Danny Reinberg. Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. *Genes Dev.*, 16(22):2893–2905, November 2002. 01123.

[128] Ru Cao, Liangjun Wang, Hengbin Wang, Li Xia, Hediye Erdjument-Bromage, et al. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science*, 298(5595):1039–1043, November 2002. 02274.

[129] Serge Saxonov, Paul Berg, and Douglas L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U.S.A.*, 103(5):1412–1417, January 2006. 00832.

[130] Elisabeth Wachter, Timo Quante, Cara Merusi, Aleksandra Arczewska, Francis Stewart, et al. Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. *Elife*, 3:e03397, September 2014. 00027.

[131] Ryan Rickels, Deqing Hu, Clayton K. Collings, Ashley R. Woodfin, Andrea Piunti, et al. An Evolutionary Conserved Epigenetic Mark of Polycomb Response Elements Implemented by Trx/MLL/COMPASS. *Molecular Cell*, 63(2):318–328, July 2016. `http://www.cell.com/molecular-cell/abstract/S1097-2765(16)30276-3`.

[132] Seth Frietze, Henriette O'Geen, Kimberly R. Blahnik, Victor X. Jin, and Peggy J. Farnham. ZNF274 recruits the histone methyltransferase SETDB1 to the 3' ends of ZNF genes. *PLoS ONE*, 5(12):e15082, December 2010. 00085.

[133] Adam Siepel and David Haussler. Phylogenetic hidden Markov models. In *Statistical Methods in Molecular Evolution*, pages 325–351. Springer, 2005. `http://link.springer.com/content/pdf/10.1007/0-387-27733-1_12.pdf`.

[134] Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8):1034–1050, January 2005. `http://genome.cshlp.org/content/15/8/1034`.

[135] RepeatMasker Open-4.0. `http : / / www . repeatmasker . org`, 2013-2015. Http://www.repeatmasker.org.

[136] Timothy L. Bailey, Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, et al. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.*, 37(Web Server issue):W202–208, July 2009. 02057.

[137] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, April 2011. 00582.

[138] Raul Urrutia. KRAB-containing zinc-finger repressor proteins. *Genome Biology*, 4:231, 2003. `http://dx.doi.org/10.1186/gb-2003-4-10-231`.

[139] Paul L. Severson, Erik J. Tokar, Lukas Vrba, Michael P. Waalkes, and Bernard W. Futscher. Coordinate H3K9 and DNA methylation silencing of ZNFs in toxicant-induced malignant transformation. *Epigenetics*, 8(10):1080–1088, October 2013. 00015.

[140] Kimberly R. Blahnik, Lei Dou, Lorigail Echipare, Sushma Iyengar, Henriette O'Geen, et al. Characterization of the contradictory chromatin signatures at the 3' exons of zinc finger genes. *PLoS ONE*, 6(2):e17121, February 2011. 00046.

[141] Jiyong Wang, Sharon T. Jia, and Songtao Jia. New Insights into the Regulation of Heterochromatin. *Trends Genet.*, 32(5):284–294, May 2016. 00002.

[142] Amos Tanay, Anne H. O'Donnell, Marc Damelin, and Timothy H. Bestor. Hyperconserved CpG domains underlie Polycomb-binding sites. *Proc. Natl. Acad. Sci. U.S.A.*, 104(13):5521–5526, March 2007.

[143] Reyad A. Elbarbary, Bronwyn A. Lucas, and Lynne E. Maquat. Retrotransposons as regulators of gene expression. *Science*, 351(6274):aac7247, February 2016. 00018.

[144] Kirsty Jamieson, Elizabeth T. Wiles, Kevin J. McNaught, Simone Sidoli, Neena Leggett, et al. Loss of HP1 causes depletion of H3K27me3 from facultative heterochromatin and gain of H3K27me2 at constitutive heterochromatin. *Genome Res.*, November 2015. `http://genome.cshlp.org/content/early/2015/12/07/gr.194555.115`.

[145] Angelo Lupo, Elena Cesaro, Giorgia Montano, Diana Zurlo, Paola Izzo, et al. KRAB-Zinc Finger Proteins: A Repressor Family Displaying Multiple Biological Functions. *Curr Genomics*, 14(4):268–278, June 2013. `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3731817/`.

[146] Pi-Xiao Wang, Xiao-Jing Zhang, Pengcheng Luo, Xi Jiang, Peng Zhang, et al. Hepatocyte TRAF3 promotes liver steatosis and systemic insulin resistance through targeting TAK1-dependent signalling. *Nature Communications*, 7:10592, February 2016. `http://www.nature.com/ncomms/2016/160217/ncomms10592/full/ncomms10592.html`.

[147] Eva Ramboer, Bram De Craene, Joery De Kock, Tamara Vanhaecke, Geert Berx, et al. Strategies for immortalization of primary hepatocytes. *Journal of Hepatology*, 61(4):925–943, October 2014. `http://www.sciencedirect.com/science/article/pii/S0168827814003948`.

[148] Elena Grassi, Ettore Zapparoli, Ivan Molineris, and Paolo Provero. Total Binding Affinity Profiles of Regulatory Regions Predict Transcription Factor Binding and Gene Expression in Human Cells. *PLoS ONE*, 10(11):e0143627, 2015. 00001.

[149] Justin Crocker, Ella Preger-Ben Noon, and David L. Stern. The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. *Curr. Top. Dev. Biol.*, 117:455–469, 2016. 00006.

[150] Sebastiaan H. Meijsing, Miles A. Pufall, Alex Y. So, Darren L. Bates, Lin Chen, et al. DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science*, 324(5925):407–410, April 2009. 00415.

[151] Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nat Biotech*, 28(10):1057–1068, October 2010. `http://www.nature.com/nbt/journal/v28/n10/abs/nbt.1685.html`.

[152] Paz Polak, Rosa Karlić, Amnon Koren, Robert Thurman, Richard Sandstrom, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, 518(7539):360–364, February 2015. 00088.

[153] C. Sardet, M. Vidal, D. Cobrinik, Y. Geng, C. Onufryk, et al. E2F-4 and E2F-5, two members of the E2F family, are expressed in the early phases of the cell cycle. *Proc. Natl. Acad. Sci. U.S.A.*, 92(6):2403–2407, March 1995. 00338.

[154] Yannick Sylvestre, Vincent De Guire, Emmanuelle Querido, Utpal K. Mukhopadhyay, Véronique Bourdeau, et al. An E2F/miR-20a autoregulatory feedback loop. *J. Biol. Chem.*, 282(4):2135–2143, January 2007. 00524.

[155] Sara Aibar, Celia Fontanillo, Conrad Droste, and Javier De Las Rivas. Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics*, 31(10):1686–1688, May 2015. 00003.

[156] Matthew D. Shirley, Joseph D. Baugher, Eric L. Stevens, Zhenya Tang, Norman Gerry, et al. Chromosomal variation in lymphoblastoid cell lines. *Hum. Mutat.*, 33(7):1075–1086, July 2012. 00012.

[157] Donald F. Conrad, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, et al. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, April 2010. 01378.

[158] Cosmas D. Arnold, Daniel Gerlach, Christoph Stelzer, Łukasz M. Boryń, Martina Rath, et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339(6123):1074–1077, March 2013. 00213.

[159] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, 10(12):1213–1218, December 2013. 00407.

[160] Assaf Rotem, Oren Ram, Noam Shoresh, Ralph A. Sperling, Alon Goren, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*, 33(11):1165–1172, November 2015. 00063.

Appendices

Appendix A

Supplementary Figures

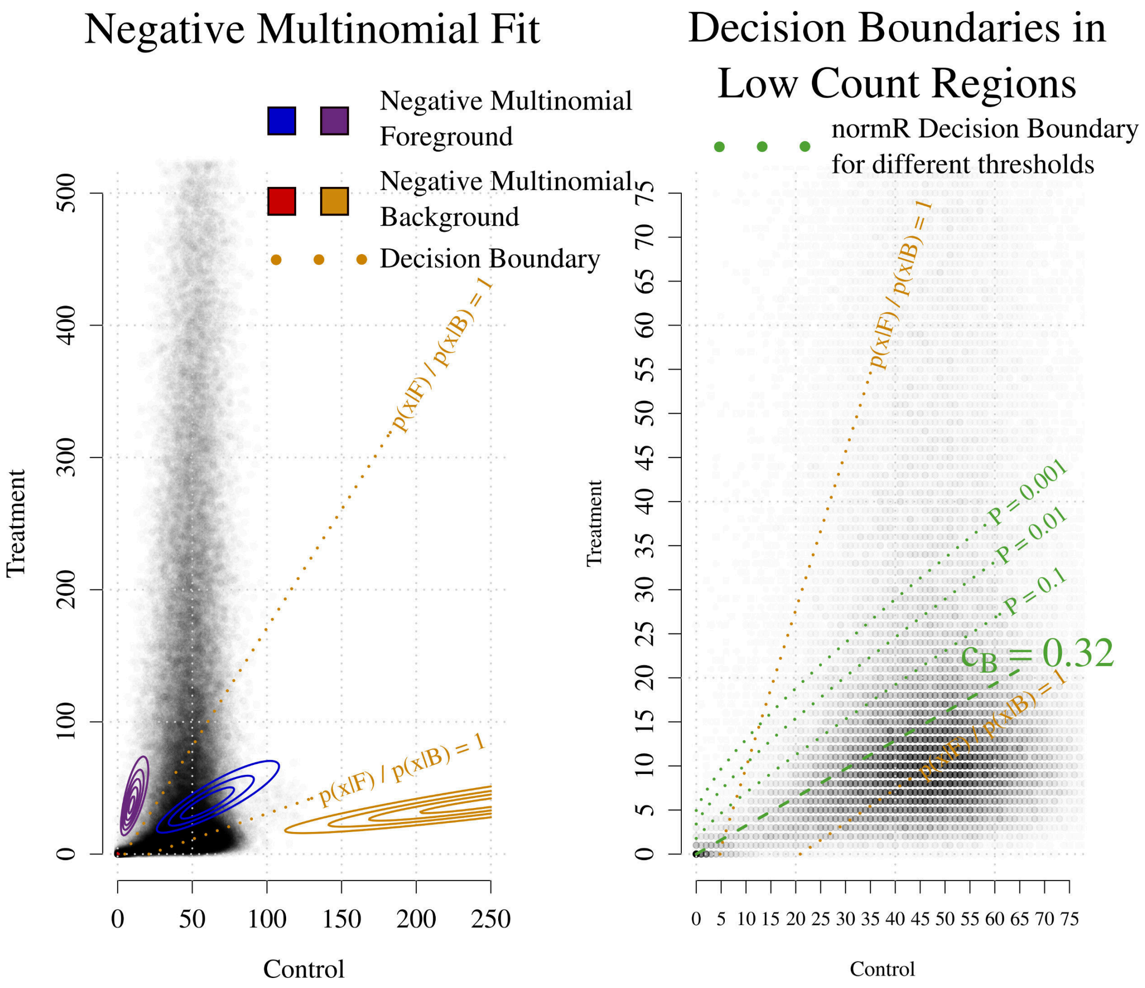

Fig. A.1 – A Negative Multinomial 4-Mixture Fit Does Not Model The Interrelation between Treatment and Control. A 4-Mixture Model of Negative Multinomials models the density of read counts in two dimensions resulting in a separation of low and high count regions (left; Decision boundary based on likelihood ratio given). Low count regions (5≥Control≤20) are not classified as background in the Negative Multinomial Mixture and the normR classification is more accurate (right). See also Fig. 3.5.

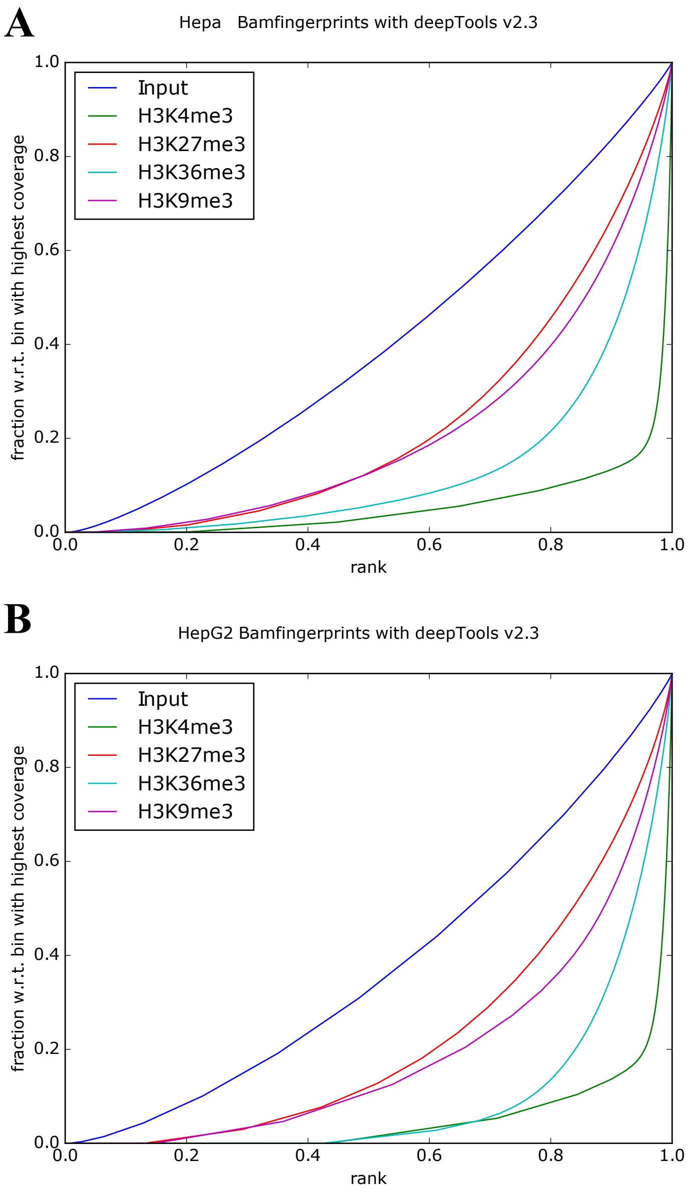

Fig. A.2 – HepG2 and Primary Human Hepatocytes ChIP-seq Data Quality Measured with BamFingerprint. Cumulative fraction of ChIP-seq, Input reads with respect to (w.r.t.) to bin with highes coverage in PHH (A) and HepG2 cells (B) computed with deepTools [110]. Input experiments are almost uniform in both cell types. ChIP-seq experiments contain bins with substantial enrichment. "PHH"/"Hepa" = Primary Human Hepatocytes.

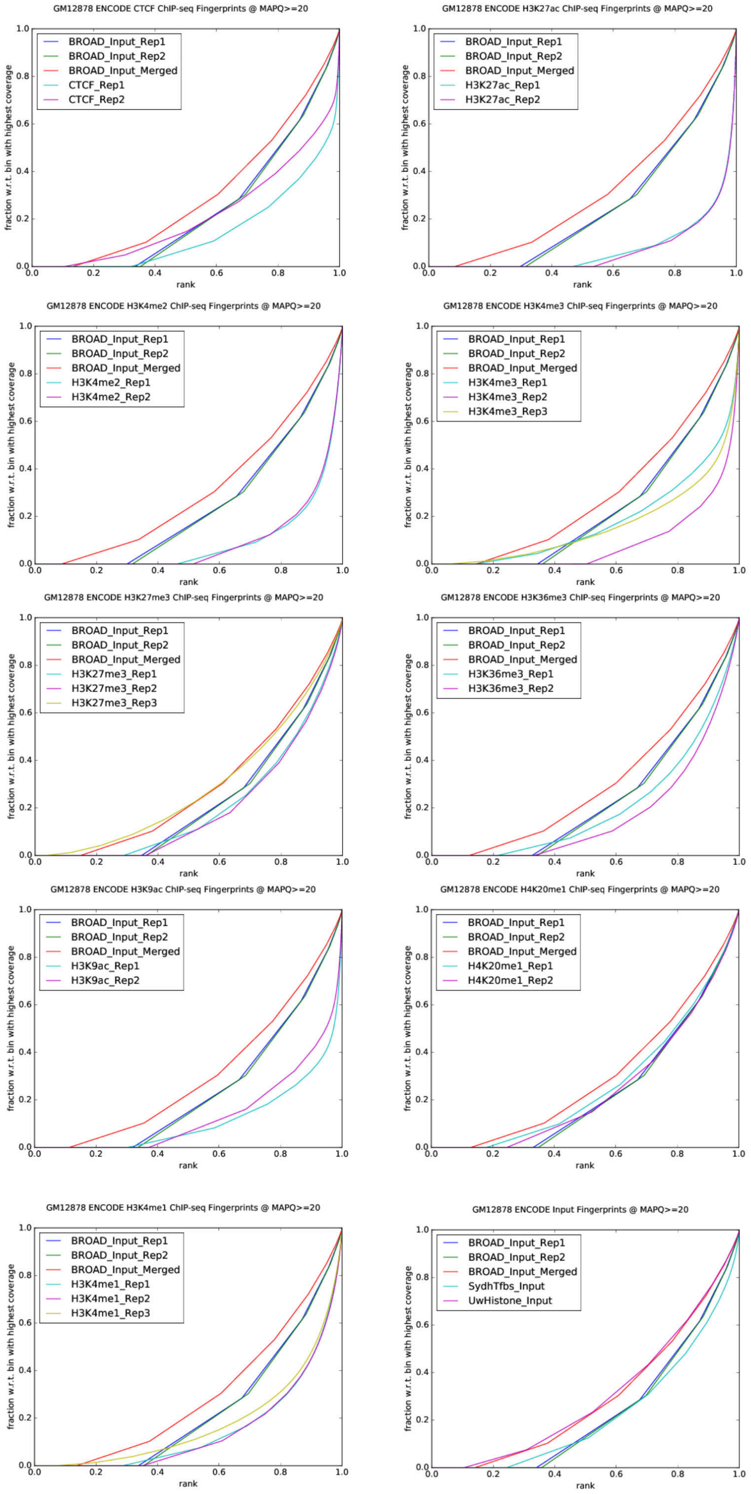

Fig. A.3 – ENCODE GM12878 ChIP-seq Data Quality Measured with BamFingerprint. Cumulative fraction of ChIP-seq, Input reads with respect to (w.r.t.) to bin with highes coverage in GM12878 cells computed with deepTools [110]. Input experiments are lack coverage in ∼30% of the genome. Pooling Input experiments resolves this issue. ChIP-seq replicates contain regions with substantial enrichment.

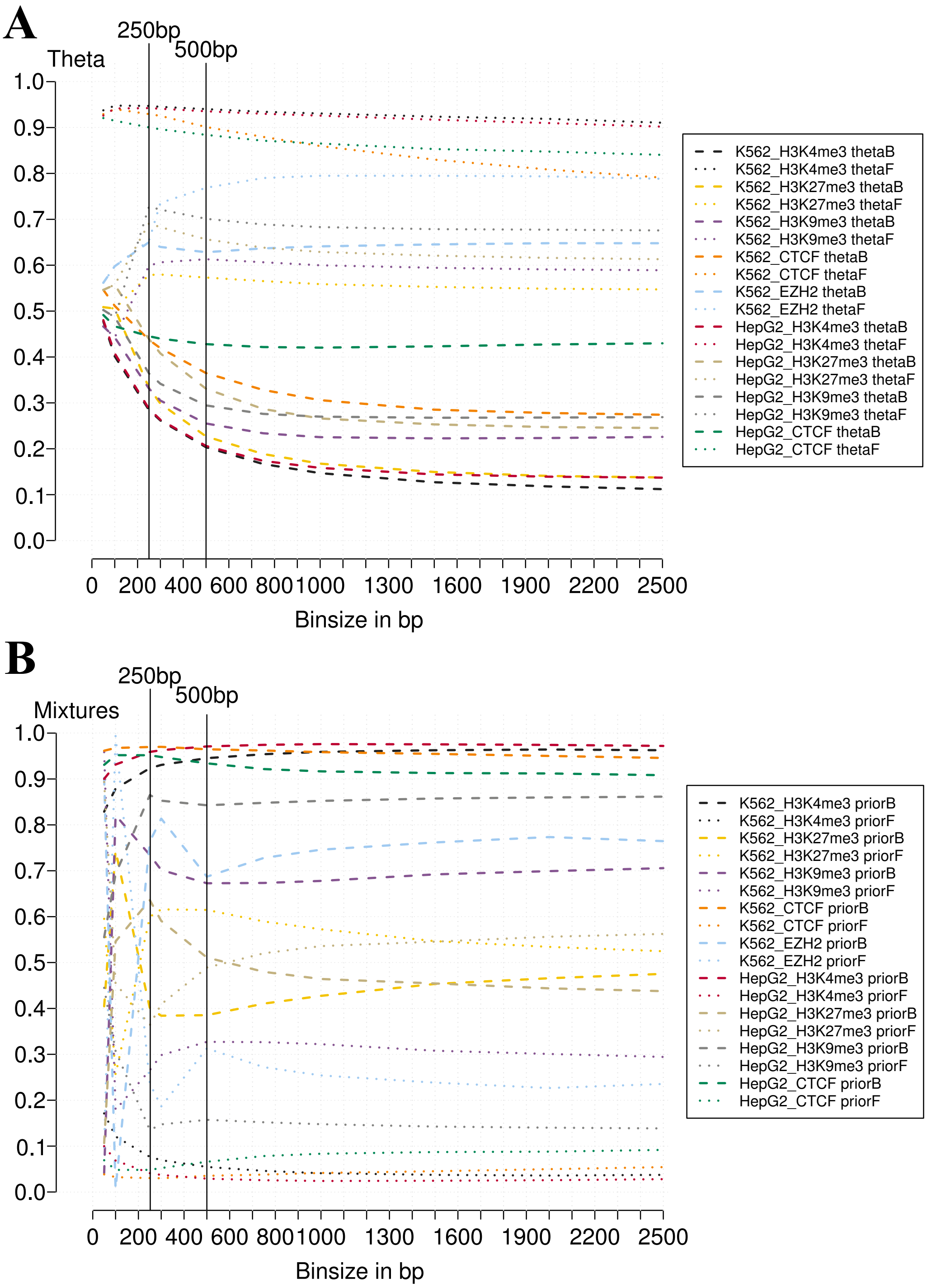

Fig. A.4 – The Dependency of the enrichR Model Fit on the Choosen Binsize. (A) Estimates for θ_E and θ_B are robust for a binsize≥500bp across all studied ChIP-seq experiments. (B) Estimates for π_E and π_B are robust for binsizes≥500bp for most ChIP-seq experiments.

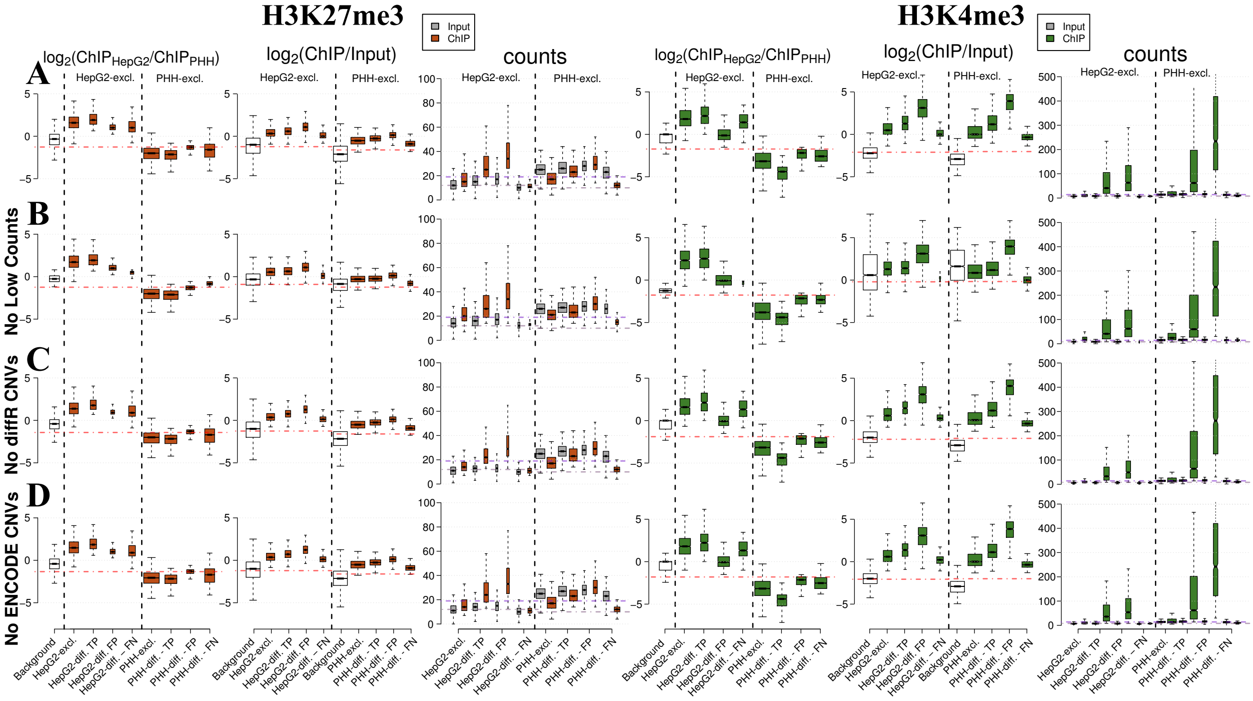

Fig. A.5 – Discrepancies between diffR and enrichR-compare are Due to Low Count Regions and Copy Number Variations Specific for HepG2 Cells. (A) (from left to right) Conditional $\log_2$ fold change, $\log_2$ fold change over respective Input and read counts for H3K27me3 and H3K4me3 in HepG2 cells and Primary Human Hepatocytes (PHH) for background, "cell type-exclusive" (enrichR-compare classified) and "cell type-differential" (diffR classified). "cell type-differential" is subclassified as true positives (TP), false positives (FP) and false negatives (FN) with respect to enrichR-compare classification. Conitional fold change and read counts in FP regions are equal or greater than the ones in TP regions suggesting genuine differences in ChIP-seq signal intensity between HepG2 cells and PHH. FN regions are characterized by low ChIP fold change over Input and low read counts. (B) Same as (A) with diffR's T Filter applied to enrichR-compare. The number of FN decreases substantially improving consistency between diffR and enrichR-compare (C) Same as (A) with diffR called CNVs removed. Most groups do not change but the number of FP in HepG2-differential calls is reduced. (D) Same as (A) with ENCODE reported CNVs in HepG2 cells removed with similar observations thatn for (C). See also Table 6.1. Whiskers extend up to 1.5-times the interquartile range. Width of boxes are propotional to the square-root of the number of regions in the groups.

Appendix B

Supplementary Tables

		1.5kb Promoter			Genes		
All regions		+	-	Odds, P-value	+	-	Odds, P-value
enrichR	enriched	47984	94467	18.29	113531	28920	3.627
	background	151829	5467798	P<2.2e-16	2920701	2698926	P<2.2e-16
MACS2	enriched	46627	79766	20.92	102294	24099	3.914
	background	153186	5482499	P<2.2e-16	2931938	2703747	P<2.2e-16
DFilter	enriched	42406	37229	39.97	65205	14430	4.13
	background	157407	5525036	P<2.2e-16	2969027	2713416	P<2.2e-16
CisGenome	enriched	37980	36547	35.48	62947	11580	4.969
	background	161833	5525718	P<2.2e-16	2971285	2716266	P<2.2e-16
SPP	enriched	46027	92768	17.65	112820	25975	4.017
	background	153786	5469497	P<2.2e-16	2921412	2701871	P<2.2e-16
BCP	enriched	47923	101178	17.03	120391	28710	3.884
	background	151890	5461087	P<2.2e-16	2913841	2699136	P<2.2e-16
MUSIC	enriched	41757	64910	22.37	87121	19546	4.096
	background	158056	5497355	P<2.2e-16	2947111	2708300	P<2.2e-16
Tool-exclusive		+	-	Odds, P-value	+	-	Odds, P-value
enrichR	enriched	1158	12206	2.651	8877	4487	1.781
	background	198655	5550059	P=5.203e-170	3025355	2723359	P=1.655e-228
MACS2	enriched	182	1878	2.699	1437	623	2.074
	background	199631	5560387	P=2.781e-29	3032795	2727223	P=3.96e-56
DFilter	enriched	302	726	11.6	613	415	1.328
	background	199511	5561539	P=4.879e-184	3033619	2727431	P=7.698e-06
CisGenome	enriched	0	0	0	0	0	0
	background	0	0	P=1	0	0	P=1
SPP	enriched	4287	35502	3.413	29147	10642	2.476
	background	195526	5526763	P<2.2e-16	3005085	2717204	P<2.2e-16
BCP	enriched	913	12387	2.057	10030	3270	2.763
	background	198900	5549878	P=1.973e-80	3024202	2724576	P<2.2e-16
MUSIC	enriched	69	840	2.287	630	279	2.03
	background	199744	5561425	P=3.051e-09	3033602	2727567	P=2.106e-24
Tool-specific		+	-	Odds, P-value	+	-	Odds, P-value
enrichR	enriched	7575	49310	4.405	41437	15448	2.431
	background	192238	5512955	P<2.2e-16	2992795	2712398	P<2.2e-16
MACS2	enriched	5884	34226	4.901	29718	10392	2.586
	background	193929	5528039	P<2.2e-16	3004514	2717454	P<2.2e-16
DFilter	enriched	3789	5490	19.57	5869	3410	1.548
	background	196024	5556775	P<2.2e-16	3028363	2724436	P=2.449e-94
CisGenome	enriched	104	299	9.687	262	141	1.671
	background	199709	5561966	P=1.649e-58	3033970	2727705	P=6.935e-07
SPP	enriched	9776	59092	4.79	52118	16750	2.829
	background	190037	5503173	P<2.2e-16	2982114	2711096	P<2.2e-16
BCP	enriched	7351	55770	3.771	48013	15108	2.887
	background	192462	5506495	P<2.2e-16	2986219	2712738	P<2.2e-16
MUSIC	enriched	2528	21498	3.303	17404	6622	2.371
	background	197285	5540767	P<2.2e-16	3016828	2721224	P<2.2e-16

Table B.1 – Transcriptional Start Site and Gene Overlap of H3K4me3 Enrichment Calls by enrichR, MACS2, DFilter, CisGenome, SPP, BCP and MUSIC. A 1.5kb promoter is defined as 750bp down- and upstream of the TSS. "+" = overlapping; "-" = non-overlapping; P-value obtained from Fisher's exact test.

		1.5kb Promoter			Genes		
All regions		+	-	Odds, P-value	+	-	Odds, P-value
enrichR	enriched	24045	535515	1.003	520070	39490	17.06
	background	99501	2221983	P=0.7132	1011345	1310139	P<2.2e-16
MACS2	enriched	18696	433224	0.9566	429676	22244	23.27
	background	104850	2324274	P=4.117e-08	1101739	1327385	P<2.2e-16
DFilter	enriched	4407	94917	1.038	97990	1334	69.09
	background	119139	2662581	P=0.01912	1433425	1348295	P<2.2e-16
CisGenome	enriched	1989	40528	1.097	42168	349	109.5
	background	121557	2716970	P=8.019e-05	1489247	1349280	P<2.2e-16
SPP	enriched	1912	23579	1.823	24308	1183	18.39
	background	121634	2733919	P=3.913e-118	1507107	1348446	P<2.2e-16
BCP	enriched	16596	390522	0.9405	388968	18150	24.98
	background	106950	2366976	P=4.402e-13	1142447	1331479	P<2.2e-16
MUSIC	enriched	17729	392437	1.01	394110	16056	28.78
	background	105817	2365061	P=0.244	1137305	1333573	P<2.2e-16
Tool-exclusive		+	-	Odds, P-value	+	-	Odds, P-value
enrichR	enriched	4684	88909	1.183	77393	16200	4.381
	background	118862	2668589	P=7.126e-27	1454022	1333429	P<2.2e-16
MACS2	enriched	296	7137	0.9255	6451	982	5.81
	background	123250	2750361	P=0.1971	1524964	1348647	P<2.2e-16
DFilter	enriched	0	0	0	0	0	0
	background	0	0	P=1	0	0	P=1
CisGenome	enriched	0	0	0	0	0	0
	background	0	0	P=1	0	0	P=1
SPP	enriched	227	479	10.59	571	135	3.728
	background	123319	2757019	P=2.794e-129	1530844	1349494	P=2.268e-53
BCP	enriched	221	2739	1.802	2530	430	5.192
	background	123325	2754759	P=7.351e-15	1528885	1349199	P=1.001e-305
MUSIC	enriched	1855	10161	4.122	11002	1014	9.622
	background	121691	2747337	P<2.2e-16	1520413	1348615	P<2.2e-16
Tool-specific		+	-	Odds, P-value	+	-	Odds, P-value
enrichR	enriched	19337	436221	0.9874	417716	37842	13
	background	104209	2321277	P=0.1145	1113699	1311787	P<2.2e-16
MACS2	enriched	13988	333927	0.9266	327319	20596	17.54
	background	109558	2423571	P=5.251e-17	1204096	1329033	P<2.2e-16
DFilter	enriched	3	24	2.79	26	1	22.91
	background	123543	2757474	P=0.1075	1531389	1349628	P=1.584e-06
CisGenome	enriched	1	25	0.8928	25	1	22.03
	background	123545	2757473	P=1	1531390	1349628	P=2.977e-06
SPP	enriched	750	2560	6.573	2670	640	3.681
	background	122796	2754938	P=2.147e-308	1528745	1348989	P=5.517e-240
BCP	enriched	11889	291235	0.9017	286622	16502	18.61
	background	111657	2466263	P=1.928e-26	1244793	1333127	P<2.2e-16
MUSIC	enriched	13022	293148	0.9905	291760	14410	21.81
	background	110524	2464350	P=0.3127	1239655	1335219	P<2.2e-16

Table B.2 – Transcriptional Start Site and Gene Overlap of H3K36me3 Enrichment Calls by enrichR, MACS2, DFilter, CisGenome, SPP, BCP and MUSIC. A “1.5kb Promoter” is defined as 750bp down- and upstream of the TSS. “+” = overlapping; “-” = non-overlapping; P-value obtained from Fisher’s exact test.

Abstract

Molecular Biology pertains to the molecular basis of the regulation of biomolecular processes in the cell, *e.g.* gene expression or the genome-wide localization of DNA-associated proteins. These molecular quantities are routinely measured by Next Generation Sequencing (NGS)-based techniques due to their genome-wide scalability and cost-efficiency. In order to discern background-regions from genomic loci that harbor a biological relevant signal, *i.e.* **difference calling**, the NGS measurements need to be corrected for technical biases with the help of a control, *i.e.* **normalization**. However, the normalization itself requires the knowledge of background regions and, consequently, difference calling and normalization are inseparable. Here, this problem is solved by the data-driven "normR" framework which models the inter-dependency of NGS measurements in background- and signal-regions as a multinomial sampling trial with a binomial mixture model. The robust normR normalization accounts for the effect of signal on the overall measurement statistic by modeling treatment and control simultaneously. In this thesis, I used normR in three studies concerning the inference of DNA-protein binding from ChIP-seq data. Firstly, the two-component "enrichR" model is shown to achieve a more sensitive enrichment calling (AUC$\geq$0.93) than six competitor methods (AUC$\leq$0.86) in low, *e.g.* H3K36me3, and high, *e.g.* H3K4me3, signal-to-noise ratio (S/N) ChIP-seq data. enrichR's enrichment calls augment the resolution and comprehensiveness of chromatin segmentations by chromHMM and its normalization improves on present *in silico* and *in vitro* ChIP-seq normalization methods. Secondly, the three-component "regimeR" model dissects enrichment into two unprecedented regimes of different signal levels. A regimeR-based analysis identified two distinct facultative and constitutive heterochromatic enrichment regimes in H3K27me3 and H3K9me3 ChIP-seq data, respectively. The identified *peak* regions (high enrichment) resemble nucleation sites for heterochromatin embedded in regions of *broad* (low) enrichment. Lastly, the three-component "diffR" model calls conditional differences in ChIP-seq enrichment between two conditions. The diffR calls in low (H3K27me3) and high (H3K4me3) S/N ChIP-seq data are confirmed by a systematic comparison to four difference callers. Overall, normR represents a robust and versatile framework for the comprehensive analysis of ChIP-seq data, yet, it can be readily applied to other NGS-based experiments like ATAC-seq, STARR-seq or RNA-seq.

Zusammenfassung

Die Molekulare Biologie studiert die molekulare Basis der Regulierung von biomolekularen Prozessen wie der Genexpression und der genomweiten Lokalisation von DNS-bindenden Proteinen. Die molekularen Größen werden mittels Next Generation Sequencing(NGS)-basierten Methoden gemessen, da diese genomweit skalierbar und kosteneffizient sind. Um Hintergrundregionen von genomischen Regionen mit einem biologisch relevanten Signal zu unterscheiden (**Differenzenbestimmung**) müssen technische Verzerrungen in den NGS Messungen mit Hilfe einer Kontrolle normalisiert werden. Jedoch benötigt eine korrekte **Normalisierung** die Identität der Hintergrundregionen und, somit, sind Differenzenbestimmung und Normalisierung untrennbar miteinander verbunden. Dieses Problem wird mit dem vorgestellten datenbasierten "normR" Modell gelöst, welches die Wechselbeziehung zwischen Zahlenwerten in Hintergrund- und Signalregionen als eine binomiale Mischverteilung modelliert. Die robuste Normalisierung von normR berücksichtigt durch gleichzeitige Modellierung von Experiment und Kontrolle den Einfluss des Signals auf die Messstatistik. In dieser Arbeit wurde normR in drei Analysen von ChIP-seq Daten verwendet um DNS-Bindestellen von Proteinen zu identifizieren. 1. Das "enrichR" Modell erreicht mit einer Mischverteilung aus zwei Komponenten eine Differenzenbestimmung, die sensitiver ist (AUC$\geq$0.93) als bei sechs anderen Programmen (AUC$\leq$0.86). Die identifizierten differentiellen Regionen erweitern die Auflösung und den Umfang von Chromatinsegmentierungen durch das chromHMM Programm. Die Normalisierung von enrichR ist besser als bekannte in vitro und in silico Normalisierungsansätze. 2. Das "regimeR" Modell mit drei Komponenten teilt die vom ChIP angereicherten Regionen in zwei Klassen mit unterschiedlicher Signalintensität. Eine Analyse mit regimeR identifiziert zwei Klassen von Anreicherung in fakultativem und konstitutivem Heterochromatin in H3K27me3 and H3K9me3 ChIP-seq Datensätzen. Die Regionen mit hoher Signalintensität sind flankiert von breiten Regionen mit niedrigem Signal und könnten Keimstellen des Heterochromatins darstellen. 3. Das "diffR" Modell identifiziert Unterschiede zwischen ChIP-seq Messungen in zwei zellulären Bedingungen. Die Ergebnisse von diffR wurden mittels eines systematischen Vergleichs zu vier anderen ChIP-seq Differenzbestimmungsprogrammen validiert. normR ist ein robustes und vielseitiges Programm zur umfassenden Analyse von ChIP-seq Daten und vermag in Zukunft eine sensitive Analyse von anderen NGS Datensätzen wie ATAC-seq, STARR-seq und RNA-seq zu ermöglichen.

Selbstständigkeitserklärung

Hiermit erläre ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsquellen und Quellen verwendet habe. Ich erkläre weiterhin, dass ich die vorliegende Arbeit oder deren Inhalt nicht in einem früheren Promotionsverfahren eingereicht habe.

Johannes Helmuth, Berlin, den 28. Juni 2017