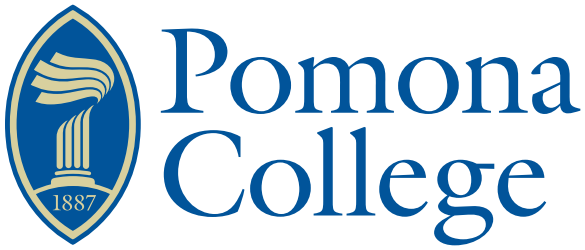

Senior Thesis in Mathematics

Selecting ChIP-Seq Normalization Methods from the Perspective of their Technical Conditions

Author:
Sara Colando

Advisor:
Dr. Johanna Hardin

Submitted to Pomona College in Partial Fulfillment
of the Degree of Bachelor of Arts

April 27, 2024

Acknowledgements

First and foremost, I would like to thank Professor Hardin for her guidance and support on this thesis and throughout my time at Pomona. Working alongside her on various research projects over the past three years has deepened my passion for statistics and has ultimately inspired me to continue to pursue statistics after graduating from Pomona. I am also thankful to Professor Schulz at Harvey Mudd College for providing helpful insights on the biological basis of ChIP-Seq and to Viet Pham from Pomona College's ITS department for helping us run our ChIP-Seq read count simulations on Pomona's High-Performance Computing system. Finally, I want to extend my appreciation to Julie Tannenbaum and Gabbrielle Johnson, as well as my family, coaches, and friends, for their unwavering support during my time at Pomona. They all have played a significant role in shaping me, and I feel incredibly grateful to have each of them in my life.

Abstract

Within the last two decades, high throughput sequencing has become one of the most popular methods for data generation within genomics, epigenomics, and transcriptomics (Lee, 2023). A popular method of high throughput sequencing is chromatin immunoprecipitation followed by high throughput sequencing, or ChIP-Seq. ChIP-Seq data provides vital insights into locations on the genome where there are differences in DNA occupancy between experimental states (i.e., differential DNA occupancy) (Wu et al., 2015). However, since ChIP-Seq data is collected experimentally, it must be normalized to assess which genomic regions have differential DNA occupancy. While normalization is an essential step in identifying genomic regions with differential DNA occupancy, the primary technical conditions underlying ChIP-Seq normalization methods have yet to be meticulously examined in the academic literature. In this thesis, we identify three primary technical conditions underlying ChIP-Seq between-sample normalization methods: (1) Symmetric Differential DNA Occupancy, (2) Equal Amount of Total DNA Occupancy, and (3) Equal Amount of Total Background Binding. We then categorize popular ChIP-Seq normalization methods based on the technical conditions they use to normalize between experimental states. A major contribution of this thesis is our ChIP-Seq read count simulation results, which validate our categorization of the normalization methods by their technical conditions. We conclude by underscoring how our findings demonstrate that not all normalization methods are equally effective on all kinds of ChIP-Seq data. Instead, our results emphasize that biologists should use their understanding of the ChIP-Seq experiment at hand to guide their choice of normalization method.

Contents

1 Introduction **1**

2 ChIP-Seq and Differential Binding Analysis **5**
2.1 ChIP-Seq Data Collection 5
2.2 Data Analysis for Differential Binding 7

3 Importance of Normalization Methods' Technical Conditions **12**
3.1 Toy Example 12

4 Normalization Methods and their Technical Conditions **17**
4.1 Normalization by Library Size 19
4.2 Normalization by Distribution 23
4.3 Normalization by Background 26
4.4 Normalization by Controls 29

5 Simulation Preliminaries **31**
5.1 The Oracle Normalization Method 32
5.2 Simulation Conditions 36
5.3 Simulation Metrics 40

6 Simulation Results **42**
6.1 General Simulation Results 42
6.2 Simulation Results: All Technical Conditions Met 46
6.3 Effect of Asymmetric Differential DNA Occupancy 47
6.4 Effect of Different DNA Occupancy 49
6.5 Effect of Different Background Binding 50

7 Conclusion **52**

8 Supplementary Materials **54**
8.1 ChIP-Seq Simulation Code 54
8.2 Confidence Intervals for Simulation Results 55
8.3 The Benjamini-Hochberg Method 57

List of Figures

2.1 ChIP-Seq Data Collection 6
2.2 Peak-Calling Illustration 8
2.3 Consensus Peaks Illustration 9

3.1 Toy Example: Amount of DNA Occupancy per Cell and Expected Proportional Shares of DNA Binding 13
3.2 Toy Example: Reads Aligned to Each Peak 14
3.3 Toy Example: Comparing Normalized Read Counts and Fold Changes 16

4.1 Observational Unit for Reads in Peaks Normalization 18
4.2 Observational Unit for Background Bin Normalization 18
4.3 Normalization by Library Size Illustration 23
4.4 Spike-in Normalization Illustration 30

5.1 Oracle Normalization Method Illustration 34
5.2 Simulation Condition Illustration 39

6.1 Simulation Results: Average False Discovery Rate 43
6.2 Simulation results: Average Power 45
6.3 Simulation Results: All Technical Conditions Met 46
6.4 Simulation Results: Effect of Asymmetric Differential DNA Occupancy 47
6.5 Simulation Results: Effect of Differences in DNA Occupancy 49
6.6 Simulation Results: Effect of Differences in Background Binding . . . 50

8.1 Simulation Results: 95% Confidence Intervals for the Average False Discovery Rate 55
8.2 Simulation Results: 95% Confidence Intervals for the Average Power . 56

Chapter 1

Introduction

Within the last two decades, high throughput sequencing has become one of the most popular methods for data generation within genomics, epigenomics, and transcriptomics (Lee, 2023). One popular method of high throughput sequencing is chromatin immunoprecipitation, followed by high-throughput sequencing, which is commonly referred to as ChIP-Seq. ChIP-Seq aims to characterize the occupancy behavior of a protein of interest (e.g., a transcription factor, histone modification, etc.) on the genome or DNA (Wu et al., 2015).[1]

In ChIP-Seq experiments, *peaks* (i.e., genomic regions enriched with DNA binding) are typically the unit of interest because ChIP-Seq experiments are often conducted to assess whether a certain genomic region is *truly* occupied by the protein of interest in the experimental state. In this thesis, we refer to *DNA occupancy* as the population parameter that we aim to estimate through ChIP-Seq experiments and refer to the sample estimate of DNA occupancy as the *DNA binding*. Often, biologists care not only about DNA occupancy but also about whether there is *differential DNA occupancy* between experimental states. That is, whether the *true* amount of DNA occupied by the protein of interest is different between experimental states for a given peak. Differential DNA occupancy is estimated via *differential binding*. A peak is considered *differentially bound* if there is a *statistically significant* difference in the amount of DNA binding between experimental states (Nakato and Sakata, 2020). The amount of DNA binding in a given genomic region is found by counting the number of aligned *reads* (i.e., sequenced transcripts) that are associated with the region. Genomic regions with higher read counts (i.e., more DNA binding) are expected to have a larger amount of DNA occupancy (Nakato and Sakata, 2020).[2]

Since ChIP-Seq data is collected experimentally, there is expected to be a difference in the read counts (and thus the observed DNA-protein binding) across experi-

[1] A detailed description of ChIP-Seq data collection is provided in Chapter 2.1.

[2] In ChIP-Seq literature, the population parameter and its estimate are both referred to as *differential DNA binding*. However, we use different terms for the parameter and estimate to make it clearer whether we were referring to the parameter (i.e., differential DNA occupancy) or the estimate (i.e., differential DNA binding) in this thesis.

mental states, regardless of whether or not there are differences in DNA occupancy. In turn, statistical models (which rely on technical conditions about the underlying structure of the true DNA-protein binding) allow us to perform hypothesis tests in order to detect statistically significant differences in the observed DNA binding across experimental states. The process of hypothesis testing to assess for differences in DNA binding across different experimental states is referred to as *differential binding analysis* (Nakato and Sakata, 2020).

However, using raw read counts across experimental states to conduct differential binding analysis is fallible for at least two reasons. First, the total reads aligned to a given peak is commonly considered a random variable (Anders and Huber, 2010). So, even if a peak has different read counts across experimental states, it could have the same amount of DNA occupancy across those experimental states. Second, non-random events can influence the number of reads aligned to a peak. Non-random events include experimental artifacts like changes in the amount of DNA loaded or the quality of the antibody used.[3] Such experimental artifacts influence the sequencing (or read) depth (i.e., the total number of reads across the entire sample) (Nakato and Sakata, 2020). Thus, differences in read counts in a given peak can arise between experimental states even if there is no difference in the DNA occupancy between the experimental states. In an attempt to ameliorate the influence of random variability and experimental artifacts on which genomic regions are classified as differentially bound and thus presumed to have different DNA occupancy between the experimental states, the raw read counts must be normalized between the samples prior to performing differential binding analysis (Steinhauser et al., 2016).

Before the read counts can be normalized between experimental states, consensus peaks must first be identified from the collected ChIP-seq data. The process of identifying peaks is commonly called *peak-calling*. *Peak-calling* aims to distinguish background regions where there is DNA occupancy from regions where there is not (Nakato and Sakata, 2020). The peaks found via peak-calling are then consolidated into a consensus peak set (Stark and Brown, 2011). After identifying the consensus peaks, a *read count matrix* is generated. In the matrix, each entry is the raw read count for a consensus peak across the different experimental states (Stark and Brown, 2011). These read counts are then normalized between samples, and differential binding analysis is performed using the normalized read counts (Nakato and Sakata, 2020).

There are various ChIP-Seq normalization methods available to biologists, and we will examine several popular normalization methods (e.g., Library Size (Dillies et al., 2012), TMM (Robinson and Oshlack, 2010), and RLE (Love et al., 2014)) in this thesis. The normalization methods we consider in this thesis are all compatible with `DiffBind`, an `R` package that performs the entire pipeline of ChIP-seq data analysis, from creating the consensus peak set across experimental states to performing differential binding analysis. However, while there are various normalization methods that

[3]Thank you to Professor Schulz at Harvey Mudd College for emphasizing this point.

biologists can employ, there is a dearth of literature specifying which normalization method should be selected within a given biological context. A major contribution of this thesis is to identify the different technical conditions of ChIP-Seq normalization methods in order to provide a framework for selecting between different normalization methods. Through our work, we identify three key technical conditions: (1) **symmetric differential DNA occupancy**, (2) **equal amount of total DNA occupancy**, and (3) **equal amount of total background binding**. Here is a brief overview of what each technical condition entails:

1. **Symmetric Differential DNA Occupancy**: there is roughly symmetric differential DNA occupancy across experimental states. That is, the number of peaks with more DNA occupancy in one experimental state is equal to the number of peaks with more DNA occupancy in the other experimental state.

2. **Equal Amount of Total DNA Occupancy**: the amount of total DNA occupancy is the same across the two experimental states. That is, each experimental state has the same amount of DNA occupancy per cell.

3. **Equal Amount of Total Background Binding**: the number of rogue background DNA binding is the same across experimental states.

While the role of normalization methods has been analyzed through the lens of their technical conditions for RNA-seq, which is another popular technique for high throughput sequencing (e.g., see Evans et al. (2016)), there is currently no analogous analysis of ChIP-Seq between-sample normalization. However, ChIP-Seq data is used to answer different types of biological questions than RNA-Seq data. RNA-Seq focuses on characterizing *true* (differential) gene expression, whereas ChIP-Seq focuses on characterizing (differential) DNA occupancy. In virtue of asking different questions, RNA-Seq and ChIP-Seq data also have different underlying structures. For one, unlike in RNA-Seq, there are no pre-defined genomic regions of interest in ChIP-Seq. Instead, the genomic regions of interest are defined through the process of peak-calling. Thus, there is an added layer of variability in ChIP-Seq data that is not present in RNA-Seq data. Therefore, we cannot directly apply the results from RNA-Seq analyses to ChIP-Seq normalization methods. Rather, ChIP-Seq normalization methods must be explicitly analyzed through the lens of their own technical conditions.

In this thesis, we identify three technical conditions underlying different between-sample ChIP-Seq normalization methods and demonstrate how violating the technical conditions can influence the biological conclusions about which peaks have differential DNA occupancy between experimental states. We first provide a biological primer on ChIP-Seq and its data analysis workflow for differential binding analysis in Chapter 2. We then illustrate why normalization is necessary as well as why satisfying the technical conditions is crucial for accurate downstream differential binding analysis in Chapter 3. Next, we categorize popular ChIP-Seq normalization methods by their

main technical conditions in Chapter 4. To examine how violating some combination of our primary technical conditions impacts the performance of different normalization methods, we simulate ChIP-Seq read count data. In particular, we investigate how violating some combination of the primary technical conditions affects the average false discovery rate, power, and absolute size factor ratios associated with various ChIP-Seq normalization methods as we vary the proportion of peaks with differential DNA occupancy. We explicate our ChIP-Seq read count simulation and associated metrics in Chapter 5 and present our simulation results in Chapter 6. We end with a discussion of how this thesis' findings underscore the importance of understanding the underlying structure of ChIP-Seq data when choosing a normalization method in order to draw meaningful biological conclusions.

Chapter 2

ChIP-Seq and Differential Binding Analysis

2.1 ChIP-Seq Data Collection

One purpose of ChIP-Seq experiments is to identify genomic regions that have differential DNA occupancy across experiment states. We consider a genomic region to have *differential DNA occupancy* across experimental states if there is a difference (on average) in the amount of DNA occupied per cell by a protein in that region across the different experimental states (Nakato and Sakata, 2020). ChIP-Seq helps identify genomic regions with differential DNA occupancy by generating *read count* data, which is meant to approximate the degree to which a protein of interest occupies a genomic region. In this thesis, we distinguish differential DNA occupancy from another important term: *differential DNA binding*. We consider a genomic region to be *differentially bound* if there is a statistically significant difference in the aligned read counts (i.e., amount of DNA binding) across the different experimental states. The following high-level procedure is done to generate ChIP-Seq data (Nakato and Sakata, 2020; Park, 2009):

Step (1): Immunoprecipitate the DNA-protein fragments with a specific antibody

Step (2): Purify the DNA

Step (3): Map the purified DNA (i.e., reads) to a reference genome

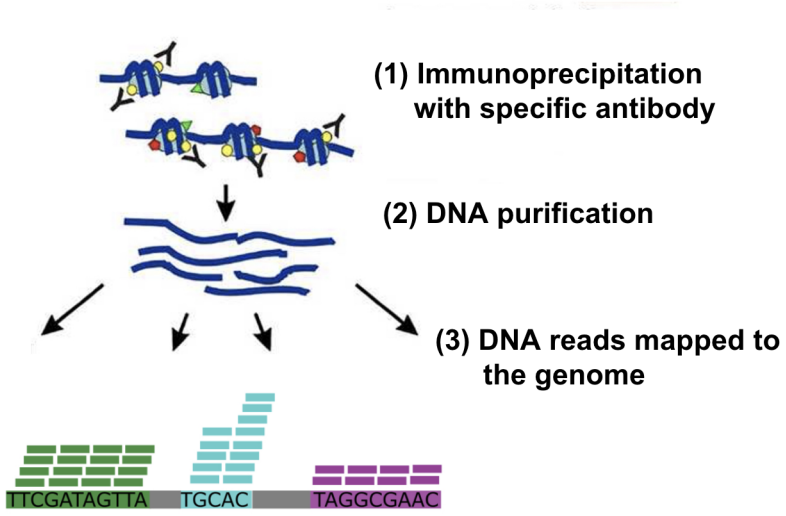

Figure 2.1: **A high-level overview of ChIP-Seq data collection.** In Step (1), the DNA-protein fragments are immunoprecipitated with a specific antibody that is known a priori to bind with the protein of interest. In Step (2), the DNA is purified, meaning that the antibody and protein of interest are removed from the DNA. In Step (3), the purified DNA fragments (which are referred to as "reads") are mapped to a reference genome. Components of this illustration were taken from Figure 1 in Lodhi and Tulin (2011).

Immunoprecipitation with a specific antibody (i.e., Step (1)) is a key step of the ChIP-Seq data collection process. In this step, an antibody that is known (or theorized) a priori to bind with the protein of interest is used to pick out the DNA fragments that the protein of interest occupies. Ideally, after immunoprecipitation, only the DNA fragments that were occupied by the protein of interest remain. These remaining DNA fragments are then purified (i.e., released from the protein of interest and antibody) in Step (2) (Park, 2009). After the DNA has been purified, the DNA fragments are mapped to a reference genome based on where the reference genome's and DNA fragments' nucleotide sequences match one another during Step (3). Each mapped DNA fragment is considered a *read*.

The data collection process is then repeated for each replicate across each experimental state. A *replicate* refers to a specific ChIP-Seq sample within a given experimental state (Höllbacher et al., 2020). Meanwhile, *experimental state* is broadly defined as the features of the experiment, which the researchers explicitly vary to analyze how the change affects the DNA occupancy of a specific protein of interest. For example, Kotipalli et al. (2021) used ChIP-Seq data to investigate how the H3K4me3 histone modification's (a type of protein's) behavior varied between normal-like and breast cancer cell lines (Kotipalli et al., 2021). For Kotipalli et al. (2021), one ex-

perimental state would be the normal-like cell lines, and the other experimental state would be the breast cancer cell lines, as this is what they are explicitly changing in their experiment to see its effects on DNA occupancy for the H3K4me3 histone modification. Meanwhile, the replicates are all the ChIP-Seq samples that are collected under the same experimental state (i.e., in normal-like cell lines or breast cancer cell lines).

2.2 Data Analysis for Differential Binding

After ChIP-Seq data collection is completed, downstream analysis can be conducted on the aligned DNA reads. A popular method of data analysis of ChIP-Seq read count data is differential binding analysis, which estimates differential DNA occupancy for a given protein of interest. Usually, the downstream analysis of ChIP-Seq data for differential binding involves the following workflow (Stark and Brown, 2011):

Step (4): Perform Peak-calling within each replicate

Step (5): Identify the consensus peak set within each experimental state

Step (6): Identify the consensus peak set across the experimental states

Step (7): Normalize across the experimental states[1]

Step (8): Perform differential binding analysis across the experimental states

Typically, the first step of ChIP-Seq data analysis is calling the peaks within each replicate (i.e., Step (4)), which is illustrated in Figure 2.2. *Peaks* are genomic regions that are significantly enriched with aligned reads in the replicate (Bailey et al., 2013). One common peak-calling software for ChIP-Seq is MACS (i.e., Model-based Analysis of ChIP-Seq) (Zhang et al., 2008). In MACS, peak-calling is done using hypothesis testing, where the expected number of reads in a genomic region is compared to the actual number of reads in that region. For MACS, the expected number of reads is calculated dynamically via a local Poisson distribution, which aims to capture any local biases in the replicate (Zhang et al., 2008). If the p-value for the region is below a pre-specified cut-off (which is 10^{-5} by default), then it is classified as a peak.[2] Any genomic region that is not called as a peak is then considered *background*.

[1]This step is usually referred to as *between-sample normalization*. Note, too, that sometimes there is another step between 6 and 7 where read counts are normalized within a given replicate rather than across experimental states. This process of normalization within a replicate is generally referred to as *within-sample normalization*.

[2]In newer versions of MACS, like MACS2 (which we use in our simulation), the p-values are adjusted for multiple comparisons using the Benjamini-Hochberg method. We describe the Benjamini-Hochberg method in more detail in Section 8. Thank you to the Harvard Biostatistics Core (HBC) for clarifying this point in their online training: `https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html`.

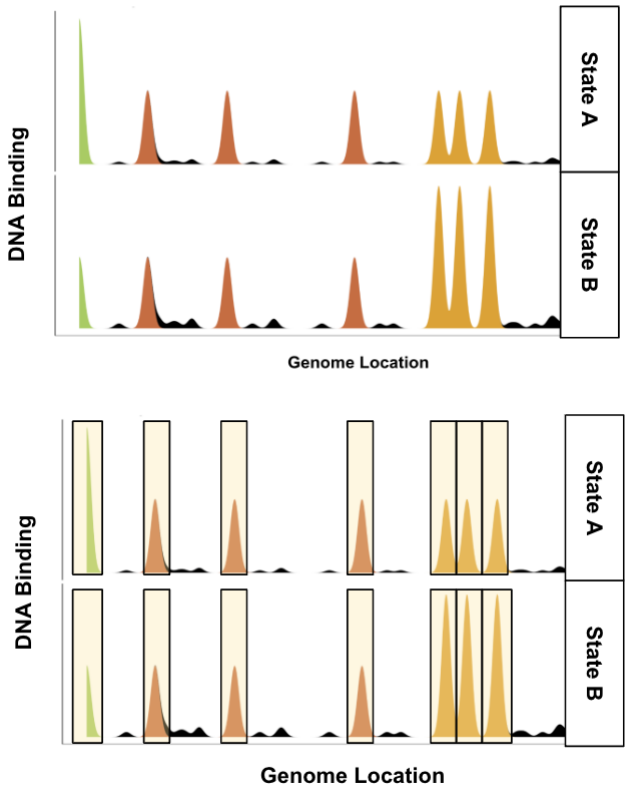

Figure 2.2: **Illustration of Step 4: peak-calling within a replicate.** The horizontal axis represents the genome location, and the vertical axis represents the number of aligned reads (i.e., DNA binding) to that specific genome location. The top plot is an illustration of what ChIP-Seq data looks like after being mapped to the genome. The bottom plot is an illustration of the peak-calling process with the ChIP-Seq replicate. The yellow-highlighted regions with the black box around them denote regions that would likely be called as peaks. All the genomic regions outside the yellow-highlighted boxes would likely be considered background.

After peak-calling has been conducted for each replicate, the next step is to generate a consensus peak set within each experimental state (i.e., Step (5)). *Consensus peaks within states* are regions that are called as peaks in a user-specified number of replicates within a given experimental state (Stark and Brown, 2011). This number (along with how many bases in different replicate peaks must be shared for it to be classified as the same peak) is determined based on how conservative the user wants to be about what to consider as a potential differentially bound region in Step (8). After the consensus peak set within each condition has been generated, the consensus peak set across experimental states is generated in Step (6). *Consensus peaks across states* are regions that are called as consensus peaks within conditions in a

user-specified number of the experimental states (Stark and Brown, 2011). In this thesis, we only consider cases where there are two unique experimental states. So, the consensus peaks across states will always be the union of the consensus peaks within each state in our thesis.[3]. Figure 2.3 provides an illustration of what would be considered consensus peaks within and across two unique experimental states.

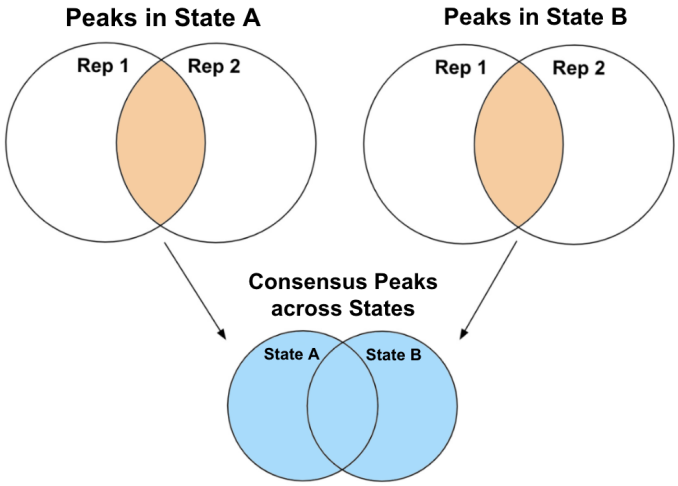

Figure 2.3: **Illustration of consensus peaks within and across experimental states.** In the illustration, we have two experimental states (A and B) and two replicates (1 and 2) in each experimental state. For the top layer of the figure, each circle denotes the specific peak set in a specific replicate. The orange regions (i.e., the intersection between replicates 1 and 2 for experimental states A and B) would be the consensus peak set *within* experimental states if we require two replicates to share a peak for it to be part of consensus peak set *within* each experimental state. In the bottom layer of the figure, we have the consensus peaks across the experimental states. The blue region (i.e., the union between the consensus peak set within conditions A and B) denotes the consensus peak set *across* the experimental states if we only require the peak to be a consensus peak set within one experimental state for it to be counted as part of the consensus peak set across the experimental states.

[3]Usually, even when there are more than two experimental states, the consensus peak set across states is defined to be the union of the consensus peaks within each experimental state so that the peaks that appear in one experimental state but not the others can still be candidates for differential DNA binding in downstream analysis. For more on this, see the following Biostars thread: `https://support.bioconductor.org/p/89517/`.

2.2.1 Between-Sample Normalization

The focus of our thesis is on ChIP-Seq normalization between experimental states, which is Step (7) in data analysis for differential binding. Paradigmatically, between-sample normalization involves calculating and then applying a size factor (which is also sometimes called a scaling factor) to the *raw* ChIP-Seq read counts associated with each peak in order to generate *normalized* read counts for the given peak. That is, the normalized read count for peak i in sample j is computed as follows:

$$(\text{Normalized Read Count})_{ij} = \frac{(\text{Raw Read Count})_{ij}}{s_j} \tag{2.1}$$

where s_j is the size factor associated with sample j.

The main goal of applying a size factor to the raw read counts for a given sample is to reduce the differences in read counts (i.e., DNA binding) between samples that result from the experimental nature of ChIP-Seq data collection rather than differences in DNA occupancy (Wu et al., 2015). Experimental artifacts that can arise in ChIP-Seq data include features such as the amount of chromatin loaded into the sequencer, the quality of the antibody, how long the replicate is in the sequencer, etc., that can influence the sequencing (or read) depth (i.e., the total number of aligned reads) for a given replicate (Nakato and Sakata, 2020). Crucially, however, ChIP-Seq between-sample normalization methods have technical conditions. These technical conditions rely on an expected ChIP-Seq data structure in order to ascertain the *expected* difference in the DNA binding in a given peak that is due to differences in DNA occupancy between the experimental conditions and normalize the ChIP-Seq data accordingly. We further explain the importance of satisfying normalization technical conditions in Chapter 3. We then categorize popular normalization methods by their technical conditions in Chapter 4. Our categorization is confirmed by our simulation results, which we present in Chapter 6

2.2.2 Differential Binding Analysis

After between-sample normalization has been performed, the normalized read counts associated with each peak in the consensus peak set across experimental states are then compared via a process called *differential binding analysis* in Step (8). Like peak-calling, *differential binding analysis* usually involves hypothesis testing to determine whether the observed difference in read counts between the samples is significant. That is, whether the observed difference in read counts across the experimental states is greater than we would expect it to be if the difference were just due to random chance (Love et al., 2014).

In our simulation, we use the `R` package `DiffBind` available through `Bioconductor` to conduct the entirety of the data analysis workflow for differential binding from Step

(4) to Step (8). By default, `DiffBind` uses DESeq2 for differential binding analysis, which was developed specifically for sequence count data such as ChIP-Seq (Love et al., 2014). In DESeq2, the expected read count for each peak region is found using a negative binomial distribution (Love et al., 2014). The p-values associated with the peaks in the consensus peak set across experimental states are then adjusted using the Benjamini-Hochberg method to control the false discovery rate for multiple comparisons at the pre-specified cut-off of 0.05 (the procedure for adjusted p-value using the Benjamini-Hochberg method is described in Chapter 8). As we discuss more in Section 5.3, in the context of ChIP-Seq differential binding analysis, the false discovery rate is the proportion of peaks that are classified as differentially bound that do *not* have differential DNA occupancy across the experimental states.

Chapter 3

Importance of Normalization Methods' Technical Conditions

Normalizing between experimental states is an essential step of data analysis for differential binding. In this chapter, we use a toy example to demonstrate how ChIP-Seq normalization methods have technical conditions and to motivate why satisfying a normalization method's technical conditions is necessary to draw meaningful biological conclusions about differential DNA occupancy.

3.1 Toy Example

Imagine that there are two experimental states, which we denote as A and B. Each experimental state only has one replicate. Moreover, three genomic regions are classified as peaks; they are denoted as Peak 1, Peak 2, and Peak 3. Figure 3.1 illustrates the total amount of DNA occupancy per cell in each of the peaks and experimental states in this toy example. In our toy example, like in many ChIP-Seq experiments, we are interested in determining which peaks have differential DNA occupancy, i.e., have a difference in the amount of DNA occupied by the protein across experimental states A and B.

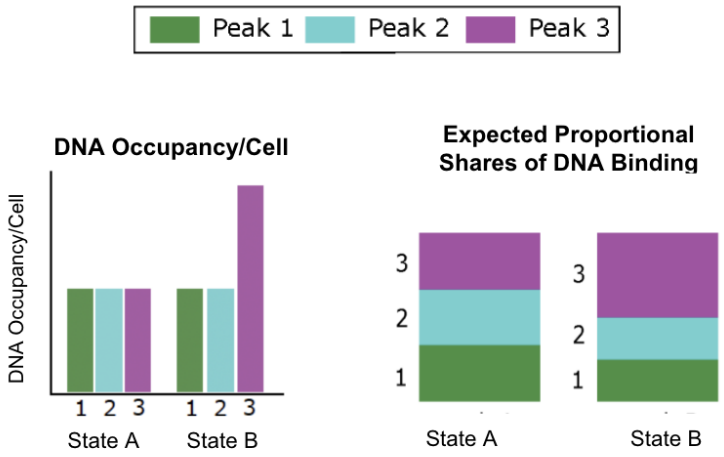

Figure 3.1: **Toy example: amount of DNA occupancy per cell and expected proportional shares of DNA binding.** In the left plot, the vertical axis denotes the total amount of DNA occupancy per cell, and the horizontal axis denotes the peak and experimental state. The right plot is the proportional shared of DNA occupancy, i.e., the expected proportional shares of DNA binding for each peak across experimental states A and B.

As depicted in the left plot of Figure 3.1, only Peak 3 has differential DNA occupancy; there is a fold change of 1/2 in the total amount of DNA occupancy associated with Peak 3 between experimental states A and B. Furthermore, it is important to note that there is a different amount of total DNA occupancy (per cell) between experimental states A and B in the toy example. Indeed, Figure 3.1 shows that State B has a higher total amount of DNA occupancy (per cell) than State A.

The right plot in Figure 3.1 breaks down the proportional shares of DNA occupancy for each peak across the experimental states, i.e., the *expected* proportional shares of DNA binding for each peak across the experimental states. In virtue of there being a difference in the total amount of DNA occupancy (per cell) across the experimental states, the fold change in the proportional shares of DNA occupancy for each peak does not match the true fold change in the total amount of DNA occupancy for each peak. For example, Peak 1 has a fold change of 4/3 in the proportional share of DNA occupancy between experimental states A and B, even though the true fold change in the amount of total DNA occupancy between experimental states A and B is 1 for Peak 1. Likewise, even though there is a true fold change of 1/2 in the total amount of DNA occupancy associated with Peak 3 between experimental states A and B, Peak 3 only has a fold change of 2/3 in the proportional share of DNA occupancy between experimental states A and B. Thus, in our toy example, the difference in the expected proportional shared of DNA binding between experimental states is not a good proxy for the differences in the total amount of DNA occupancy

per cell between the experimental states.

However, in real-life experiments, researchers do not know the actual amount of DNA occupancy (per cell) and how it differs between experimental states. Rather, the amount of DNA occupancy per cell would be approximated through the number of aligned reads to a given peak in an experimental state. Figure 3.2 provides the reads aligned to each peak in our toy example.[1] Note that, experimentally, the total number of aligned reads is the same across the experimental states. There is no reason that the total DNA occupancy would be the same as the total binding, given that DNA binding is a random variable and impacted by experiment-specific features (e.g., length of time in the sequencer and quality of the used antibody).

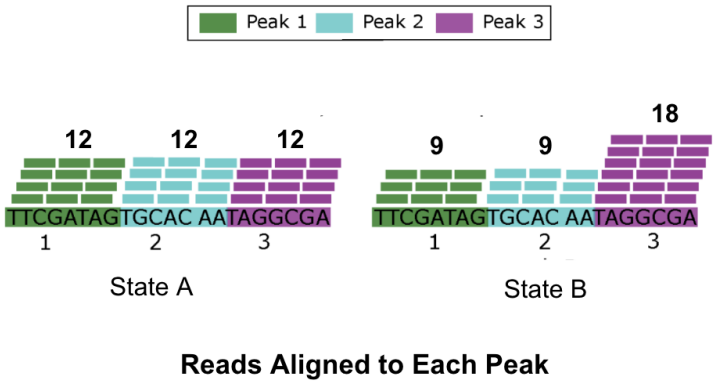

Figure 3.2: **Toy example: reads aligned to each peak.** The numbers above each peak represent the total number of reads aligned to that peak in the given experimental state.

Suppose we conducted differential binding analysis on the raw read counts given in Figure 3.2. The raw read count for each peak is different between experimental states A and B. Thus, it is likely that conducting differential binding analysis on the raw reads given in Figure 3.2 would classify all of Peaks 1, 2, and 3 as differentially bound. Moreover, Peak 3 would have a fold change of 12/18, or 2/3, in the amount of DNA binding between experimental states A and B. However, this misaligns with Peak 3's fold change in DNA occupancy between experimental states A and B, which is 1/2 (see Figure 3.1). Therefore, normalizing the read counts between experimental states is essential to draw accurate biological conclusions about differential DNA occupancy from the read count data.

However, not every normalization method will be effective on our toy read count data depicted in Figure 3.2. For example, suppose we normalize the read count data

[1]Note that in actual ChIP-Seq experiments, the peaks are separated (like in Figure 2.3) rather than right new to each other as we depict in our toy example in Figure 3.2.

with Library Size (Reads in Peaks).[2] Library Size (Reads in Peaks) uses the total number of reads aligned to the peaks as the size factor for the sample (see Equation (4.2)). The read count for each peak in the sample is then divided by this size factor s_j to create its normalized read count (see Equation (2.1)). Thus, the normalized read count for each peak represents its proportional share of actual DNA binding relative to the other peaks in the experimental state under Library Size (Reads in Peaks). Noting that the total number of reads is 36 in both experimental states in our toy example (see Figure 3.2), we have that $s_A = s_B = 36$. Using the raw read counts in Figure 3.2 and these size factors, we calculate the normalized read counts and estimated fold change for Library Size (Reads in Peaks) in Table 3.1.

	Norm. read count in State A	Norm. read count in State B	Est. fold change based on norm. read count	True fold change based on DNA occupancy
Peak 1	1/3	1/4	4/3	1
Peak 2	1/3	1/4	4/3	1
Peak 3	1/3	1/2	2/3	1/2

Table 3.1: **Library Size (Reads in Peaks) normalized read counts and fold changes.** The columns correspond to the experimental state, estimated fold change (based on the normalized read counts), and the true fold change based on the DNA occupancy (per cell) depicted in 3.1. The rows correspond to the peak number. The normalized read count for the specific peak and experimental state combination is found by dividing the raw read count by the size factor. The estimated fold change is then calculated by dividing the normalized read count in experimental state A by the normalized read count in experimental state B.

By dividing each raw read count by the total number of reads in the sample, Library Size (Reads in Peaks) posits that the changes in the proportional shares of DNA binding between experimental states correspond to the changes in the total amount of DNA occupancy between experimental states. However, changes in the proportional shares of DNA binding only would track the well with changes in the total amount of DNA occupancy when there is the same total amount of DNA occupancy across the experimental states (see Figure 4.3 for a further illustration of this point). Therefore, Library Size (Reads in Peaks) relies on the technical condition that there is the same total amount of DNA occupancy (per cell) across the experimental states, which is *not* met in our toy example (see Figure 3.1).

Since the technical condition underlying Library Size (Reads in Peaks) is violated in our toy example, when we compare the normalized read count under Library Size

[2]Library Size (Reads in Peaks) is further described in Section 4.1.

(Reads in Peaks) across experimental states A and B, peaks that do not have differential DNA occupancy are likely to be identified as differentially bound. Further, even when the peaks that have differential DNA occupancy are identified as differentially bound, the estimated fold change in DNA binding does not align with the true fold change in DNA occupancy between experimental states. Figure 3.3 emphasizes these points.

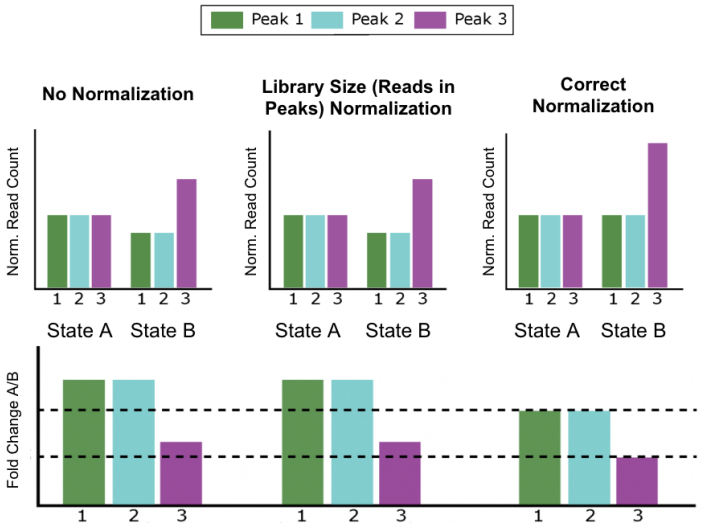

Figure 3.3: **Toy Example: Comparing normalized read counts and fold changes.** The top plot denotes the normalized read counts associated with each peak across experimental states A and B. The subplots going left to right correspond to no normalization (i.e., using the raw read counts), Library Size (Reads in Peaks) normalization, and correct normalization (i.e., normalization that returns the left plot in Figure 3.1). The bottom plot denotes the fold change A/B for each peak based on the normalized read counts. The subplots going left to right correspond to no normalization, Library Size (Reads in Peaks) normalization, and the correct normalization.

Therefore, our toy example underscores the importance of normalizing ChIP-Seq read count data as well as why it is important to satisfy the technical conditions underlying a normalization method in order to draw accurate biological conclusions about differential DNA occupancy from ChIP-Seq read count data. In the next chapter, we will categorize various ChIP-Seq normalization methods by their technical conditions. We then confirm our categorization by violating different combinations of three primary technical conditions in ChIP-Seq read count simulations and assess our empirical outcomes in Chapter 6.

Chapter 4

Normalization Methods and their Technical Conditions

In this chapter, we identify the primary technical conditions underlying various ChIP-Seq between-sample normalization methods. We then categorize the normalization methods based on the technical conditions we identified. Before we delve into the normalization methods and their respective technical conditions, however, two key terms must be defined:

1. *Reads in Peaks Normalization:* describes normalization methods that use only the reads from the consensus peak set across experimental states (Stark and Brown, 2011). An illustration of the observational unit for reads in peak normalization methods is given in Figure 4.1.

2. *Background Bin Normalization:* describes normalization methods in which the chromosomes are partitioned into large genome "bins" (that are roughly 15,000 base pairs long). The size factors are then computed from the bins (Stark and Brown, 2011). An illustration of the observational unit for background bin normalization methods is given in Figure 4.2.[1]

[1]Note that only chromosomes that have at least one peak in the consensus peak set across experimental states are partitioned into background bins. However, we are only using one chromosome in our simulation. So, all genomic regions belong to some background bin in our simulation (Stark and Brown, 2011).

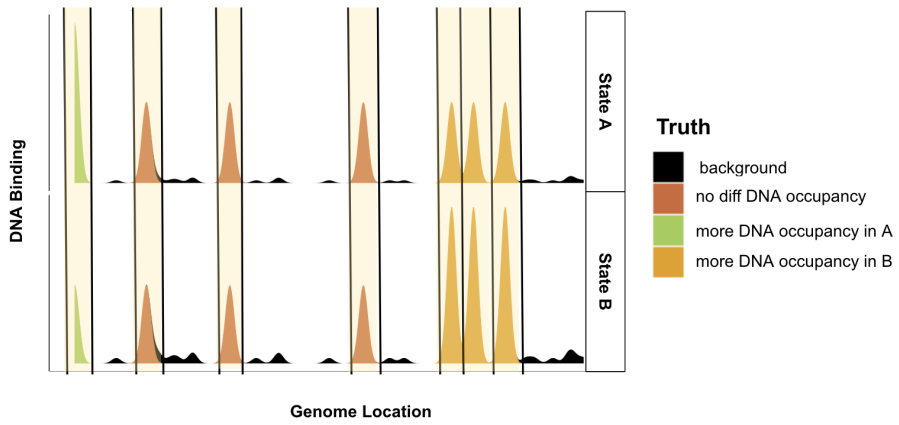

Figure 4.1: **Observational unit for *Reads in Peaks Normalization.*** The vertical axis represents the DNA binding, and the horizontal axis represents the genome location of the DNA binding. The plots are partitioned based on the experimental states A and B. The genomic regions highlighted in yellow are the ones that would be used for *Reads in Peaks Normalization*. The black vertical lines denote the boundaries of each observational unit. Notably, the background regions (colored in black) are not considered in Reads in Peaks Normalization.

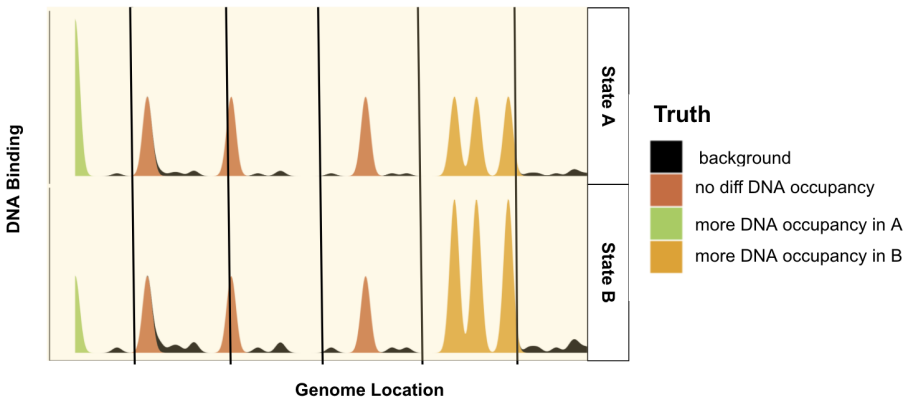

Figure 4.2: **Observational Unit for *Background Bin Normalization.*** The vertical axis represents the DNA binding, and the horizontal axis represents the genome location of the DNA binding. The plots are partitioned based on the experimental states A and B. The genomic regions highlighted in yellow are the ones that would be used for *Background Bin Normalization*. The black vertical lines denote the boundaries of the observational unit. Notably, for Background Bin Normalization, the genome is divided up via "mile markers," which are roughly 15,000 base pairs apart.

Now that key terms have been defined, we present our categorization of popular ChIP-Seq normalization methods by their technical conditions. Through the technical conditions, we develop four overarching categories of normalization methods:

1. Normalization by Library Size (Section 4.1)
2. Normalization by Distribution (Section 4.2)
3. Normalization by Background (Section 4.3)
4. Normalization by Control (Section 4.4)

For each category of normalization method, we provide an overview of the category, its primary technical conditions, popular ChIP-Seq normalization methods in the category, and finally, the motivation for normalizing ChIP-Seq read count data with a normalization method in that category.

4.1 Normalization by Library Size

The goal of normalization by Library Size is to remove any differences in DNA binding that are due to experimental artifacts rather than differential DNA occupancy. Normalization by Library Size attempts to remove the differences in DNA binding that are due to experimental artifacts by scaling each sample based on its total read count (Dillies et al., 2012).

4.1.1 Technical Conditions

- **Equal Amount of Total DNA Occupancy:** the amount of total DNA occupancy is equal across the two experimental states. That is, each experimental state has the same amount of DNA occupancy (per cell).

4.1.2 Methods

1. **Library Size (Background Bins)** first calculates the total raw read count in each sample (across both peak and background regions). The sample is then normalized by dividing the raw read count in each peak by the total raw read count. (Dillies et al., 2012). Thus, for Library Size (Background Bins), the size factor for sample j (i.e., s_j) is equivalent to sample j's total raw read count. That is:

$$s_j = \sum_{b=1}^{B} k_{bj} \tag{4.1}$$

where B is the number of background bins in sample j and k_{bj} is the raw read count associated with background bin b in sample j.

2. **Library Size (Reads in Peaks)** is similar to Library Size (Background Bins), but instead of summing all the raw read counts, Library Size (Reads in Peaks) only sums the raw read counts associated with the consensus peak set across experimental states to get the sample's total raw peak read count. The sample is then normalized by dividing the raw read count in each peak by the sample's total raw peak read count (Dillies et al., 2012). Thus, for Library Size (Reads in Peaks), the size factor for sample j is equivalent to sample j's total raw peak read count. That is:

$$s_j = \sum_{i=1}^{N} k_{ij} \tag{4.2}$$

where N is the number of background bins in sample j and k_{ij} is the raw read count associated with peak i in sample j.

3. **MAnorm2** involves a two-step normalization procedure that normalizes one sample against another (Tu et al., 2020). The first step of MAnorm2's normalization procedure is normalizing the samples within a given experimental state. The second step is normalizing across the experimental states. Both normalization steps employ the same general normalization algorithm. Namely, the raw read count associated with a given peak in a specific sample is normalized by applying a linear transformation to the $\log_2$ raw read count associated with that peak in the given sample. Let k_{ij} denote the raw read count associated with peak i in sample j. Then, in MAnorm2, the normalized read count for peak i in sample j relative to sample l, which we denote as k^*_{ij}, is defined as follows:

$$k^*_{ij} = \alpha_j + \beta_j \cdot \log_2(k_{ij}) \tag{4.3}$$

Notably, unlike other considered normalization methods, MAnorm2 does not generate a general size factor s_j that is applied to the entire sample j. Instead, each peak's normalized read count is a linear function of its raw read count for MAnorm2. The linear coefficients α_j and β_j associated with sample j in Equation (4.3) are calculated by comparing the raw read counts associated with the common peaks across two samples. A peak is considered a *common peak* if it is called a peak under both samples (e.g., if it is a peak in both sample j and sample l) (Tu et al., 2020).[2]

[2]In our simulation, which is described in more detail in Chapter 5, every peak is simulated as a *common peak*.

The coefficients α_j and β_j in Equation (4.3) are defined as follows:

$$\alpha_j = \text{mean}_i\Big(\log_2([k_l])) - \beta_j \cdot \text{mean}_i(\log_2([k_j]))\Big) \tag{4.4}$$

$$\beta_j = \frac{sd_i(\log_2([k_l]))}{sd_i(\log_2([k_j]))} \tag{4.5}$$

where $[k_j]$ is a vector with entries consisting of the raw read count associated with common peak i in sample j, and sd_i denotes the sample standard deviation of the raw read counts across the common peaks i in a given sample.

By defining the coefficients α_j and β_j in this way, MAnorm2 forces the average of the normalized $\log_2$ fold changes in the common peak regions across samples j and l to be zero and the covariance between the normalized $\log_2$ fold changes and the normalized absolute $\log_2$ signal intensities in the common peak regions across samples j and l to be zero as well (i.e., it forces the normalized $\log_2$ fold changes in the common peak regions and the normalized absolute $\log_2$ signal intensities in the common peak regions to be independent of one another) (Tu et al., 2020). Let $[M^*]$ denote the vector of the normalized $\log_2$ fold changes in the common peak regions across samples j and l, and $[A^*]$ denote a vector of the normalized absolute $\log_2$ signal intensities in the common peak regions across samples j and l. Then, $[M^*]$ and $[A^*]$ are computed as follows:

$$[M^*] = [k_j^*] - \log_2([k_l]) \tag{4.6}$$

$$[A^*] = \frac{1}{2}\Big(\log_2([k_l]) + [k_j^*]\Big) \tag{4.7}$$

where $[k_j^*]$ is a vector consisting of the normalized read counts in sample j associated with each peak i (i.e., k_{ij}^*), and k_l is a vector consisting of the raw read counts associated with sample l for each peak i.

Then, given how α_j and β_j are defined, the $[M^*]$ and $[A^*]$ vectors associated with the two samples will satisfy the following set of equations under MAnorm2:

$$\text{mean}_i([M^*]) = 0 \tag{4.8}$$

$$\text{cov}_i([M^*], [A^*]) = 0 \tag{4.9}$$

Since MAnorm2 sets the average normalized $\log_2$ fold change between samples j and l to be zero (see Equation (4.8)), MAnorm2 relies on the technical condition

that there is an equal amount of total DNA occupancy in the common peaks across the experimental states. Otherwise, the mean $\log_2$ fold change across the common peaks would be skewed toward the sample that has more total DNA occupancy in its common peaks. Indeed, Tu et al. (2020) state that MAnorm2 makes the assumption that there is "no global change" in DNA occupancy in the common peak regions across the two samples (p. 132). Moreover, if a sample has both common and non-common peaks, then MAnorm2 must rely on the technical condition that the effects of experimental artifacts on the common peaks are the same as the effects of experimental artifacts on the non-common peaks in order to use the coefficients α_j and β_j to normalize each peak in sample j via Equation (4.3).

4.1.3 Motivation

When there is the same total DNA occupancy technical condition across experimental states, then we expect a peak to have the same proportional share of DNA binding across the experimental states (see Figure 4.3 for an illustration of this point) and, moreover, for the mean $\log_2$ fold change between two different samples to be zero. Thus, insofar as there is an equal amount of total DNA occupancy across experimental states, we would expect that peaks with different proportional shares of DNA binding across experimental states would also have differential DNA occupancy across the states. Likewise, when there is an equal amount of total DNA occupancy across experimental states, the true fold change in the amount of DNA occupancy between experimental states would have a mean of zero.

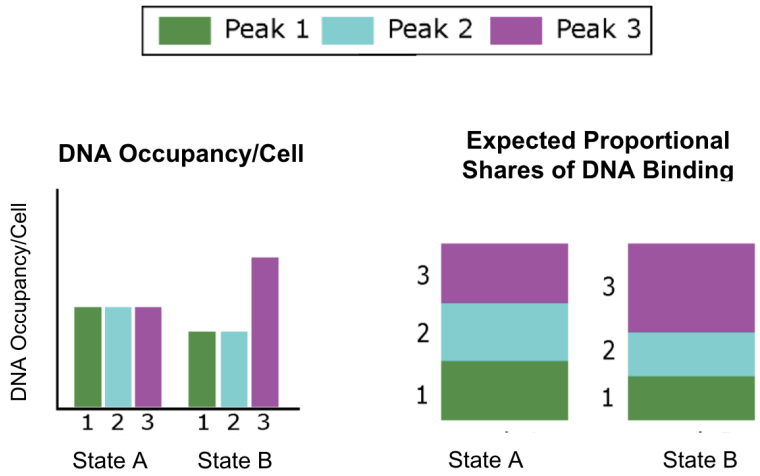

Figure 4.3: **Illustration of Normalization by Library Size.** The left plot represents the true DNA occupancy per cell, and the right plot represents the expected proportional share of DNA binding (i.e., the proportional shares of DNA occupancy). For both plots, the horizontal axis denotes the experimental state (i.e., A or B). In the left plot, the vertical axis represents the DNA occupancy per cell. In the right plot, the vertical axis represents the proportional share of DNA binding. Note that the amount of DNA occupancy across experimental states is *equal*. Thus, if ChIP-Seq data was repeatedly collected, we would expect that, on average, the proportional share of DNA binding in a given peak would match the proportional share of DNA occupancy per cell in that peak. For example, Peak 3 has an A/B fold change of 2/3 in the total DNA occupancy per cell in the left plot, as well as an A/B fold change of 2/3 in the expected proportional shared of DNA binding in the right plot.

4.2 Normalization by Distribution

Normalization by Distribution works by comparing some aspect of the distribution of read counts (or a function of read counts) across experimental states in order to calculate the size factor for each sample. If the effects of experimental artifacts on the peaks with differential DNA occupancy and without differential DNA occupancy are the same, then we can normalize the entire sample by comparing the read counts in the peaks that are estimated to not have differential DNA occupancy across the experimental states.

4.2.1 Technical Conditions

- **Peaks with Differential DNA Occupancy and Peaks without Differential DNA Occupancy Behave the Same:** the effects of experimental artifacts on peaks with differential DNA occupancy are the same as the effects of experimental artifacts on peaks without differential DNA occupancy.

- **Symmetric Differential DNA Occupancy:** there is roughly symmetric differential DNA occupancy across experimental states. That is, the number of peaks with more DNA occupancy in one experimental state is equal to the number of peaks with more DNA occupancy in the other experimental state.

4.2.2 Methods

1. **TMM (Reads in Peaks)** or "Trimmed Mean of the M-values for Reads in Peaks" first selects a reference sample to be the basis of comparison for the other samples. Then, the $\log_2$ fold change of signal magnitudes and the absolute signal magnitudes for each peak in the consensus peak set, relative to this reference sample, are calculated (Robinson and Oshlack, 2010).

   Let k_{ij} be the raw read count associated with peak i in sample j and N_j be the total number of reads associated with the consensus peak set in sample j (i.e., $N_j = \sum_i k_{ij}$). Finally, let r denote the sample selected as the reference. Then, the $\log_2$ fold change for peak i in sample j, relative to reference r, is defined as:

$$M_{ij}^r = \frac{\log_2(k_{ij}/N_j)}{\log_2(k_{ir}/N_r)} \tag{4.10}$$

   While, the $\log_2$ absolute signal magnitude for peak i in sample j, relative to reference r, is defined as:

$$A_{ij}^r = \frac{1}{2}\log_2\left(\frac{k_{ij}}{N_j} \cdot \frac{k_{ir}}{N_r}\right) \tag{4.11}$$

   The M_{ij}^r and A_{ij}^r values are trimmed twice such that only peaks with central M_{ij}^r and A_{ij}^r values remain (Robinson and Oshlack, 2010). The purpose of trimming the M_{ij}^r and A_{ij}^r values is to approximate the set of peaks without differential DNA occupancy. However, for trimming the M_{ij}^r and A_{ij}^r values to accurately approximate the set of peaks without differential DNA occupancy, there must be symmetric differential DNA occupancy across the experimental states. We will refer to the set of peaks remaining after the trimming as Q. The set Q is used to calculate the size factor for sample j (i.e., s_j), which is scaled by the reference sample r:

$$\log_2(s_j) = \frac{\sum_{i \in Q} w_{ij}^r M_{ij}^r}{\sum_{i \in Q} w_{ij}^r} \tag{4.12}$$

where,

$$w_{ij}^r = \frac{N_j - k_{ij}}{N_j k_{ij}} + \frac{N_r - k_{ir}}{N_r k_{ir}} \tag{4.13}$$

Thus, TMM (Reads in Peaks) calculates the size factor for sample j using only features about peaks in set Q, which aims to estimate the set of peaks that do not have differential DNA occupancy across the experimental states. Therefore, TMM (Reads on Peaks) relies on the effects of experimental artifacts on the peaks without differential DNA occupancy, which is the same as the effects of experimental artifacts on the peaks with differential DNA occupancy.

2. **RLE (Reads in Peaks)** or "Relative Log Expression for Reads in Peaks" begins by finding the ratio between the raw read count associated with peak i in a given sample and the geometric mean of the raw read count associated with peak i across all the samples. The process of calculating the ratio between the raw read count and the geometric mean of the raw read counts is repeated until a ratio has been generated for every peak in the consensus peak set. Then, the median ratio is computed for each sample and serves as its size factor. The size factor is applied to the raw read count in each peak to normalize the sample (Love et al., 2014). Then, the size factor for sample j (s_j) under RLE (Reads in Peaks) is calculated as follows:

$$s_j = \text{median}_i \left(\frac{k_{ij}}{(\prod_{l=1}^{m} k_{il})^{1/m}} \right) \tag{4.14}$$

where k_{ij} is the raw read count associated with peak i in sample j and m is the total number of samples across all experimental states.

In Equation (4.14), the denominator is the genomic mean of the raw read counts associated with peak i across all the samples. This geometric mean serves as the read count associated with peak i in a pseudo-reference sample (Love et al., 2014). The read count for each peak in the pseudo-reference sample is compared to the raw read count associated with peak i in sample j (i.e., k_{ij}). Next, the median ratio of the raw read counts is taken across all the peaks. The purpose of taking the median ratio is to approximate a ratio that is associated with a peak that does not have differential DNA occupancy (Love et al., 2014). So, RLE (Reads in Peaks) relies on the assumption that the median peak will not have differential DNA occupancy across the experimental states. However, when there is asymmetric DNA occupancy coupled with a high proportion of peaks

with differential DNA occupancy, the median peak might have differential DNA occupancy. Thus, RLE (Reads in Peaks) relies on the technical condition that there is symmetric differential DNA occupancy across the experimental states to generate samples' size factors (Love et al., 2014). Moreover, given that RLE (Reads in Peaks) generates each sample's size factor by approximating a peak without differential DNA occupancy, it follows that RLE (Reads in Peaks) also relies on the technical condition that peaks without differential DNA occupancy behave the same as peaks with differential DNA occupancy.

4.2.3 Motivation

Normalization by Distribution posits that we can estimate the size factor for a sample by using by using a measure of center, such as the median, to estimate peaks that do not have differential DNA occupancy. Moreover, if the effects of experimental artifacts on the peaks that have differential DNA occupancy are the same as the effects of experimental artifacts on the peaks that do not have differential DNA occupancy, then we can normalize the entire sample based on the size factor that is calculated by estimating the peaks that do *not* have differential DNA occupancy across the experimental states.

4.3 Normalization by Background

Normalization by background first partitions the genome into long segments that are roughly 15,000 base pairs long (Stark and Brown, 2011). These long genomic regions are referred to as *background bins*. For Normalization by Background, the background bins, rather than the peaks in the consensus peak set across experimental states, then serve as the unit of normalization (see Figure 4.2).

4.3.1 Technical Conditions

- **Previous Technical Conditions:** the aforementioned technical conditions for any given method must be met by the background bins across the experimental states.
- **Same Total Amount of Background Binding:** the number of rogue reads is the same across experimental states.

4.3.2 Methods

1. **Library Size (Background Bins)** normalization with Library Size (Background Bins) is described in Section 4.1. To summarize: a sample is normalized using Library Size (Background Bins) by dividing the raw read count associated

with each peak by the total raw read count summed over the background bins (Dillies et al., 2012).

2. **TMM (Background Bins)** or "Trimmed Mean of the M-values using Background Bins" works very similarly to TMM (Reads in Peaks). It first selects a reference sample to be the basis of comparison for the other samples. Then, the fold changes and absolute signal magnitudes relative to the reference sample are calculated and used to normalize the raw read counts associated with each peak in the sample (Robinson and Oshlack, 2010). The major difference between TMM (Reads in Peaks) and TMM (Background Bins) is the unit used for normalization. In TMM (Background Bins), the units used for normalization are the sample's background bins rather than its peaks.

   Let the $\log_2$ fold change for background bin b between sample j and the reference sample r be denoted as M^r_{bj}. Further, let k_{bj} denote the reads aligned to bin b in replicate j, and N_j denote the total number of reads aligned to replicate j. Likewise, let k_{br} denote the reads aligned to bin b in the reference replicate r and N_r denote the total number of reads aligned to reference replicate r. Then, M^r_{bj} is defined as follows:

$$M^r_{bj} = \frac{\log_2 (k_{bj}/N_j)}{\log_2(k_{br}/N_r)} \tag{4.15}$$

   While, the absolute signal magnitude for background bin b in sample j, which we denote as A^r_{bj}, is defined as:

$$A^r_{bj} = \frac{1}{2} \log_2 \left(\frac{k_{bj}}{N_j} \cdot \frac{k_{br}}{N_r} \right) \tag{4.16}$$

   Like with TMM (Reads in Peaks), the M^r_{bj} and A^r_{bj} are trimmed twice. As a result, only background bins with central M^r_{bj} and A^r_{bj} values remain (Robinson and Oshlack, 2010). The purpose of trimming the M^r_{bj} and A^r_{bj} values is so that only the background bins with approximately the same amount of DNA binding across experimental states are used to generate the sample's size factor. Yet, when there is a different total amount of background binding across experimental states, background bins can have high M^r_{bj} and A^r_{bj}, even if there is no difference in the total DNA occupancy across the experimental states in the background bin. Thus, TMM (Background Bins) relies on the technical condition that there is the same total amount of background binding across the experimental states. We denote the set of background bins left after the two rounds of trimming as Q_b. The set Q_b is used to create the size factor for replicate j, scaled by reference r (i.e., s_j) that can be applied to the raw read count for each peak i in sample j:

$$\log_2(s_j) = \frac{\sum_{i \in Q_b} w_{bj}^r M_{bj}^r}{\sum_{i \in Q_b} w_{bj}^r} \tag{4.17}$$

where,

$$w_{bj}^r = \frac{N_j - k_{bj}}{N_j k_{bj}} - \frac{N_r - k_{br}}{N_r k_{br}} \tag{4.18}$$

3. **RLE (Background Bins)** or "Relative Log Expression using Background Bins" uses the same procedure as RLE (Reads in Peaks), but instead of peaks, background bins are the unit used for normalization. RLE (Background Bins) begins by calculating a ratio between the raw read counts for a specific background bin and the geometric mean of the read count associated with the same background bin across all of the samples (Love et al., 2014). The median ratio then serves as the size factor for the sample. That is, the size factor for sample j (s_j) under RLE (Background Bins) is calculated as follows:

$$s_j = \text{median}_b \left(\frac{k_{bj}}{(\prod_{l=1}^{m} k_{bl})^{1/m}} \right) \tag{4.19}$$

where k_{bj} is the read count associated with background bin b in replicate j and m is the total number of replicates across all experimental states.

The purpose of taking the median ratio over all the background bins is to approximate a background bin where there is no difference in DNA occupancy between the experimental states (Love et al., 2014). However, when there is a difference in the amount of total background binding between experimental states, background bins that do not have differences in DNA occupancy between experimental states **do** have differences in DNA binding between experimental states. As a result, the median ratio might not return a background bin where there is the same amount of DNA occupancy between experimental states when if is a difference in the total amount of background binding across the experimental states. Therefore, RLE (Background Bins) also relies on the technical condition that there is an equal amount of background binding across the experimental states.

4.3.3 Motivation

Given that peaks are usually only a few hundred base pairs long, we would not expect to identify a (non-trivial) global change in the DNA occupancy (per cell) in the background bins across the experimental states (Stark and Brown, 2011). Hence, any observed difference in the amount of DNA binding is expected to be the result of experimental artifacts rather than true biological differences in the amount of DNA occupancy per cell between the experimental states. So, it is beneficial to generate

a sample's size factor by leveraging features of its background bins rather than its peaks.

4.4 Normalization by Controls

When the technical conditions listed in Sections 4.1.1, 4.2.1, and 4.3.1 are violated, Normalization by Controls can enable us to properly normalize between ChIP-Seq samples. Normalization by Controls works by generating the size factors using *control peaks*. *Control peaks* are peaks where we have a priori knowledge of the DNA occupancy behavior (Bonhoure et al., 2014). A common type of control peak is a *negative control*. In a *negative control*, we would expect any differences in the DNA binding across experimental states in the control peak region to be the result of experimental artifacts rather than differences in DNA occupancy (Bonhoure et al., 2014).

4.4.1 Technical Conditions

- **Existence of Controls:** the controls needed for the experiment exist, and their DNA occupancy behavior is what we understand it to be (e.g., a negative control).
- **Controls Behave like Peaks:** the effects of experimental artifacts on the control peaks reflect the effects of experimental artifacts on all the peaks.
- **Previous Technical Conditions:** the aforementioned technical conditions for any given normalization method must be met by the control.

4.4.2 Methods

1. **Spike-in Normalization** works by modifying the ChIP-Seq data collection procedure. Namely, before immunoprecipitation is done, chromatin from another organism is injected into each sample. The peaks resulting from the spiked-in chromatin are then used to generate the size factors for the entire sample using a normalization method, like those we previously described (e.g., TMM (Reads in Peaks), RLE (Reads in Peaks), Library Size (Reads in Peaks)). The choice of which normalization method is ultimately based on our a priori knowledge of the DNA occupancy behavior for the spike-in control. Given that the size factor calculated through the spiked-in peaks is applied to all the peaks, spike-in normalization relies on the technical condition that the experimental artifacts have the same effect on the spiked-in peaks and the peaks of interest.

Figure 4.4 provides a cartoon example of Spike-in Normalization.[3]

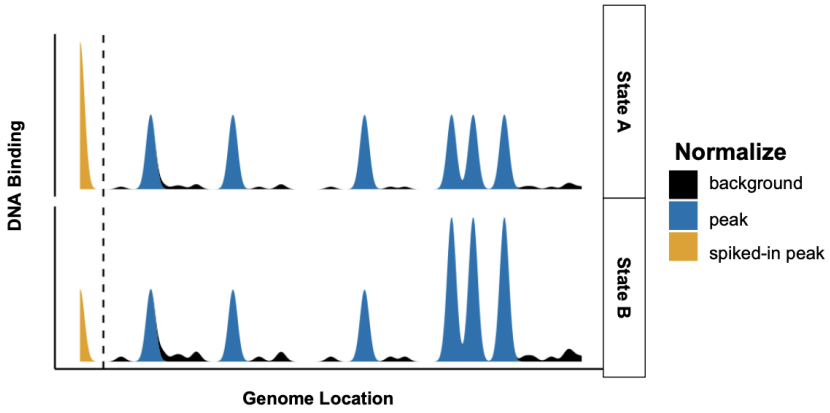

Figure 4.4: **Spike-in Normalization Illustration.** The vertical axis represents the number of reads associated with different regions on the genome. The spiked-in control peak in this cartoon example is colored orange. Here, the spiked-in peak serves as a negative control. Thus, we would expect there to be the same number of reads in the spiked-in peak across conditions A and B. We would use the fact that the spike-in control is a negative control to create a size factor using some normalization method like TMM (Reads in Peaks). The calculated size factor would then be applied to the entire sample to normalize the peaks of interest in that sample.

4.4.3 Motivation

Unlike for the peaks of interest, we have an a priori understanding of the spiked-in controls' DNA occupancy behavior across the experimental states (Bonhoure et al., 2014). So, if the DNA binding across experimental states misaligns with what we expect, then we can take this misalignment to be the result of experimental artifacts rather than differences in DNA occupancy. If the effects of experimental artifacts on the peaks of interest are equivalent to the effects of experimental artifacts on the control peaks, then we can apply the size factor calculated using the spike-in control peaks to the entire sample.

[3]Unlike in the illustration provided in Figure 4.4, there are usually several such spiked-in peaks in real-life experiments rather than just one.

Chapter 5

Simulation Preliminaries

We use simulated ChIP-Seq read count data to assess our categorization of the popular ChIP-Seq normalization methods given in Chapter 4. We chose to simulate the ChIP-Seq read count data rather than use real ChIP-Seq read count data for two primary reasons. First, unlike with experimental data, we are omniscient about which peaks have differential DNA occupancy in simulated data. Second, we are also omniscient about which combination of the technical conditions are being violated in simulated ChIP-Seq read count data. As such, simulating ChIP-Seq read count data enables us to more accurately estimate the effects of violating various combinations of technical conditions on various ChIP-Seq normalization methods. However, there are also shortcomings to simulating rather than using real data. Notably, we cannot realistically approximate spike-in control peaks in our ChIP-Seq read count simulation because our simulation does not presume characteristics about the protein of interest, chromosome, or the kind of antibody used for immunoprecipitation. So, we cannot realistically simulate spike-in controls that are sensitive to such characteristics.

Following the work of Anders and Huber (2010) as well as Lun and Smyth (2016), we use a negative binomial distribution to simulate the total read count in a given peak. The mean read count for peak i in sample j (i.e., μ_{ij}) in the negative binomial distribution for the read count is simulated to be the following:

$$\mu_{ij} = \frac{p_{ij} \cdot (fc)_{ij}}{(\text{basesum})_j} \cdot (\text{library multiplier})_j \cdot 600,000 \tag{5.1}$$

In Equation (5.1), $600,000$ serves as our base number of aligned reads in a given sample (i.e., library size), which is then multiplied by a scalar (i.e., $(\text{library multiplier})_j$) in order to calculate the total number of aligned reads for the specific sample j in our simulation. The $(\text{library multiplier})_j$ term is added to account for the fact that sometimes the total library size of samples is different in virtue of experimental artifacts, e.g., leaving one sample in the sequencer for longer than another. Like Evans et al. (2016), we simulate $(\text{library multiplier})_j$ by using a truncated normal distribution with a mean of 1, a standard deviation of 0.2, and an upper and lower bound

of 1.4 and 0.6, respectively. In Equation (5.1), p_{ij} denotes the proportional share of DNA occupancy in peak i in sample j before we simulate differential DNA occupancy. We calculate p_{ij} by dividing the total number of aligned reads in sample j (i.e., $(\text{library multiplier})_j \cdot 600,000$) by the total number of peaks in the sample. Differential DNA occupancy is then simulated by multiplying each peak i in sampler j by $(fc)_{ij}$, which represents the fold change in DNA occupancy for peak i in sample j relative to a sample in a different experimental state. Notably, when a peak does not have differential DNA occupancy between experimental states, $(fc)_{ij}$ will be one. Finally, the value $(\text{basesum})_j$ is defined as the dot product of $[p_j]$, which is a column vector consisting of the original proportional share of DNA occupancy (i.e, p_{ij}), and $[(fc)_j]^\intercal$, which is a row vector consisting of the multiplicative change in the amount of DNA occupancy per cell in peak i between experimental states A and B (i.e., $(fc)_{ij}$). Equation (5.2) provides the formula for calculating the basesum for sample j using $[p_j]$ and $[(fc)_j]$.

$$(\text{basesum})_j = [(fc)_j]^\intercal \cdot [p_j] \tag{5.2}$$

To further elucidate how the basesum for a given sample is calculated, let us suppose we had three peaks in two samples denoted as A and B. In this example, $[p_A] = [p_B] = [1/3, 1/3, 1/3]$. Moreover, $[(fc)_A] = [1, 1, 1]$ and $[(fc)_B] = [1, 1, 2]$, meaning that only the third peak has differential DNA occupancy, with two times more DNA occupancy in B than A. Using Equation (5.2), it follows that $(\text{basesums})_A = 1$ and $(\text{basesums})_B = 4/3$. We will build upon this simple example in the next section when we illustrate the Oracle normalization method in Figure 5.1.

5.1 The Oracle Normalization Method

Under ideal circumstances, perfect normalization would entail that all and only the peaks that have differential DNA occupancy are classified as differentially bound. However, peaks can be classified as differentially bound even though the peak does not have differential DNA occupancy due to other steps in the differential binding analysis workflow beyond between-sample normalization (e.g., a peak with differential DNA occupancy might not be called in Step (4), or differential binding analysis generates a high adjusted p-value to a peak with differential DNA occupancy in Step (8)). Therefore, a normalization method might still be performing well even if some peaks that have differential DNA occupancy are not classified as differentially bound or, likewise, if there are many peaks that are classified as differentially bound that do not have differential DNA occupancy. As such, we develop an omniscient normalization method, which we refer to as the *Oracle*, to serve as our basis of comparison in the simulation results. The Oracle size factor for sample j in a given simulation iteration is defined as follows, where m is the total number of samples across the two experimental states:

$$s_j = \frac{(\text{normFactor})_j}{\prod_{h=1}^{m} (\text{normFactor})_h{}^{(1/m)}} \tag{5.3}$$

where,

$$(\text{normFactor})_j = \frac{(\text{library multiplier})_j}{(\text{basesum})_j} \tag{5.4}$$

Importantly, when we divide the mean read count for peak i under sample j (i.e., μ_{ij}) by the normalization factor for sample j (i.e., $(\text{normFactor})_j$), we get the following:

$$\frac{\mu_{ij}}{(\text{normFactor})_j} = p_{ij} \cdot (fc)_{ij} \cdot 600,000 \tag{5.5}$$

Since p_{ij} (i.e., the base proportion of DNA occupancy in peak i in sample j) and $600,000$ (i.e., the base library size) are constant across different samples in our simulation, the only value that changes in Equation (5.5) between experimental states is $(fc)_{ij}$, i.e., the fold change in DNA occupancy for peak i in sample j relative to a sample in the other experimental state. Therefore, when we compare read counts that are normalized by the Oracle across experimental states, we are directly estimating the fold change in DNA occupancy across experimental states, i.e., differential DNA occupancy. Equation (5.3) simply standardizes the normalization factor for sample j relative to the other samples' size factors by dividing the normalization factor for sample j by the geometric mean of the normalization factors. Our toy example illustrated in Figure 5.1 further demonstrates how the Oracle successfully normalizes raw read count data by leveraging each sample's basesum and library multiplier.

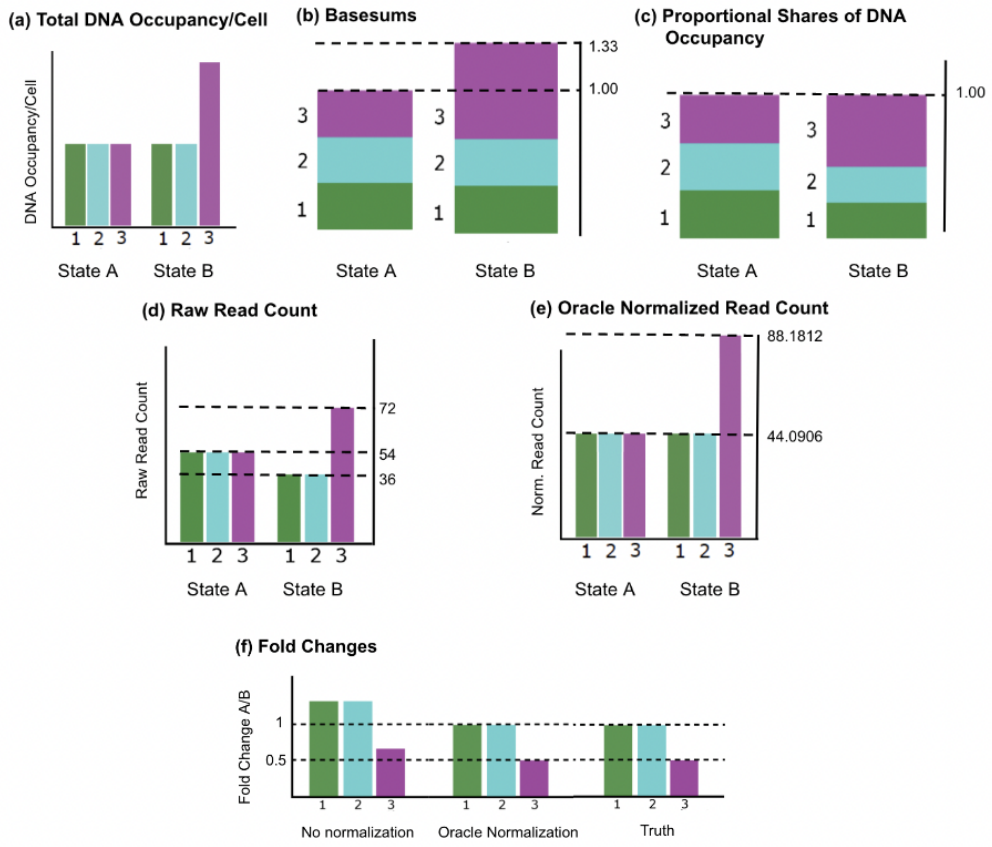

Figure 5.1: **Illustration of Oracle Normalization Method.** In our toy example, there are two experimental states, A and B, and three peaks in each experimental state, denoted as 1,2, and 3. The library multipliers for State A and State B are 0.9 and 0.8, respectively. Plot **(a)** represents the DNA occupancy (per cell) for each peak in States A and B. Plot **(b)** is the corresponding basesums for States A and B. Plot **(c)** provides the proportional shares of DNA occupancy for the peaks for States A and B. The raw read counts are generated through the general process outlined in Equation (5.1) and are depicted in Plot **(d)**. Using the raw read counts, Peaks 1 and 2 appear differentially bound even though they do not have different DNA occupancy in Plot **(f)**. Additionally, even though Peak 3 has 1/2 as much DNA occupancy in State A as compared to B in Plot **(a)**, it appears to have 2/3 as much in Plot **(f)**. Note that when we normalize our toy data using the Oracle, the fold changes associated with the Oracle normalized read counts in Plot **(e)** match the *true* fold changes in the DNA occupancy between the states depicted in Plot **(f)**.

We can find the Oracle normalization factors associated with States A and B in our toy example in Figure 5.1 by using Equation (5.4) and leveraging that the library multipliers for State A and State B are 0.9 and 0.8, respectively.

$$(\text{normFactor})_A = \frac{0.9}{1} = 0.9 \tag{5.6}$$

$$(\text{normFactor})_B = \frac{0.8}{\frac{4}{3}} = 0.6 \tag{5.7}$$

Then, we can find the Oracle size factor for each experimental state by using Equation (5.3).

$$s_A = \frac{0.9}{\sqrt{0.6 \cdot 0.9}} = 1.2247 \tag{5.8}$$

$$s_B = \frac{0.6}{\sqrt{0.6 \cdot 0.9}} = 0.8165 \tag{5.9}$$

Dividing the raw read counts for States A and B (given in Plot **(d)** of Figure 5.1) by their respective Oracle size factors of 1.04447 and 0.95743 gives us the Oracle normalized read counts given in Plot **(e)** of Figure 5.1. Table 5.1 shows the calculations for the oracle normalized read counts.

	Norm. read count state A	Norm. read count state B
Peak 1	$54/1.2247 = 44.0906$	$36/0.8165 = 44.0906$
Peak 2	$54/1.2247 = 44.0906$	$36/0.8165 = 44.0906$
Peak 3	$54/1.2247 = 44.0906$	$72/0.8165 = 88.1812$

Table 5.1: **Oracle Normalized Read Counts.** The columns correspond to the experimental state associated with the raw read count, and the rows correspond to the peak associated with the raw read count. Each entry is the Oracle normalized read count for the specific peak, which is found by multiplying the raw read count depicted in Plot **(d)** of Figure 5.1 by the size factors calculated in Equations (5.8) and (5.9).

Using the Oracle normalized read counts computed in Table 5.1, we would conclude that Peaks 1 and 2 are **not** differentially bound and that Peak 3 **is** differentially bound, with two times more DNA binding in State B than State A. As shown in Plots **(e)** and **(f)** of Figure 5.1, the Oracle-normalized DNA binding results perfectly align with the amount of DNA occupancy (per cell) depicted in Plot **(a)** of Figure 5.1. Therefore, our toy example demonstrates that our Oracle normalization method is omniscient and, thus, will serve as an effective basis for comparison with other normalization methods in our simulation.

5.2 Simulation Conditions

In each simulation, we have two different experimental states (which we refer to as A and B) and four replicates within each experimental state. We only simulate transcription factor binding, which is known to generate narrower and taller peaks than other proteins (e.g., histone modifications) (Lun and Smyth, 2016). In our ChIP-Seq simulation, we violate every combination of the following technical conditions:

1. **Symmetric Differential DNA Occupancy**: there is roughly symmetric differential DNA Occupancy across experimental states. That is, there is the same number of regions with more DNA occupancy in one experimental state than another across the experimental states.

2. **Same Total DNA Occupancy**: the amount of total DNA occupancy is the same across the experimental states. That is, each experimental state has the same amount of DNA occupancy per cell.

3. **Same Total Amount of Background Binding**: the number of rogue DNA binding is the same across the two experimental states.

As such, we have eight unique simulation conditions in total. In what follows, we provide more details about each simulation condition. Table 5.2 and 5.2 further elucidate the simulation conditions.

Symmetry, Equal DNA Occupancy, Equal Background Binding:

- *Peaks:* 50% of the peaks that have differential DNA occupancy have 2 times more DNA occupancy in experimental state A. The other 50% of the peaks that have differential DNA occupancy have 2 times more DNA occupancy in experimental state B.

- *Background:* The amount of background binding within each background bin is, on average, 50% of the total DNA occupancy for a peak that does not have differential DNA occupancy in both experimental states. The total amount of background binding is uniformly distributed over the background bins in each sample.

Symmetry, Different DNA Occupancy, Equal Background Binding

- *Peaks:* 50% of the peaks with differential DNA occupancy have 4 times more DNA occupancy in experimental state A. The other 50% of the peaks with differential DNA occupancy have 2 times more DNA occupancy in experimental state B.

- *Background:* The amount of background binding within each background bin is, on average, 50% of the total DNA occupancy for a peak that does not have differential DNA occupancy in both experimental states. The total amount of background binding is uniformly distributed over the background bins in each sample.

Asymmetry, Equal DNA Occupancy, Equal Background Binding

- *Peaks:* 12.5% of the peaks with differential DNA occupancy have 8 times more DNA occupancy in experimental state A. The other 87.5% of the peaks with differential DNA occupancy have 2 times more DNA occupancy in experimental state B.

- *Background:* The amount of background binding within each background bin is, on average, 50% of the total DNA occupancy for a peak that does not have differential DNA occupancy in both experimental states. The total amount of background binding is uniformly distributed over the background bins in each sample.

Asymmetry, Different DNA Occupancy, Equal Background Binding

- *Peaks:* 12.5% of the peaks with differential DNA occupancy have 2 times more DNA occupancy in experimental state A. The other 87.5% of the peaks with differential DNA occupancy have 2 times more DNA occupancy in experimental state B.

- *Background:* The amount of background binding within each background bin is, on average, 50% of the total DNA occupancy for a peak that does not have differential DNA occupancy in both experimental states. The total amount of background binding is uniformly distributed over the background bins in each sample.

Symmetry, Equal DNA Occupancy, Different Background Binding

- *Peaks:* 50% of the peaks with differential DNA occupancy have 2 times more DNA occupancy in experimental state A. The other 50% of the peaks with differential DNA occupancy have 2 times more DNA occupancy in experimental state B.

- *Background:* The amount of background binding within each background bin is set to be, on average, 50% of the total DNA occupancy for a peak that does not have differential DNA occupancy in experimental state A and 25% of the DNA occupancy in a peak that does not have differential DNA occupancy in experimental state B. The total amount of background binding is uniformly distributed over the background bins in each sample.

Symmetry, Different DNA Occupancy, Different Background Binding

- *Peaks:* 50% of the peaks with differential DNA occupancy have 4 times more DNA occupancy in experimental state A. The other 50% of the peaks with differential DNA occupancy have 2 times more DNA occupancy in experimental state B.

- *Background:* The amount of background binding within each background bin is set to be, on average, 50% of the total DNA occupancy for a peak that does not have differential DNA occupancy in experimental state A and 25% of the DNA occupancy in a peak that does not have differential DNA occupancy in experimental state B. The total amount of background binding is uniformly distributed over the background bins in each sample.

Asymmetry, Equal DNA Occupancy, Different Background Binding

- *Peaks:* 12.5% of the peaks with differential DNA occupancy have 4 times more DNA occupancy in experimental state A. The other 87.5% of the peaks with differential DNA occupancy have 2 times more DNA occupancy in experimental state B.

- *Background:* The amount of background binding within each background bin is set to be, on average, 50% of the total DNA occupancy for a peak that does not have differential DNA occupancy in experimental state A and 25% of the DNA occupancy in a peak that does not have differential DNA occupancy in experimental state B. The total amount of background binding is uniformly distributed over the background bins in each sample.

Asymmetry, Different DNA Occupancy, Different Background Binding

- *Peaks:* 12.5% of the peaks with differential DNA occupancy have 2 times more DNA occupancy in experimental state A. The other 87.5% of the peaks with differential DNA occupancy have 2 times more DNA occupancy in experimental state B.

- *Background:* The amount of background binding within each background bin is set to be, on average, 50% of the total DNA occupancy for a peak that does not have differential DNA occupancy in experimental state A and 25% of the DNA occupancy in a peak that does not have differential DNA occupancy in experimental state B. The total amount of background binding is uniformly distributed over the background bins in each sample.

Symmetric	Equal DNA occ.	% with more DNA occ. in A	% with more DNA occ. in B	fold change A	fold change B
✓	✓	50%	50%	2	2
✓	X	50%	50%	4	2
X	✓	12.5%	87.5%	8	2
X	X	12.5%	87.5%	2	2

Table 5.2: **Summary of *Peaks* Simulation Conditions.** The breakdown of the percent of peaks with differential DNA occupancy that have more DNA occupancy in experimental states A and B, as well as the fold change in such peaks between the experimental states based on whether there is symmetric differential DNA occupancy (i.e., symmetry) and equal total DNA occupancy in the simulation condition.

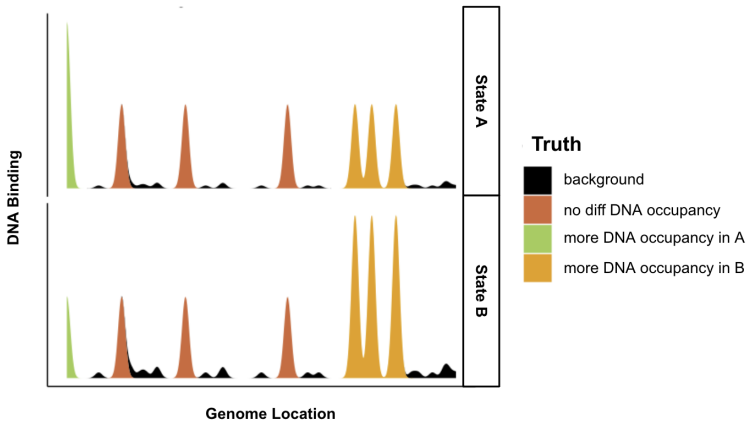

Figure 5.2: **Illustration of the *Asymmetry, Different DNA Occupancy, and Different Background Binding* Simulation Condition.** The vertical axis represents the DNA binding (i.e., read count), and the horizontal axis represents the location of the region on the genome. There is only one peak (colored green) simulated to have more DNA binding in state A than B, whereas three peaks (colored yellow) are simulated to have more DNA binding in State B than in State A. Therefore, the differential DNA occupancy is asymmetric across experimental states. The total DNA binding in the peak regions (i.e., the sum of the yellow, green, and orange regions) is simulated to be greater in State B than in State A. Therefore, there is a difference in total DNA occupancy across the experimental states. Finally, the overall background binding (colored black) is greater in state B than in state A. Hence, there is a difference in total background binding across the experimental states.

Another crucial aspect of our simulation is that we also vary the proportion of peaks with differential DNA occupancy between 0.05 and 0.95 with a step size of 0.05. We ran the simulation one hundred times (i.e., we generated one hundred independent datasets) for *each* combination of the simulation conditions and proportion of peaks with differential DNA occupancy.

5.3 Simulation Metrics

For every simulated condition and proportion of peaks with differential DNA occupancy, we calculate three metrics for each between-sample normalization method. First, we find the average false discovery rate associated with each normalization method. The false discovery rate (FDR) for a single simulation iteration and normalization method is calculated as follows, where N_{DB} denotes the number of peaks classified as differentially bound:

$$\text{FDR} = \frac{N_{DB} \text{ of those with differential DNA occupancy}}{N_{DB}} \tag{5.10}$$

In our simulation, we set the FDR cut-off to 0.05. Hence, the p-values are adjusted using the Benjamini-Hochberg method using the 0.05 cut-off, and only peaks with adjusted p-values below 0.05 are considered differentially bound. Given that we set the FDR cut-off to be 0.05, we will judge the ability of different normalization methods to control the average FDR at a level of 0.05 in our simulation.

The second metric we find is the average power associated with each normalization method. The power for a single simulation iteration and normalization method is calculated as follows (N_{DB} still denotes the number of peaks classified as differentially bound):

$$\text{Power} = \frac{N_{DB} \text{ of those with differential DNA occupancy}}{\text{Number of peaks with differential DNA occupancy}} \tag{5.11}$$

Given that the Oracle normalization method serves as the basis of comparison in our simulation, the third simulation metric directly compares the size factors generated by other normalization methods to the Oracle's size factor. We call this metric the Average Absolute Size Factor Ratio relative to the Oracle (i.e., AASFRO). To calculate AASFRO for a given normalization method, we first need to calculate the Absolute Size Factor Ratio relative to the Oracle (i.e., ASFRO) for each sample. Let s_{jem} denote the size factor for sample j under experimental state e for some normalization method m within a given simulation iteration. Let o_{je} denote the Oracle's size factor for sample j under experimental state e for that same simulation iteration. Then, the absolute size factor ratio relative to the Oracle for sample j under experimental state e for the normalization method m (i.e., $(\text{ASFRO})_{jem}$) in a given simulation iteration is defined as follows:

$$(\text{ASFRO})_{jem} = \begin{cases} \frac{s_{jem}}{o_{ij}} & \text{if } \frac{s_{jem}}{o_{je}} \geq 1 \\ \frac{o_{je}}{s_{jem}} & \text{otherwise} \end{cases} \tag{5.12}$$

By splitting the calculation of $(\text{ASFRO})_{jem}$ into the cases outlined in Equation (5.12), we ensure that $(\text{ASFRO})_{jem}$ is always greater than one and thus the *absolute* size factor ratio relative to the Oracle. After we calculate $(\text{ASFRO})_{jem}$, we calculate the *average* absolute size factor ratio relative to the Oracle under normalization method m for a given simulation iteration, which we abbreviate as $(\text{AASFRO})_m$:

$$(\text{AASFRO})_m = \frac{1}{N_e \cdot N_j} \cdot \sum_e \left(\sum_j (\text{ASFRO})_{jem} \right) \tag{5.13}$$

where N_e is the number of samples in experimental state e, and N_e is the total number of experimental states in a simulation iteration, which is always 2 in the simulation results we present in the next chapter.

Chapter 6

Simulation Results

In this chapter, we present our simulation results under different simulation conditions, focusing on the three metrics we described in Section 5.3: (1) false discovery rate, (2) power, and (3) the average absolute size factor ratio relative to the Oracle (i.e., AASFRO). We created our ChIP-Seq read count data simulation by expanding on code from Lun and Smyth (2016) as well as (Evans et al., 2016). Our code to run our simulations and generate the figures we present in this chapter is available in our GitHub repository.

6.1 General Simulation Results

We first examine how the average false discovery rate and power for different normalization methods change over the proportion of peaks with differential DNA occupancy across all eight simulation conditions. Figure 6.1 demonstrates how the average false discovery rate changes over the proportion of peaks with differential DNA occupancy for all eight simulation conditions. Meanwhile, Figure 6.2 demonstrates how the average power changes over the proportion of peaks with differential DNA occupancy for all eight simulation conditions. Note that the average false discovery rate will naturally go down as we have a higher proportion of peaks with differential DNA occupancy because there are fewer peaks that can be classified as differentially bound that do not have differential DNA occupancy. Likewise, the average power naturally increases as the proportion of peaks with differential DNA occupancy increases because there are fewer peaks that can classified as differentially bound that do not have differential DNA occupancy. In the appendix, we include versions of Figures 6.1 and 6.2 with 95% confidence interval bands around the average metric values for each normalization method.

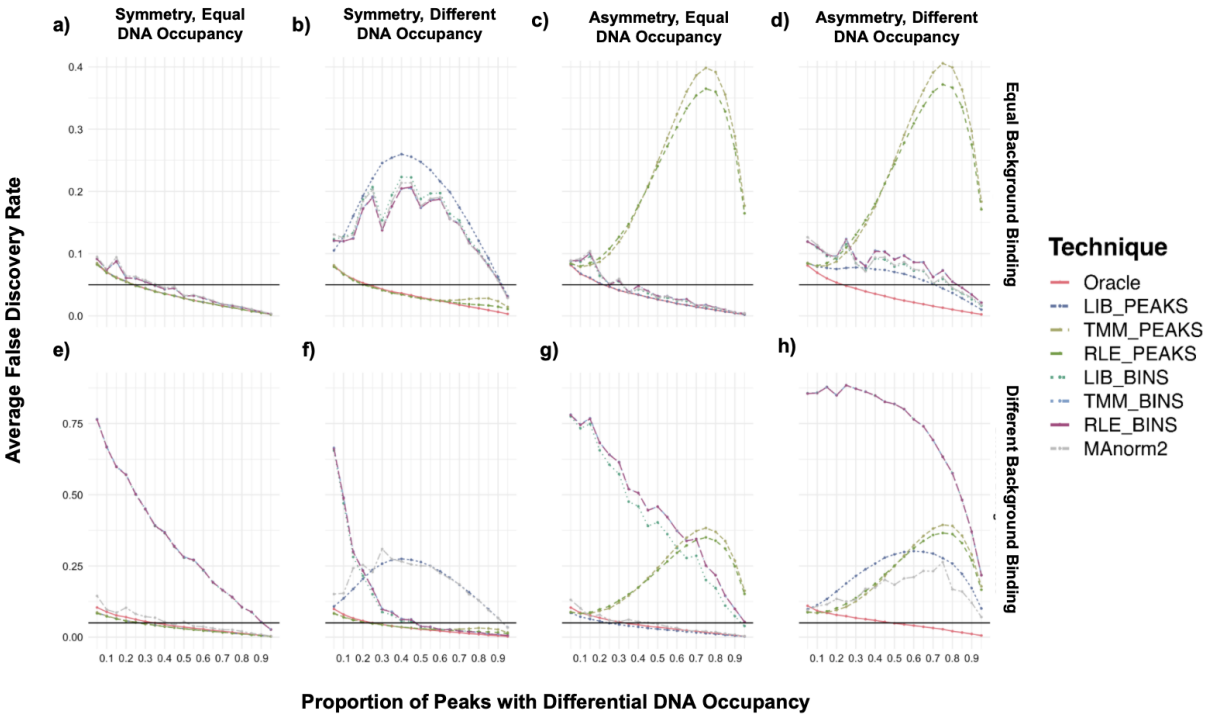

Figure 6.1: **Average false discovery rate under different simulation conditions.** The average false discovery rate for each normalization factor, faceted by simulation condition. The horizontal axis is the proportion of peaks that have differential DNA occupancy, and the vertical axis is the average false discovery rate. Note that the scale of the vertical axis changes between the subplots in the top and bottom rows. Each plot has a black horizontal line at 0.05. When a curve is below that black line, then the associated normalization method (on average) successfully controls the false discovery rate at a level of 0.05 for the given proportion of peaks with differential DNA occupancy. We consider normalization methods that track with the Oracle (the solid red line) to be doing well with respect to controlling the average false discovery rate.

In Figure 6.1, we see that when all technical conditions are met (Plot **(a)**), all normalization methods track closely with the Oracle and control the average false discovery rate at a level of 0.05. However, once we violate some of the technical conditions, some normalization methods start to diverge from the Oracle. For example, in Plot **(c)** of Figure 6.1, the average false discovery rate becomes much higher for TMM (Reads in Peaks) and RLE (Reads in Peaks) than the Oracle and the 0.05 cutoff as the proportion of peaks with differential DNA occupancy increases. Crucially, TMM (Reads in Peaks) and RLE (Reads in Peaks) perform poorly with respect to controlling the average false discovery rate whenever asymmetric differential DNA

occupancy is violated (see also Plots **(d), (g), (h)** in Figure 6.1), and especially so when there is a high proportion of peaks with differential DNA occupancy. Thus, our simulation results indicate that both TMM (Reads in Peaks) and RLE (Reads in Peaks) do a poor job of controlling the average false discovery rate when symmetric DNA occupancy is violated, confirming that TMM (Reads in Peaks) and RLE (Reads in Peaks) rely on the technical condition that there is a symmetric amount of differential DNA occupancy across experimental states.

Moreover, when there is a different total amount of total DNA occupancy across experimental states (see Plots **(b), (d), (f), (h)** of Figure 6.1), several methods do a poor job of controlling the average false discovery rate. In particular, Library Size (Reads in Peaks), MAnorm2, Library Size (Background Bins), RLE (Background Bins), and TMM (Background Bins) all have an average false discovery rate that is much higher than the 0.05 threshold, and the Oracle's. These results confirm that Library Size (Reads in Peaks), Library Size (Background Bins), and MAnorm2 all rely on the technical condition that there is an equal amount of total DNA occupancy across the experimental states.

Further, in virtue of how bins are defined (see Figure 4.2), a substantial difference in the total amount of DNA occupancy in the peaks can lead to the same background bin having different total amounts of DNA binding across the experimental states. This explains why we see that TMM (Background Bins) and RLE (Background Bins) also perform poorly when there is a difference in total DNA occupancy even when there is not a difference in the total amount of background binding across experimental states (e.g., in Plot **(b)** of Figure 6.1). Additionally, TMM (Background Bins), RLE (Background Bins), and Library (Background Bins) also do a poor job of controlling the average false discovery rate, relative to the 0.05 cut-off as well as the Oracle when there is a different amount of background binding across conditions (see Plots **(e), (f), (g), (h)** in Figure 6.1). Hence, our simulation results validate that Library Size (Background Bins), TMM (Background Bins), and RLE (Background Bins) rely on the technical condition that there is an equal amount of total background binding across the experimental states.

However, it is also important to examine how violating the primary technical conditions underlying ChIP-Seq normalization methods impacts whether peaks with differential DNA occupancy are classified as differentially bound in downstream analysis. Thus, we present Figure 6.2 to investigate the average power associated with ChIP-Seq normalization methods as we vary the proportion of peaks with differential DNA occupancy.

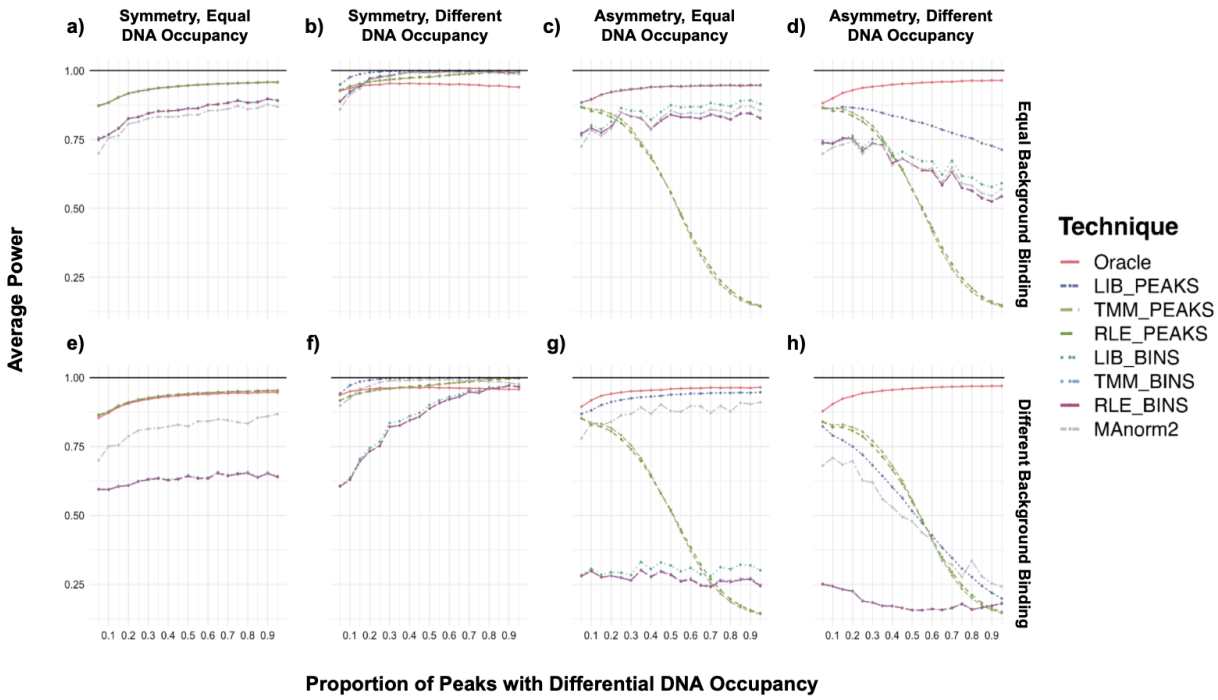

Figure 6.2: **Average power under different simulation conditions.** The average power for each normalization method, faceted by simulation condition. The horizontal axis is the proportion of peaks with differential DNA occupancy, and the vertical axis is the average power. Note that the highest power possible is 1, which only occurs when all peaks with differential DNA occupancy are classified as differentially bound. Normalization methods with higher power are considered to be performing better.

By comparing Figures 6.1 and 6.2, we see that for certain normalization methods, while the average false discovery rate increases as the proportion of peaks with differential DNA occupancy increases, the average power *decreases*. For example, when there is asymmetric differential DNA occupancy across the experimental states, the average power associated with TMM (Reads in Peaks) and RLE (Reads in Peaks) drops as the proportion of peaks with differential DNA occupancy increases (see Plots **(c), (d), (g), (f)** in Figure 6.2). Moreover, RLE (Background Bins), TMM (Background Bins), and Library (Background Bins) have consistently lower average power than the Oracle in Plots **(e), (g), (h)** of Figure 6.2 when there is a difference in background binding, except when there is symmetry and a different total amount of DNA occupancy. These results indicate that normalization by background generally leads to a low proportion of the peaks with differential DNA occupancy being classified as differentially bound when the total background binding is different between the experimental states.

While Figures 6.1 and 6.2 help us understand how the average false discovery rate and power vary for different normalization methods across all of the simulation conditions, we can better see the effects of violating the technical conditions on individual normalization methods by looking at the different metrics one simulation condition at a time. Thus, we present our simulation results, only looking at one simulation condition at a time in Sections 6.2, 6.3, 6.4, and 6.5.

6.2 Simulation Results: All Technical Conditions Met

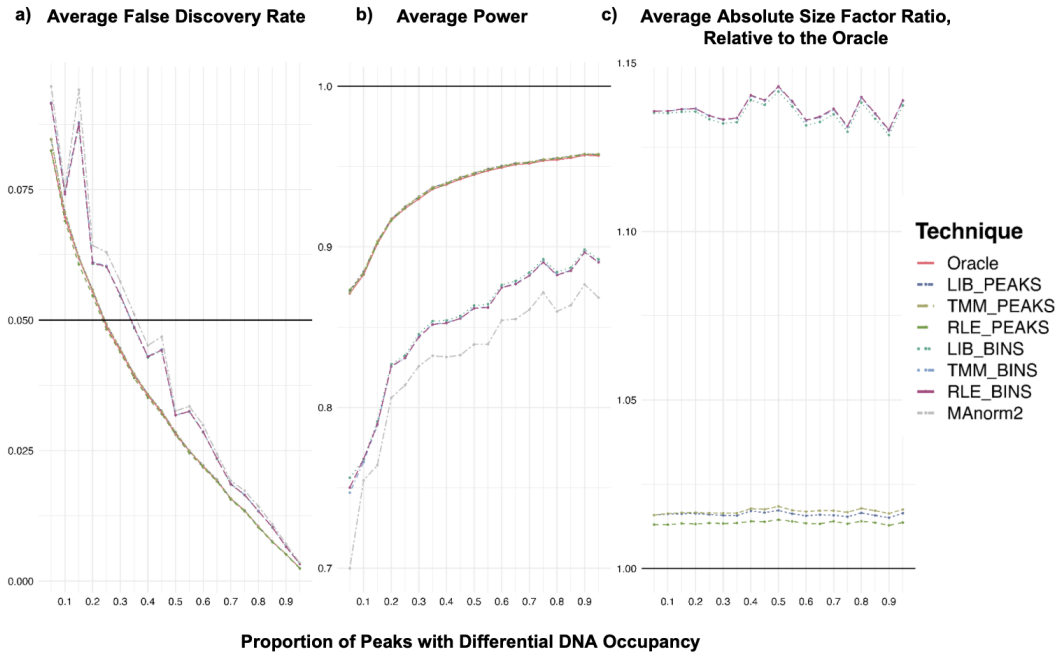

Figure 6.3: **Simulation panels where all technical conditions are met.** The horizontal axis is the proportion of peaks with differential DNA occupancy. The vertical axis is the value for the simulation metric. The figure is faceted such that each subplot represents one of the three simulation metrics. Note the scale change between each facet in the figure. Plot **(a)** is the average false discovery rate. The horizontal black line in Plot **(a)** is the 0.05 threshold we set for the FDR. Plot **(b)** is the average power. The horizontal black line at 1 in Plot **(b)** represents the highest average power possible. Plot **(c)** is the average absolute size factor ratio relative to the Oracle. The horizontal line at 1 in Plot **(c)** represents the Oracle's size factor ratio with itself. Thus, a normalization method tracks closely with the Oracle if its curve is close to the horizontal black line at 1 in Plot **(c)**. Recall that MAnorm2 does not have a size factor and, thus, is not plotted in Plot **(c)**.

In Figure 6.3, all the normalization methods track relatively closely with the Oracle with respect to the average false discovery rate (Plot **(a)**). Moreover, while some normalization methods have lower power, on average, than other normalization methods, all normalization methods have an average power above 70% (Plot **(b)**). Interestingly, RLE (Background Bins) and Library Size (Background Bins) have size factors that are very different from the Oracle's, on average (Plot **(c)**). Yet, RLE (Background Bins) and Library Size (Background Bins) still seem to be performing well with respect to the average false discovery rate and power.

6.3 Effect of Asymmetric Differential DNA Occupancy

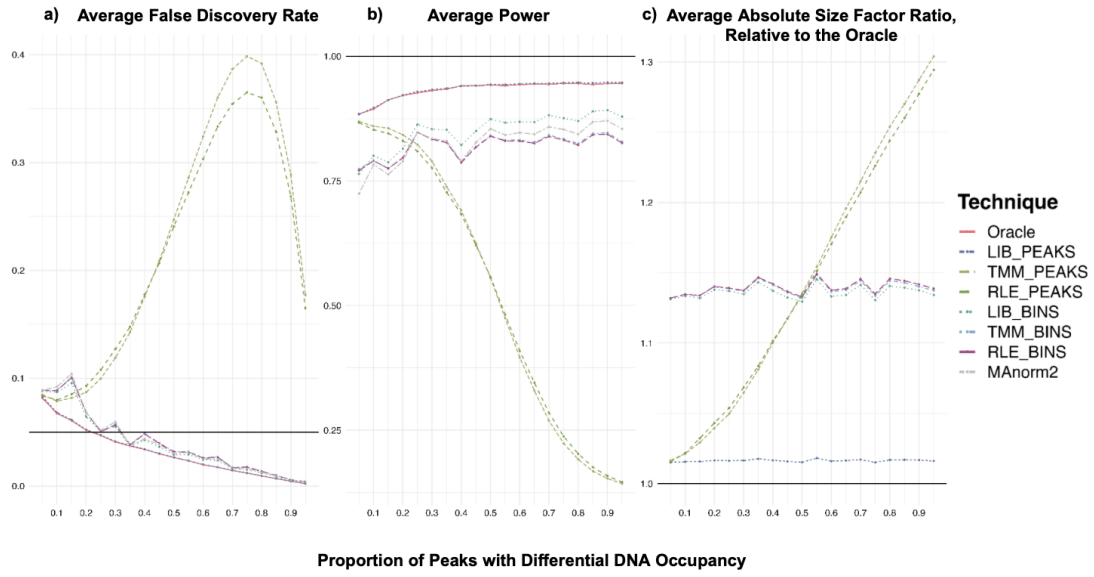

Figure 6.4: **Effect of asymmetric differential DNA occupancy on normalization methods.** The horizontal axis is the proportion of peaks with differential DNA occupancy. The vertical axis is the mean value for the simulation metric. The figure is faceted such that each facet represents one of the three simulation metrics. Note the scale change between each facet in the figure. Plot **(a)** is the average false discovery rate. The horizontal black line in Plot **(a)** is the 0.05 threshold we set for the FDR. Plot **(b)** is the average power, and the horizontal black line at 1 represents the highest average power possible. Plot **(c)** is the average absolute size factor ratio relative to the Oracle. The horizontal line at 1 represents the Oracle's size factor ratio with itself. Thus, a normalization method has a size factor close to the Oracle's if its curve is close to the horizontal black line at 1 in Plot **(c)**. Recall that MAnorm2 does not have a size factor, and thus, its curve is not present in Plot **(c)**.

Figure 6.4 depicts the average false discovery rate, average power, and average absolute size factor ratio relative to the Oracle when only the technical condition of symmetric differential DNA occupancy across experimental states is violated. As noted in Section 6.1, RLE (Reads in Peaks) and TMM (Reads in Peaks) do a poor job of controlling the average false discovery rate relative to the 0.05 cut-off as well as the Oracle, especially when there is a high proportion of peaks with differential DNA occupancy (see Plot **(a)** of Figure 6.4). Moreover, the average power associated with RLE (Reads in Peaks) and TMM (Reads in Peaks) decreases to below 25% as the proportion of peaks with differential DNA occupancy increases (see Plot **(b)** of Figure 6.4). The size factors associated with RLE (Reads in Peaks) and TMM (Reads in Peaks) also diverge (on average) from the Oracle's as the proportion of peaks with differential DNA occupancy increases (see Plot **(c)** of Figure 6.4). Notably, we see that the average absolute size factor ratios associated with RLE (Reads in Peaks) and TMM (Reads in Peaks) become greater than the average absolute size factor ratios associated with other normalization methods when the proportion of peaks with differential DNA occupancy passes 50%. When more than 50% of the peaks have differential DNA occupancy, we expect the median (and mean) peak to have differential DNA occupancy. Thus, given TMM (Reads in Peaks) and RLE (Reads in Peaks) both rely on the central peak not having differential DNA occupancy to properly normalize the read counts, it makes sense that average absolute size factor ratios associated with TMM (Reads in Peaks) and RLE (Reads in Peaks) become greater than the average absolute size factor ratios associated with other normalization methods when the proportion of peaks with differential DNA occupancy passes 50%.

6.4 Effect of Different DNA Occupancy

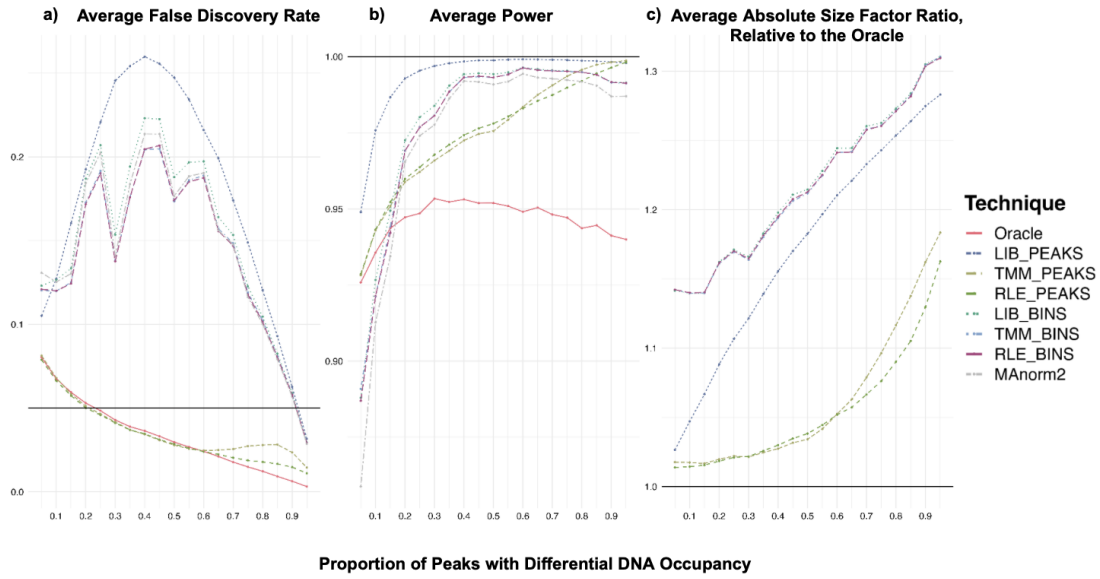

Figure 6.5: **Effect of different total DNA occupancy on normalization methods.** The horizontal axis is the proportion of peaks with differential DNA occupancy. The vertical axis is the average value for the simulation metric. The figure is faceted such that each facet represents one simulation metric. Note the scale change between each facet in the figure. Plot **(a)** is the average false discovery rate. The horizontal black line in Plot **(a)** is the 0.05 threshold we set for the FDR. Plot **(b)** is the average power with the horizontal black line at 1 representing the highest possible average power. Plot **(c)** is the average absolute size factor ratio relative to the Oracle. The horizontal line at 1 in Plot **(c)** represents the Oracle's size factor ratio with itself. Thus, a normalization method has a size factor close to the Oracle's if its curve is close to the horizontal black line at 1 in Plot **(c)**. Recall that MAnorm2 does not have a size factor, and thus, its curve is not present in Plot **(c)**.

Figure 6.5 depicts the average false discovery rate, power, and average absolute size factor ratio relative to the Oracle when only the technical condition of an equal amount of total DNA occupancy across the experimental states is violated. As discussed in Section 6.1, several methods perform poorly when an equal amount of total DNA occupancy across experimental states is violated. As seen in Plot **(a)** of Figure 6.5, the normalization methods that perform poorly with respect to the average false discovery rate are TMM (Background Bins), Library Size (Background Bins), RLE (Background Bins), Library Size (Reads in Peaks), and MAnorm2. Despite these normalization methods performing poorly with respect to the average false discovery rate, all the normalization methods have an average power above 85% (see Plot

(b) of Figure 6.5). In other words, the normalization methods all result in a large proportion of the peaks with differential DNA occupancy being classified as differentially bound. The average absolute size factor ratio relative to the Oracle increases for all the normalization methods as the proportion of peaks with differential DNA occupancy increases (see Plot **(c)** of Figure 6.5). However, TMM (Background Bins) and RLE (Background Bins) uniformly have the highest average absolute size factor relative to the Oracle.

6.5 Effect of Different Background Binding

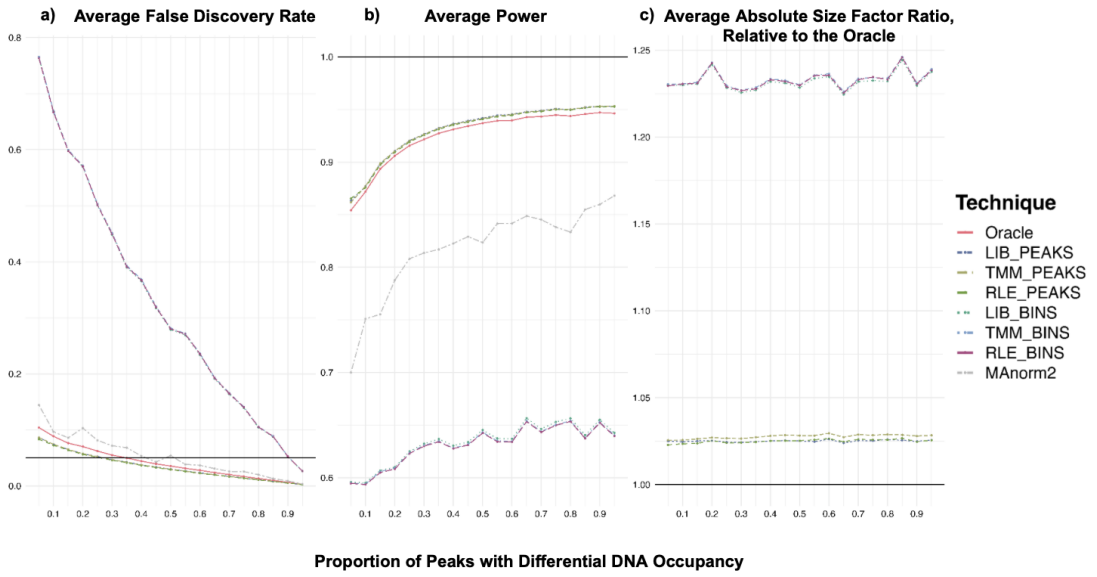

Figure 6.6: **Effect of different background binding on normalization methods.** The horizontal axis is the proportion of peaks with differential DNA occupancy. The vertical axis is the average value for the simulation metric. The figure is faceted such that each facet represents one simulation metric. Note the scale change between each facet in the figure. Plot **(a)** is the average false discovery rate. The horizontal black line in Plot **(a)** is the 0.05 threshold we set for the FDR. Plot **(b)** is the average power with the horizontal black line at 1 representing the highest average power possible. Plot **(c)** is the average absolute size factor ratio relative to the Oracle. The horizontal line at 1 in Plot **(c)** represents the Oracle's size factor ratio with itself. Thus, a normalization method has a size factor close to the Oracle's if its curve is close to the horizontal black line at 1. Recall that MAnorm2 does not have a size factor, and thus, its curve is not present in Plot **(c)**.

Figure 6.6 depicts the average false discovery rate, average power, and average absolute size factor ratio relative to the Oracle when only the technical condition of an equal amount of background binding across experimental states is violated. TMM (Background Bins) and RLE (Background Bins) perform the worst across all three metrics when there is a different amount of total background binding across the experimental states. That is, TMM (Background Bins) and RLE (Background Bins) have the highest average false discovery rate, lowest power, and highest average absolute size factor ratio relative to the Oracle in Plots **(a), (b), (c)** of Figure 6.6. Notably, we see in Plot **(c)** of Figure 6.6 that the average absolute size factor ratio relative to the Oracle remains constant for TMM (Background Bins) and RLE (Background Bins) across different proportions of peaks with differential DNA occupancy. Given that the average absolute size factor ratio relative to the Oracle remains constant as the proportion of peaks with differential DNA occupancy increases indicates that the average false discovery rate is decreasing for TMM (Background Bins) and RLE (Background Bins) in virtue of there being fewer peaks to falsely discover as differentially bound rather than TMM (Background Bins) and RLE (Background Bins) performing better at a higher proportion of peaks with differential DNA binding when an equal amount of total background binding is violated.

Chapter 7

Conclusion

Data collected via chromatin immunoprecipitation followed by high throughput sequencing (i.e., ChIP-Seq) provides vital insights into locations on the genome where there are differences in DNA occupancy between experimental states (i.e., differential DNA occupancy) (Wu et al., 2015). However, since ChIP-Seq data is collected experimentally, differences in DNA binding can arise due to experimental artifacts or random chance rather than differential DNA occupancy. Hence, ChIP-Seq data must be normalized to accurately assess which genomic regions have differential DNA occupancy. While normalization is an essential step in identifying genomic regions with differential DNA occupancy, there is a dearth of literature dedicated to examining the primary technical conditions that underlie ChIP-Seq between-sample normalization methods.

In this thesis, we identified three primary technical conditions that ChIP-Seq between-sample normalization methods rely upon, (1) symmetric differential DNA occupancy, (2) equal amount of total DNA occupancy, and (3) equal amount of total background binding. We then categorized various ChIP-Seq normalization methods by their technical conditions. A major contribution of this thesis is our ChIP-Seq read count simulation, where we examine the effects of violating different combinations of the primary technical conditions on ChIP-Seq normalization methods to corroborate our categorization. In particular, we examined how the average false discovery rate, power, and average absolute size factor relative to the Oracle changed for the various normalization methods when we violated different combinations of the primary technical conditions and varied the proportion of peaks with differential DNA occupancy.

Our simulation results (presented in Chapter 6) underscore how the choice of ChIP-Seq normalization method impacts the biological conclusions drawn from the read count data. For example, if a researcher were to use TMM (Reads in Peaks) when there is asymmetric differential DNA occupancy, then a large proportion of the peaks that are classified as differentially bound in their analysis would have no differential DNA occupancy across the experimental states (see Plot **(a)** of Figure 6.4). Thus, they might falsely conclude that there is differential DNA occupancy in various

genomic regions even when there is not. Therefore, we suggest that researchers use their prior understanding of the experiment at hand to guide their choice of ChIP-Seq normalization method.

A possible extension to our analysis would be to compare various ChIP-Seq normalization methods on real ChIP-Seq read count data that has spiked-in peaks. Using real ChIP-Seq read count data with spiked-in peaks would enable us to compare normalization with spike-ins to other ChIP-Seq normalization methods, which was not possible in this thesis' analysis in virtue of how we simulated ChIP-Seq read counts. Another direction of future work could be to examine the technical conditions underlying other ChIP-seq normalization methods that we did not consider in our analysis, such as CisGenome (Ji et al., 2011), CCAT (Xu et al., 2010), ChIPIN (Polit et al., 2021), and NCIS (Liang and Keleş, 2012). We hope that our thesis provides a framework for such analysis.

Chapter 8

Supplementary Materials

8.1 ChIP-Seq Simulation Code

The code we created to simulate ChIP-Seq read count data and generate figures presented in Chapter 6 can be accessed through our GitHub repository. We would like to thank Lun and Smyth (2016) and Evans et al. (2016), whose publicly available code helped us create the ChIP-Seq read count simulation provided in our GitHub repository.

8.2 Confidence Intervals for Simulation Results

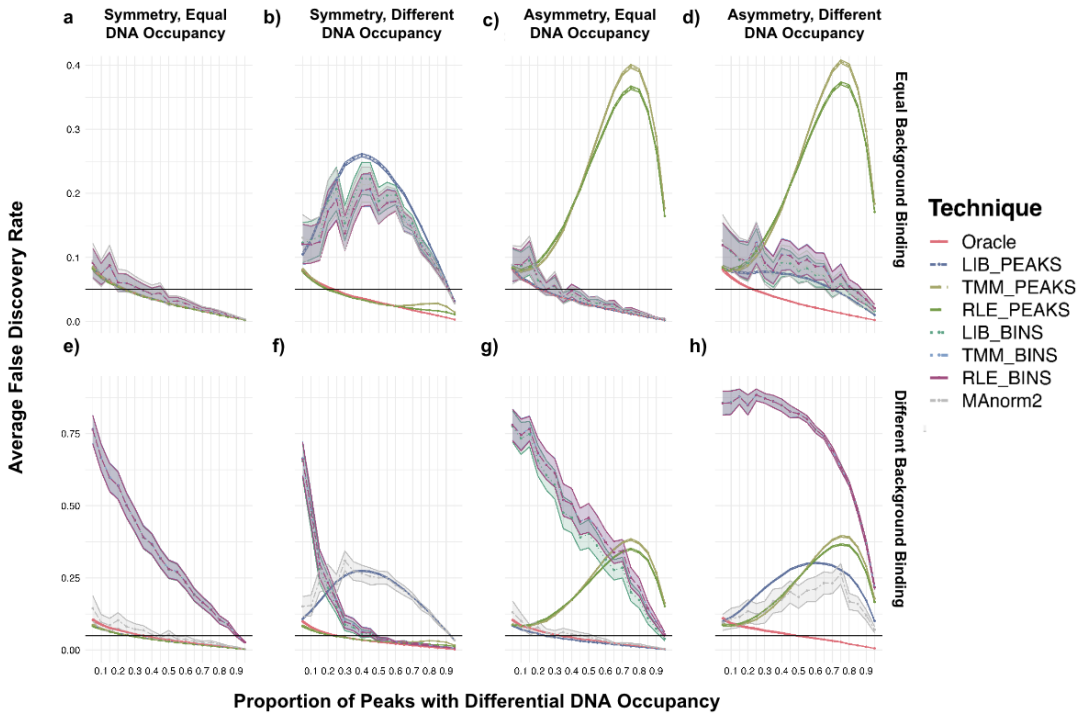

Figure 8.1: **95% confidence intervals for the average false discovery rate.** 95% Confidence Intervals for the average false discovery rate associated with each normalization method, faceted by simulation condition. The horizontal axis is the proportion of peaks that have differential DNA occupancy, and the vertical axis is the average false discovery rate. Each plot has a black horizontal line at 0.05, which is the cut-off we specified for the false discovery rate. The shaded regions denote the 95% confidence interval for the average false discovery rate associated with a given normalization method at each proportion of peaks with differential DNA occupancy. The 95% confidence interval for the mean false discovery rate is calculated as follows: $\overline{FDR} \pm 2 \cdot \frac{sd(FDR)}{\sqrt{100}}$.

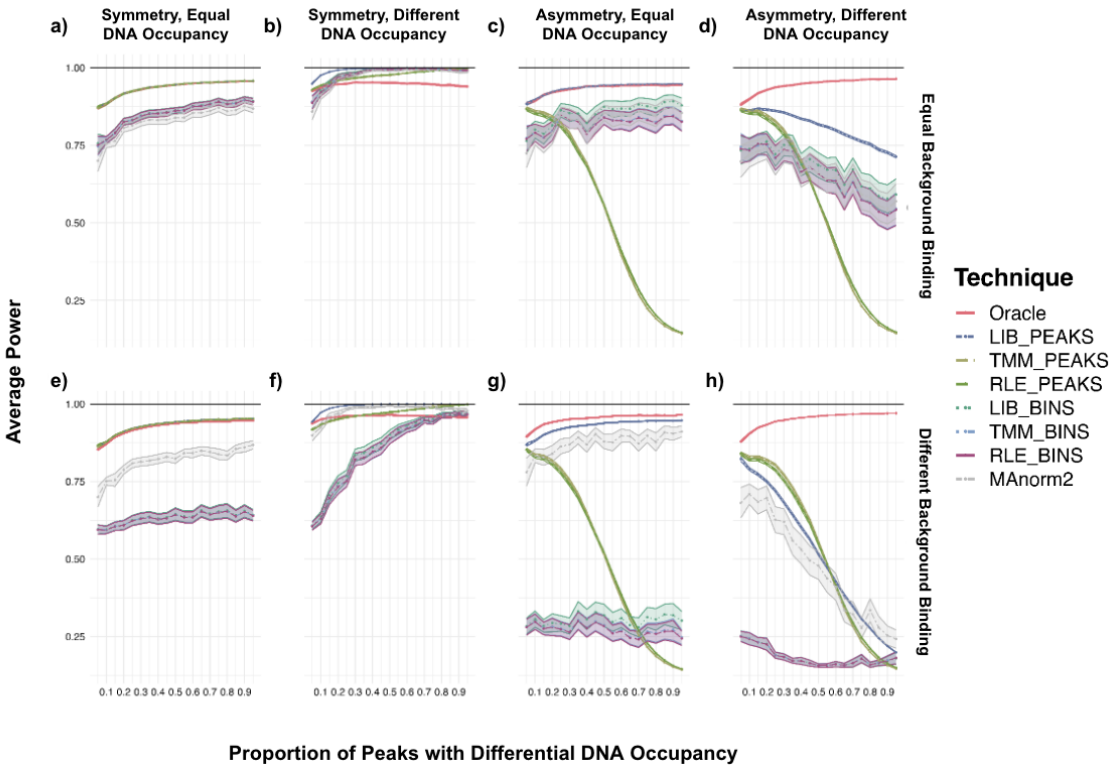

Figure 8.2: **95% confidence intervals for the average power.** 95% Confidence Intervals for the average power associated with each normalization method, faceted by simulation condition. The horizontal axis is the proportion of peaks that have differential DNA occupancy, and the vertical axis is the average power. Each plot has a black horizontal line at 1, which denotes the highest possible average power. The shaded regions denote the 95% confidence interval for the average power associated with the given normalization method at each proportion of peaks with differential DNA occupancy. The 95% confidence interval for the mean power is calculated as follows: $\overline{\text{Power}} \pm 2 \cdot \frac{sd(\text{Power})}{\sqrt{100}}$

8.3 The Benjamini-Hochberg Method

The Benjamini-Hochberg method seeks to control the false discovery rate across multiple comparisons (Benjamini and Hochberg, 1995). Let $p_1, p_2, \ldots, p_k$ be the (raw) p-values across k hypothesis tests. The p-values across the k hypothesis tests are then ranked, with the lowest p-value receiving the highest rank. Let $p_{(1)}, p_{(2)}, \ldots, p_{(k)}$ be the p-values ordered by rank, and i denote the largest p-value rank where the following equation holds:

$$p_{(i)} \leq \alpha \cdot \frac{i}{k} \tag{8.1}$$

where α is the false discovery rate that we want to maintain across the k hypothesis tests.

Then, the p-values $p_{(1)}, \ldots, p_{(i)}$ are considered *statistically significant* under the Benjamini-Hochberg method at the false discovery rate threshold of α (Benjamini and Hochberg, 1995).

References

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

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. (2013). Practical guidelines for the comprehensive analysis of chip-seq data. *PLoS computational biology*, 9.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1).

Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., Willis, I. M., Herr, W., Hernandez, N., Delorenzi, M., and Consortium, C. (2014). Quantifying chip-seq data: a spiking method providing an internal reference for sample-to-sample normalization. *Genome Res.*, 7.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Gall, C., Schaëffer, B., Le Crom, S., Guedj, M., and Jaffrézic, F. (2012). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14.

Evans, C., Hardin, J., and Stoebel, D. (2016). Selecting between-sample rna-seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics*, 19.

Höllbacher, B., Balázs, K., Heinig, M., and Uhlenhaut, H. (2020). Seq-ing answers: Current data integration approaches to uncover mechanisms of transcriptional regulation. *Computational and Structural Biotechnology Journal*, 18.

Ji, H., Jiang, H., Ma, W., and Wong, W. H. (2011). Using cisgenome to analyze chip-chip and chip-seq data. *Current Protocols in Bioinformatics*, 33(1).

Kotipalli, A., Banerjee, R., Kasibhatla, S., and Joshi, R. (2021). Analysis of h3k4me3-chip-seq and rna-seq data to understand the putative role of mirnas and their target genes in breast cancer cell lines. *Genomics & Informatics*, volume 19.

Lee, J.-Y. (2023). The principles and applications of high-throughput sequencing technologies. *Development & Reproduction*, 27.

Liang, K. and Keleş, S. (2012). Normalization of ChIP-seq data with control. BMC Bioinformatics*, 13.*

Lodhi, N. and Tulin, A. V. (2011). PARP1 Genomics: Chromatin Immunoprecipitation Approach Using Anti-PARP1 Antibody (ChIP and ChIP-seq)*. Humana Press.*

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biology*, 15.*

Lun, A. and Smyth, G. (2016). From reads to regions: A bioconductor workflow to detect differential binding in chip-seq data. F1000Research*, 4.*

Nakato, R. and Sakata, T. (2020). Methods for chip-seq analysis: A practical workflow and advanced applications. Methods*, 187.*

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

Polit, L., Kerdivel, G., Gregoricchio, S., Esposito, M., Guillouf, C., and Boeva, V. (2021). CHIPIN: Chip-seq inter-sample normalization based on signal invariance across transcriptionally constant genes. BMC Bioinformatics*, 22(1).*

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

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

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

Tu, S., Li, M., Haojie, C., Tan, F., Xu, J., Waxman, D., Zhang, Y., and Shao, Z. (2020). Manorm2 for quantitatively comparing groups of chip-seq samples. Genome Research*, 31.*

Wu, D.-Y., Bittencourt, D., Stallcup, M. R., and Siegmund, K. D. (2015). Identifying differential transcription factor binding in chip-seq. Frontiers in Genetics*, 6.*

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

Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., Nusbaum, C., Myers, R., Brown, M., Li, W., and Liu, S. (2008). Model-based analysis of chip-seq (macs). Genome biology*, 9.*