Statistical analysis of high-throughput sequencing count data

Michael I. Love

June 2013

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Gutachter:
Prof. Dr. Martin Vingron
PD. Dr. Peter N. Robinson

1. Referent: Prof. Dr. Martin Vingron

2. Referent: PD. Dr. Peter N. Robinson

Tag der Promotion: 24.9.2013

Preface

All of the work presented in this thesis grew out of collaborations with other researchers. For each chapter, I briefly summarize my contribution and acknowledge the contributions of others.

Chapter 2 represents a conceptual framework for modeling read counts using various distributions. These ideas grew out of conversations with Ho-Ryun Chung at the Max Planck Institute for Molecular Genetics (MPIMG) in Berlin and Simon Anders at the European Molecular Biology Laboratories (EMBL) in Heidelberg.

Chapter 3 was published in *Statistical Applications in Genetics and Molecular Biology* [1]. The idea for detecting copy number variants in exome-enriched sequencing data was proposed by Stefan Haas and with Alena van Bömmel various methods were tested and evaluated. My contribution was developing the hidden Markov model, implementing the software and testing the performance. I wish to acknowledge the X-linked intellectual disabilities project team at MPIMG including H.-Hilger Ropers, Vera Kalscheuer, Ruping Sun, Anne-Katrin Emde, Wei Chen, Hao Hu and Tomasz Zemojtel, who provided helpful discussions.

Chapter 4 resulted from a 5 month visit to the group of Wolfgang Huber at EMBL in Heidelberg. Simon Anders proposed the idea of incorporating priors for dispersion and log fold change into the *DESeq* framework. My contribution was to implement these new statistical methods as a new package *DESeq2*, with closer integration with core Bioconductor packages. I would like to acknowledge all the members of the Huber group for helpful discussions.

Chapter 5 resulted from a collaboration with the Transcriptional Regulation Group of Sebastiaan Meijsing at the MPIMG. I would like to thank Stephan Starick who initially proposed to investigate the interaction between glucocorticoid receptor and the chromatin landscape. My contribution was the statistical analysis presented in the chapter. Sebastiaan Meijsing provided valuable feedback during the evolution of the project. I wish to acknowledge the contributions of Morgane Thomas-Chollier, Katja Borzym, Sam Cooper and Ho-Ryun Chung.

Acknowledgments

I would like to thank my supervisors Martin Vingron, Stefan Haas and Knut Reinert for their advice and support during my PhD. I am grateful to Martin Vingron for allowing me the freedom in my research topics and for providing such stimulating research environment at the MPIMG in Berlin. I am grateful to Stefan Haas for his valuable biological insight and for helping me through many revisions of manuscripts. I am grateful to the following people for proofreading chapters of my thesis: Stefan Haas, Peter Arndt, Sebastiaan Meijsing and Ruping Sun, and to Johannes Helmuth, Stefan Haas and Juliane Perner for helping me write the Zusammenfassung. I wish to thank Peter N. Robinson for agreeing on short notice to be one of the thesis readers and to be part of my thesis committee.

I wish to thank Kirsten Kelleher and Hannes Luz for helping me get settled in Berlin, and Kirsten for assistance along the steps of the PhD and for creating a friendly atmosphere in the IMPRS program. I thank the entire Computational Molecular Biology group at MPIMG and the Algorithmic Bioinformatics group at Freie Universität for many interesting conversations and lots of helpful feedback. I would like to thank my office mates Ruping Sun, Ho-Ryun Chung, Juliane Perner, and Stefanie Schöne for the interesting scientific conversations. I would like to thank Wolfgang Huber for allowing my visit to his group, Simon Anders for many fruitful discussions, and the entire Huber group for making my time in Heidelberg so pleasant.

Finally, I want to thank my parents Sue and Cliff, my brother David, and my wife Zuzka for their unending support throughout my thesis. Especially Zuzka who had to listen to many rambling thoughts about DNA and the cell.

Michael I. Love Berlin, June 2013

Contents

Chapter 1

Introduction

1.1 Biological introduction

1.1.1 DNA

DNA is referred to as "the blueprint for life", as it is the one molecule one could extract from an organism and possibly produce a nearly identical copy of that organism. For example, in 1958, when the idea that DNA was the molecular basis of heredity was still quite new, Gurdon et al. [2] successfully cloned a frog using only the nucleus (and therefore the nucleic acids) of an adult cell. DNA can be roughly divided into two functional groups: genes and regulatory elements. Genes are regions of DNA which are "transcribed" by an enzyme, RNA polymerase, into RNA molecules, some of which will be used to make proteins (messenger RNA, or "mRNA") and others which serve enzymatic roles as RNAs alone (such as those involved in the ribosome, the cell's protein synthesizing machinery). Regulatory elements, such as "promoters" and "enhancers" are stretches of DNA where regulatory proteins called "transcription factors" can bind and then influence the transcription of nearby genes. Promoters are regions near the transcription start sites (TSS) of genes and enhancers are typically further away from the TSS. The probability of binding is determined by the match of the shape of the protein with the particular sequence of DNA, as well as the presence of other proteins or molecules bound to DNA nearby. Recently, large-scale efforts have been made to identify all of the genes and regulatory elements of the human genome, by assaying transcription and protein binding across many different cell types [3]. This identification effort is critical for the fields of molecular biology and medicine, because variation in the DNA sequence of both genes and regulatory elements contributes to an organism's phenotype, including the propensity to suffer from diseases.

1.1.2 RNA

The transcription of DNA into mRNA is the first step in the central dogma of molecular biology: DNA $\rightarrow$ mRNA (transcription) and mRNA $\rightarrow$ protein (translation). When a gene is transcribed into mRNA, one strand of DNA, the "template" strand is used to make a complementary copy of RNA (with uracil taking the place of thymine). The other DNA strand is referred to as the "coding" strand as it contains the same sequence as the mRNA molecule. Information about the mRNA transcript abundance in the cell, or mRNA "expression", is valuable, as it can be used to classify cells and tissues into different states, such as those coming from healthy or diseased tissue [4–6]. Furthermore, sets of

genes with similar function can be discovered by comparison of mRNA expression across various conditions [7, 8]. mRNA is not a direct copy of stretches of DNA, but certain regions called "introns" (a concatenation of "intragenic regions") are removed from the mRNA transcript during a process called "splicing", leaving the remaining pieces which are called "exons". Alternative transcripts can be formed from different combinations of exons of a single gene. The frequency with which exons of a gene are included in mRNA transcripts varies between cells of different tissues, as well as between healthy and diseased tissue [9].

1.1.3 Chromatin

The cell nucleus of multicellular organisms does not contain naked DNA, but instead DNA is typically wrapped around protein complexes called nucleosomes. The DNA, nucleosomes, and other proteins attached to DNA is collectively referred to as "chromatin" (where the "chroma-" refers to the ability of chromatin to be colored by dyes). In 1973, researchers first observed that chromatin consists of a repeating pattern of $\sim$ 200 base pairs of DNA wrapped around a nucleosome [10]. Nucleosomes themselves are composed of 8 proteins called histones. Nucleosomes serve many roles in the cell, with perhaps the most important role being the packaging of long molecules of DNA into tight coils (called a "30 nm fiber"), and coils of coils. Only through this packaging is it possible to contain 6 billion base pairs of DNA (which stretched out would be around 2 meters) in the cell nucleus (with a diameter on the order of micrometers).

Nucleosomes can also help direct regulatory proteins to appropriate regulatory elements, such as promoters and enhancers, mediated through small molecular signals on the N-terminal tails of the histone proteins which make up the nucleosome. Modifications such as the addition or removal of methyl or acetyl groups to the histone tails can be performed by certain enzymes and recognized by other proteins. The histone-modification-recognizing proteins can then help attract or stabilize transcription factors to certain regulatory elements, resulting in increased or decreased transcription rates of nearby genes.

In addition to the modifications mentioned above, the presence or absence of nucleosomes themselves can influence whether a protein can access the DNA at a regulatory element. This property is referred to as "chromatin accessibility". Regulatory elements which are in use, such as the promoters of actively transcribed genes, are particularly accessible. Due to their functional relevance, researchers are therefore eager to assay histone modifications and chromatin accessibility along the genome in various tissues of the body and in different disease states.

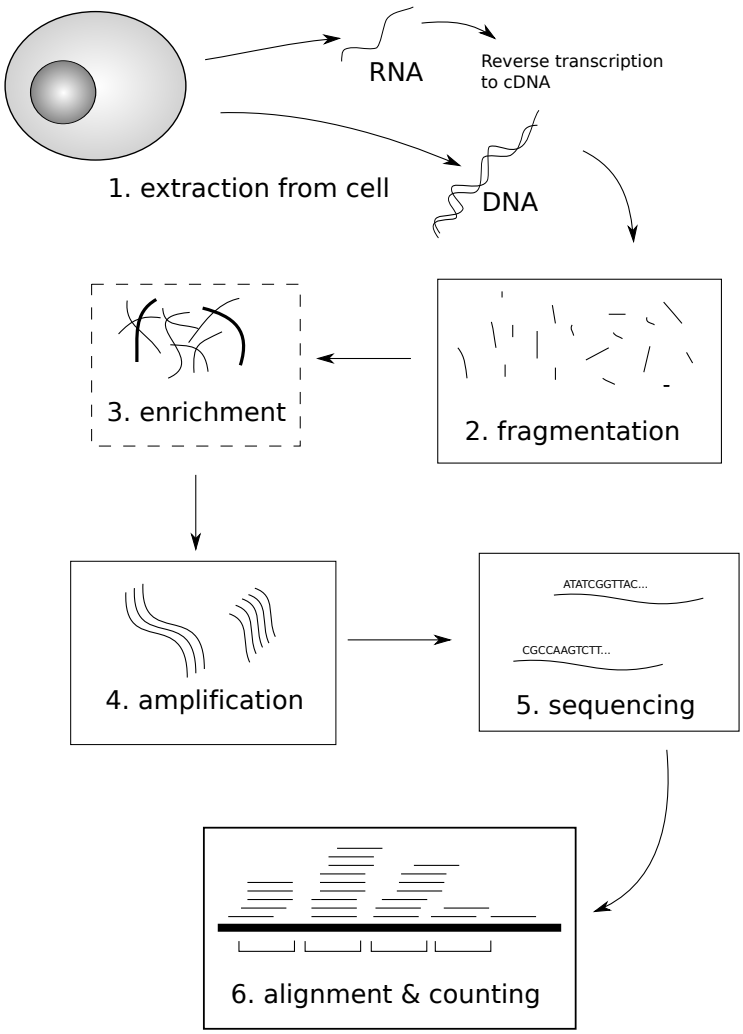

Figure 1.1: Diagram of a high-throughput sequencing protocol. In step 1, DNA (or RNA) is extracted from the nucleus of the cell. In step 2, DNA is fragmented (sheared or sonicated) into smaller fragments. In step 3, certain DNA fragments might be "enriched" over others, meaning that they are selected out of the pool and carried to the next step. For example, "exome enrichment" selects for fragments covering the exons. The box for step 3 is dashed as it is optional. In step 4, the fragments are duplicated at varying effficiency through several cycles of polymerase chain reaction (PCR). This PCR step may introduce bias in the abundance of fragments, based on the DNA sequence composition. In step 5, the fragments are "sequenced", which refers to the identification (with some error) of the individual base pairs of the fragments using a sequencing machine. Information representing the strings of identified base pairs, referred to as "reads", is then stored in computer memory, often with accompanying quality scores, which give some sense of the confidence in the identification of the correct base pair by the machine. In the final step 6, algorithms are used to align the reads (thin stacked lines) to unique locations in the reference genome (thick line). Reads can then be counted in non-overlapping genomic ranges, depicted here below the reference genome.

1.2 Descriptions of experiments

1.2.1 High-throughput sequencing

The use of high-throughput sequencing in order to assay properties of the cell has grown at a remarkable pace in the past decade, as the cost of sequencing has dropped by more than 4 orders of magnitude.[1]. This technology, which was primarily developed in order to sequence and assemble genomes, has since been extended in combination with other lab protocols to produce a plethora of "*-Seq" protocols, including RNA-Seq, ChIP-Seq, MeDIP-Seq, BS-Seq, DNase-Seq and FAIRE-Seq (reviewed in [11]). In each case, a property of the cell is assayed or quantified using identifiable strings of the reference genome as an addressing system. The remarkable achievement is that, with one experiment, quantitative measurements are made of all regions of the genome at once (except those regions which are not identifiable at the length of the DNA sequenced.) For concrete examples, I introduce the four protocols covered in this thesis, DNA-Seq, RNA-Seq, ChIP-Seq and DNase-Seq, and describe the kind of information which these protocols can provide about the cell.

1.2.2 Genotyping: DNA-Seq

DNA-Seq is a protocol which allows for the detection of variations between the genome of an individual being studied and the reference genome. These variations can include single nucleotide variants (SNV), insertions, deletions, inversions, or even large regions which are duplicated or deleted (copy number variants, or CNV). The DNA-Seq protocol was rapidly developed in order to help complete the assembly of the human genome. While techniques to sequence DNA, such as Sanger sequencing, have existed for more than 30 years, I refer in this thesis to high-throughput sequencing, wherein millions of small fragments of DNA are sequenced simultaneously. Figure 1.1 displays a typical sequencing protocol.

After sequencing, many DNA-Seq reads are aligned to the reference genome, as in step 6 of Figure 1.1. If the individual being sequenced has, at a certain genomic location, a different nucleotide than the reference genome on one of their chromosomes, this will appear in a visualization of the aligned reads as a column with a mix of the reference nucleotide and the alternative nucleotide. Such SNVs might have relevance for disease, especially if they occur in coding genes or regulatory elements [12].

In this thesis, I will focus on detecting copy number variation from DNA-Seq reads. Copy number variants can be detected as regions with higher counts of reads (duplications) or lower counts of reads (deletions) in genomic ranges, compared to neighboring regions or to the read counts from a reference sample [13–15]. This is depicted in the first panel in Figure 1.2. Copy number variants are especially of interest to geneticists if they overlap coding genes, as this could lead to increased or decreased abundance of the protein pro-

[1] http://www.genome.gov/sequencingcosts/

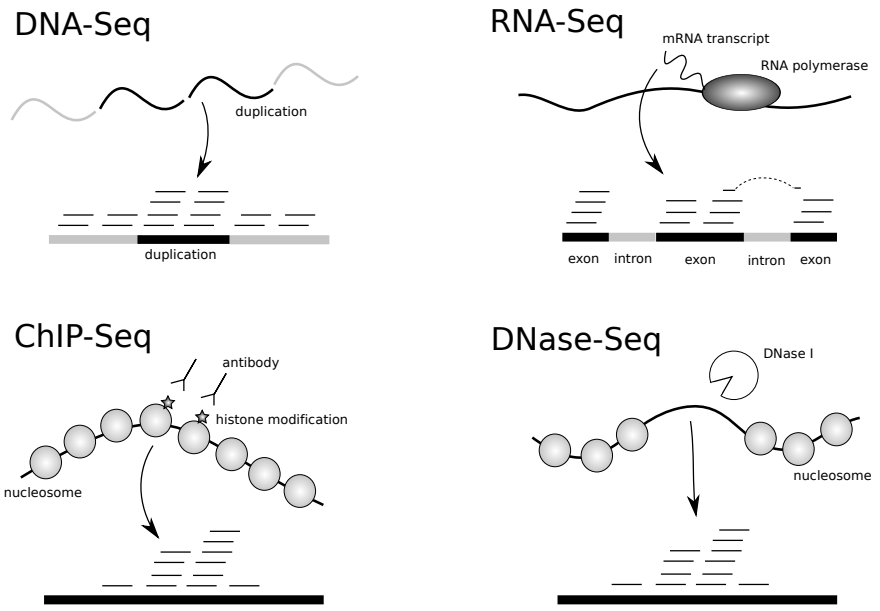

Figure 1.2: Various experiments which can be performed using high-throughput sequencing technology. Below each schematic of the molecules and enzymes is shown how the resulting sequencing reads (thin short lines) would align to the reference genome (thick line). The accumulation of reads represents different kinds of information for each experiment.

duced by that gene. If the gene is truncated or otherwise altered by the copy number variant, this could lead to alterations of the protein.

1.2.3 Gene expression: RNA-Seq

RNA-Seq is used to identify mRNA transcripts, including novel transcripts and transcripts with alternative exons, and to measure the abundance of transcripts [16–18]. There are a few critical differences between the DNA-Seq and RNA-Seq protocols, firstly that the mRNA must be reverse transcribed (using an enzyme called "reverse transcriptase") into cDNA (complementary DNA), so that it can be sequenced. RNA-Seq protocols can be either "unstranded", in which case reads from both the template strand and coding strand of the gene are generated, or "strand-specific" in which case reads align either to the template strand or the coding strand, depending on protocol steps. Secondly, it is common in RNA-Seq to enrich for RNA molecules which end with a long string of adenosines (referred to as a "poly(A) tail") before the reverse transcription. This effectively enriches the resulting pool for mRNA molecules over the highly abundant rRNA (ribosomal RNA) and tRNA (transfer RNA).

It should be noted that only a portion of the reads produced by mRNA transcripts will align easily to the genome; those reads which fall completely within exons (shown in the second panel of Figure 1.2) can be easily aligned. Reads which include the junctions between adjacent exons will either need to be aligned to a transcriptome (the gene regions

with introns removed), or to be aligned using spliced read alignment algorithms [19, 20] which allow for large deletions in the reference genome (the introns). An example of this kind of read is diagrammed in Figure 1.2, with a curved dashed line connecting the two ends which overlap exons.

Once the RNA-Seq reads have been aligned to the genome, potential novel transcripts can be identified as regions with accumulations of reads. Alternative exon usage can be detected when a spliced read alignment assigns one read part to one exon, and then skips over an exon before assigning the other read part to the following exon. Finally, transcript or gene abundance (summarizing over the different transcripts of a gene) can be measured as the number of reads aligning to all the exons of a transcript or gene, depicted in the second panel of Figure 1.2. These read counts can be used to estimate the number of mRNA transcripts in the original cell after normalizing for gene length [17], to compare levels of mRNA transcripts across samples [21–23], and to detect alternative exon usage [23, 24].

1.2.4 Chromatin state: ChIP-Seq and DNase-Seq

Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) is a protocol used to infer the binding affinity of proteins to DNA. ChIP-Seq read counts along the genome can be used to localize protein binding sites and to infer quantitatively how frequently the binding is occurring [25]. In the ChIP-Seq protocol, proteins associated with DNA are crosslinked to the DNA. Antibodies are then used to enrich for DNA fragments associated with proteins of interest. After the enrichment, the crosslinks are reversed and the remaining steps are identical to DNA-Seq. The resulting pattern of sequenced reads is depicted in the third panel of Figure 1.2, where accumulations of reads indicate regions of the genome which were bound by the protein in some of the cells in the sample.

DNase-seq is a protocol used to determine regions of chromatin accessibility along the genome. DNA in regions where nucleosomes are not tightly packed is generally more accessible to proteins, and is therefore preferentially cleaved by endonucleases like DNase I. The regions are therefore referred to as "DNase hypersensitive sites", or DHS. In order to measure chromatin accessibility along the entire genome, chromatin extracted from the nucleus is treated with the DNase I and the resulting fragments are sequenced [26, 27], as depicted in the fourth panel of Figure 1.2. As with ChIP-Seq, the resulting pattern of reads aligned to the reference genome indicates which regions were accessible to proteins in some of the cells in the sample. DNase hypersensitivity, as measured by DNase-seq, has been used to characterize human cell lines, revealing cell-type-specific regulatory elements [28–30].

1.3 Most experiments are ensemble averages

In the high-throughput sequencing assays mentioned above, it is important to keep in mind that the resulting counts of reads typically do not come from individual cells. It is useful to

borrow the concept of the "ensemble average" from statistical mechanics; observations are an average over a period of time, and over a population of cells, possibly heterogeneous.

For an example of time averaging, consider that mRNA transcripts have a lifespan in the cell after they are transcribed and processed. The counts of reads from a typical RNA-Seq experiment therefore provide information about the steady-state population of mRNA transcripts. Unless a specific protocol is used to enrich for nascent transcripts, these counts alone cannot be used to estimate the number of transcripts being produced at a given moment. For an example of population averaging, consider that in a single cell, a section of DNA on a single chromosome is either accessible or not. The number of DNase-Seq reads at a given genomic location therefore provide a continuous measure of the accessibility over a population of cells. It is important to keep in mind that these counts can be used to infer time- and population-averages.

Remarkably, researchers are now moving toward analysis of single cells [31, 32], as well as toward protocols which reduce the biases associated with amplification [33]. If the technical difficulties can be overcome, biologists and bioinformaticians can hope for future datasets where the resulting sequenced read counts are directly proportional to the number of molecules in the original cell. However, even in this case, statistical modeling of read counts will be useful in order to estimate the biological variation across a sample of individual cells.

1.4 Opportunities and challenges of genomic count data

The adoption of high-throughput sequencing machines to assay various properties of the cell has radically changed the kind and scale of information which is available to biologists. The ability to generate genome-wide maps from a single experiment means that experiments need not only be performed to test hypotheses, but can be performed to generate hypotheses. For a bioinformatician or statistician, genome-wide datasets allow for the development of new normalization techniques and new estimators, which harness shared information across many genomic observations and across samples.

A number of challenges also arise in moving to sequencing data as an all-purpose assay. There are known biases affecting the amount of sequencing reads, many of which arise from the amplification step. One well-studied bias is the dependence of read counts on the GC content, the number of G's and C's in the DNA fragment [34]. Another problem with sequence-based assays is the reliance on a reference genome as an addressing system for all reads. Reads which do not align uniquely to one position in the reference genome, so called "ambiguously mapped" reads, present a problem for estimation and hypothesis testing. One solution is to discard these reads, however this comes at the cost of disregarding some portion of the reference genome, the size of which is inversely related to the length of the sequenced reads.

1.5 Thesis objective and structure

1.5.1 Objective

The goal of this thesis is to formulate statistical models for sequence count data, which help to identify signals of interest while accounting for technical artifacts. Across three different experimental protocols, DNA-Seq, RNA-Seq and sequencing-based chromatin assays, discrete distributions are used with a covariate-dependent mean parameter and a parameter relating to the variance. These models all take advantage of the genome-wide scale of experiments by sharing information across genomic ranges in order to answer biologically relevant questions about the cell.

1.5.2 Structure

In Chapter 2, I introduce the statistical framework for the thesis, in particular the discrete distributions, such as the binomial, the Poisson and the negative binomial distributions. I list some advantages to modeling with discrete distributions rather than working with transformed counts or using non-parametric methods.

Chapter 3 describes a hidden Markov model (HMM) for detecting copy number variants in exome-enriched DNA-Seq data. The HMM has negative binomial emission distributions, wherein the hidden state is the copy number state of the sample at a particular genomic range.

Chapter 4 describes a generalized linear model (GLM) for detection of differential gene expression from RNA-Seq read counts. Novel techniques are introduced for sharing information across genes to improve the estimation of dispersion and fold changes.

Chapter 5 describes a hierarchical Bayes model for the associations between transcription factor binding and chromatin and sequence features. The model is constructed in such a way to allow for comparison of these associations across experiments and across cell types.

Chapter 2

Motivation of discrete distributions for sequencing counts

2.1 Multinomial, binomial and Poisson distributions

In this chapter, I motivate the use of various discrete distributions for modeling the counts of reads mapping to a genomic range. In high-throughput sequencing experiments, the raw data are millions of reads which are typically aligned to locations in the genome. Consider K_i, the number of sequencing reads which can be assigned to a particular region i. A sequenced read can be thought of as a draw of a colored ball from an urn, where there are as many colors as regions i and the urn represents a large pool of DNA fragments. The probabilities $\vec{p}$ for drawing a ball of each color from the urn are given by the proportions of DNA fragments arising from each genomic region[1]. If the total number of sequenced reads N is much smaller than the number of DNA fragments, then the vector of counts $\vec{K}$ will follow a multinomial distribution. Technical replicates of the experiment, where one draws repeatedly from the large pool of DNA fragments, should look like independent and identically distributed draws of the multinomial distribution. This presumes that the proportions $\vec{p}$ are not changed by each draw, which is a safe assumption when the number of DNA fragments needed for sequencing is much smaller than the number of DNA fragments in the large pool.

The random varible K_i, when considered alone, is distributed as a binomial random variable, with number of trials N equal to the total number of sequenced reads and success probability p_i, the proportion of fragments in the large pool arising from region i. As the number of sequenced reads becomes large and the probability p_i shrinks, the binomial distribution converges to a Poisson distribution with mean equal to Np_i. As shown in Figure 2.1, these two distributions are very similar already with $N = 100$, and $p_i = 1/10$. In high-throughput sequencing experiments N is typically greater than one million reads. In RNA-Seq experiments, with reads counts for each gene, p_i will be typically less than $1/1000$. This justification for the use of the Poisson distribution is empirically supported by Marioni et al. [18], who show that for technical replicates of RNA-Seq experiments, 99.5% of genes are well approximated by a Poisson distribution. It is suggested that multiple technical replicates can be added together as a single sample for the purpose of RNA-Seq analysis. This recommendation follows from the fact that the sum of independent Poisson random variables is also distributed as a Poisson random variable with a new mean equal to the sum of the individual means.

[1] To be more specific, the probabilities depend on both the proportions of fragments from each genomic region and on the efficiency of the fragment to be successfully sequenced.

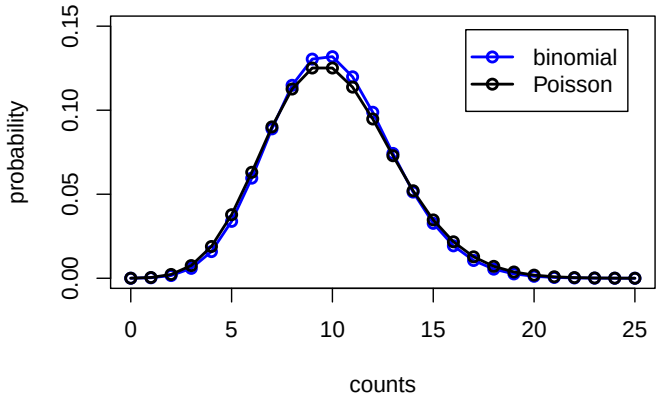

Figure 2.1: The binomial distribution with 100 trials and probability of success 1/10 and the Poisson distribution with mean 10. With a large number of trials and small probabilities, the binomial is well approximated by a Poisson. The discrete probabilities are joined by lines for ease of visualizing multiple distributions.

2.2 Overdispersion and Poisson mixture distributions

2.2.1 Overdispersion

While the idealized experiment described above, with repeated draws from a very large pool of DNA fragments, may be appropriate for the case of technical replicates, these assumptions are not appropriate for "biological replicates". Biological replication of an experiment implies that a new pool of DNA fragments is generated, which will not have an identical probability vector $\vec{p}$ of proportions from various regions of the genome. So while the Poisson approximation of the binomial distribution still holds for an individual sample – as N is still large and p_i small – the expected value for the counts for feature i and sample j, $E(K_{ij})$, will vary for each sample j. A distribution where one of the parameters itself varies according to a distribution is referred to as a compound or mixture distribution[2]. When the variance of a Poisson mixture is greater than the mean, the counts are said to be "overdispersed" with respect to a Poisson distribution. In the following sections I consider two Poisson mixture distributions.

In order to build a Poisson mixture distribution, the distribution for the mean parameter should satisfy the following two properties: (1) the distribution should have support on the non-negative real numbers, as it represents the expected value of a sequence of non-negative integers; (2) the distribution should have at least two parameters in order to specify both the mean and the variance. Property (2) allows the Poisson mixture distribution to include (in the limit) the Poisson distribution, when the variance of the mean parameter goes to zero. This is useful in the case that the biological replicates are actually more like technical replicates, i.e. that the proportions $\vec{p}$ between the pools of fragments are near identical.

[2]In this thesis, I will refer to Poisson mixtures, as the term "compound Poisson" is used to refer to a sum of N random variables, when N is Poisson distributed.

2.2.2 Poisson gamma / negative binomial distribution

One distribution which satisfies the two properties above is the gamma distribution. The Poisson-gamma mixture distribution, a Poisson distribution with a gamma distributed mean parameter, is most commonly referred to as the negative binomial distribution, shown in Figure 2.2. The name "negative binomial" refers to the fact that this distribution arises as well when counting the number of Bernoulli trials which must occur before a certain number of failures have occurred. However, this interpretation as a sequence of trials is not relevant for the context of this thesis. The density for a random variable $K \sim \mathrm{NB}(\mu, \alpha)$, with a mean parameter $\mu > 0$ and a dispersion parameter $\alpha > 0$ is defined by:

$$P(K = k) = \frac{\Gamma(k + 1/\alpha)}{k!\,\Gamma(1/\alpha)} \left(\frac{\mu}{\mu + 1/\alpha}\right)^k (1 + \mu\alpha)^{-1/\alpha} \tag{2.1}$$

In this parametrization, the dispersion parameter α is equal to the inverse of the number of failures, or "size", in the formulation as a sequence of Bernoulli trials. The mean and variance is then given by:

$$\mathrm{E}(K) = \mu, \quad \mathrm{Var}(K) = \mu + \alpha\mu^2 \tag{2.2}$$

The negative binomial was described first in ecological contexts, for calculating the number of a particular species in various locations or over time; the initial papers on estimators for the dispersion parameter are set in this ecological context [35–37]. Robinson et al. [21] and Anders and Huber [22] have suggested the negative binomial for differential analysis of sequence count data, including RNA-Seq. As $\alpha \to 0$, the negative binomial converges to a Poisson distribution, so satisfying property (2).

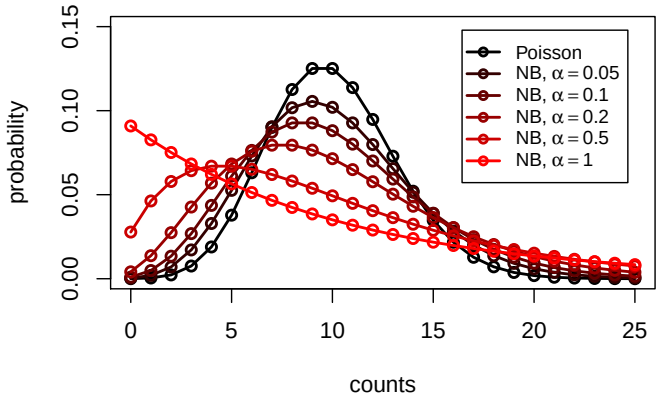

Figure 2.2: Poisson and negative binomial distributions, all with a mean value of 10. As the dispersion parameter, α goes to zero, the negative binomial distribution converges to a Poisson.

2.2.3 Poisson log normal distribution

Another possible distribution for the mean parameter of the Poisson distribution is the log normal distribution, resulting in a Poisson log normal distribution [38]. While the Poisson log normal does not have a closed-form distribution as the negative binomial, a random sample can be easily constructed by generating a normal random variable, $Z \sim \mathcal{N}(m, \sigma^2)$, and then generating counts $K \sim \text{Pois}(e^Z)$. Note that $\text{E}(e^Z) = e^{m+\sigma^2/2}$, i.e. the mean of e^Z is not simply the exponentiated mean of the normal random variable Z. While the negative binomial and Poisson log normal look very similar for low values of α and σ^2, as in Figure 2.3, they diverge for higher values of these two parameters. As the dispersion increases, the negative binomial puts increasing weight on the probability, $P(k = 0)$, however the Poisson log normal puts more weight on the right tail, seen in Figure 2.3. In this thesis, I will use for some applications the negative binomial and for other applications the Poisson log normal, with mathematical and computational convenience being the deciding factor[3].

2.2.4 Additional approaches involving discrete distributions

Two other approaches to count data, not used in this thesis but worth noting, are the zero-inflated negative binomial and the quasi-Poisson. The zero-inflated negative binomial is a negative binomial distribution with an additional peak of variable height at $P(k = 0)$. This distribution has been successfully applied to ChIP-Seq data by Rashid et al. [39], and is useful for count data which contain many zeros and many large values as well, which can be incompatible with the negative binomial or Poisson log normal distributions. The quasi-Poisson approach has been successfully used for RNA-Seq data by Lund et al. [40]. In this case the parameters are fit as if the counts were generated from a Poisson distribution, however all inference is performed on a log likelihood which is scaled by the estimated overdispersion [41]. Because the scaled log likelihood does not correspond to a proper likelihood, it is referred to as a "quasi-likelihood".

2.3 Comparison of discrete distributional modeling with alternatives

2.3.1 Why use discrete distributions?

At this point, one could ask what is gained by using discrete distributions for modeling sequence count data. One could alternatively use non-parametric methods such as permutation tests, or transform the data and use familiar statistical tests for normally distributed data, such as t-tests. In this section, I describe the differences and advantages from using

[3]The negative binomial has a closed-form distribution and so derivatives can be taken in order to optimize parameters. The Poisson log normal, however, is easy to implement in a hierarchical Bayes setting.

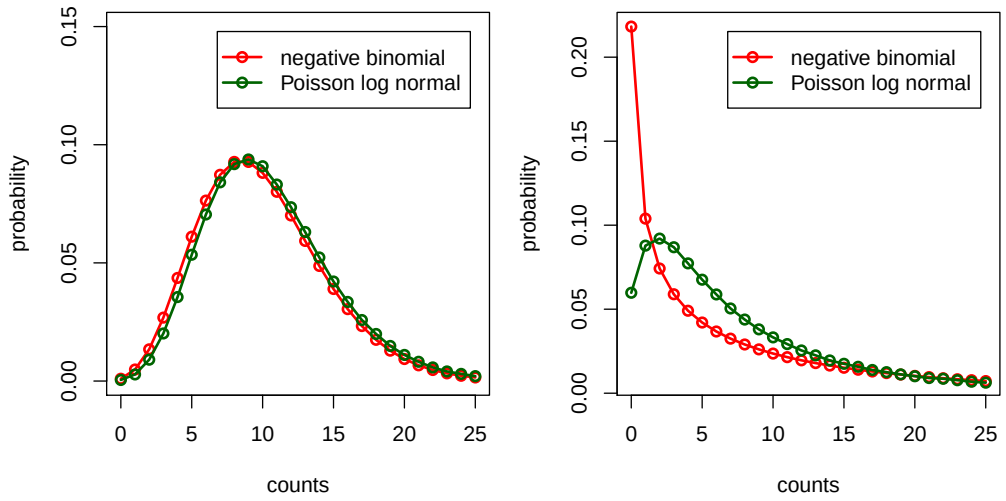

Figure 2.3: On the left, the negative binomial ($\mu = 10$ and $\alpha = 0.1$) and Poisson log normal distribution ($m = \log(10)$ and $\sigma^2 = 0.3$). On the right, the negative binomial ($\mu = 10$ and $\alpha = 2$) and Poisson log normal distribution ($m = \log(6)$ and $\sigma^2 = 1.05$), which have approximately equal mean and variance.

discrete distributional models over non-parametric approaches or transformations followed by tests based on normal assumptions.

2.3.2 Non-parametric modeling

If a model is "parametric" it means that the data is assumed to follow a certain distribution, which has one or more parameters which must be estimated. Non-parametric models attempt to make inferences about the data without making assumptions about the underlying distributions. Non-parametric approaches are therefore desirable, as new datasets will surely arise which do not exactly fit the parametric assumptions of even well-tested methodologies. A number of interesting non-parametric approaches have been proposed for analyzing high-throughput experiments, which include comparing the ranks of raw data between experiments [42, 43] and using resampling strategies [44]. I offer two reasons why one might chose a parametric model, despite the expected robust performance of non-parametric alternatives.

One advantage for parametric models is that they offer easily interpretable estimates of relationships between variables. For example, parametric models for analyzing differences in the means of counts between groups offer fold change estimates (or "effect size" estimates) with estimates of standard error or posterior distributions for model parameters. While rank-based methods or resampling strategies also offer statistical inference on the null hypothesis of no difference between groups, the calculation of effect size is not built into the method.

Another advantage of parametric models is the ability to handle small sample size and complex experimental design. Rank-based methods or resampling strategies generally rely on asymptotic behavior of a certain test statistic, which can break down due to the

granularity of datasets with small sample size. In the case of small sample size, rank-based methods cannot take into account the size of differences between groups when performing statistical inference. For example, the Mann-Whitney-Wilcoxon test (a rank-based test) assigns the same test statistic and p-value to a comparison of counts $\{1, 2\}$ vs $\{3, 4\}$ and $\{1, 2\}$ vs $\{100, 101\}$, because the ranks are identical in either case. Permutation tests are limited by the size of the set of all permutations in order to generate p-values. With small sample size and thousands of genomic ranges to test across, significant differences might be lost due to multiple test correction [45]. While it might be difficult to estimate parameters from datasets with small sample size, I will describe in later chapters how information can be pooled across genomic ranges in order to generate robust estimates of parameters. Finally, parametric models such as linear models and generalized linear models allow for the analysis of experiments with complex designs, such as paired tumor/normal samples, where confounding variables are controlled for in order to estimate the effect of a variable of interest. Non-parametric models cannot always be extended easily to accomodate these more complex designs.

2.3.3 Parametric modeling on transformed counts

Another approach instead of using discrete distributions would be to apply a transformation on counts and then use parametric models which are appropriate for real-valued data. For example, in order to compare counts between groups, one could take the log of the raw counts – adding a pseudocount in order to produce finite output even with counts of zero – and then use t-tests or F-tests which are appropriate for normally distributed data. In this section, I will briefly present some troubles which might arise from transformed data, and demonstrate differences in statistical power of different approaches using simulation.

Central to the question of how and whether to transform, is the issue of how the variance of the counts changes for subgroups with low or high mean count. Raw counts and transformed counts are often "heteroskedastic", meaning that the variance is different for different subgroups of the data. Heteroskedasticity is a problem for linear regression for example, which assumes that the dependent variable y is a linear combinations of columns of X plus independent error terms ϵ. The linear regression model is misspecified if the distribution of ϵ has a dependence on columns of X or on y.

For Poisson distributed data, the variance is equal to the mean, and for negative binomial the dependence is given by Eq. 2.2, so in both cases the counts are heteroskedastic. For Poisson distributed data, it is recommended to take the square root of the counts in order to stabilize the variance for different mean values. If the counts are overdispersed, the square root will no longer stabilize the variance. This is shown in Figure 2.4, where the variance for random samples of Poisson and negative binomial distributed counts is plotted against the mean parameter μ. More sophisticated approaches should be used instead, which estimate the dependence of the variance on the mean over many genomic ranges. "Variance stabilizing transformations" are one method, which model the variance-mean dependence and use this to derive a transformation [46, 47]. Another approach is to address

the variance-mean dependence further downstream, during the statistical inference steps [48].

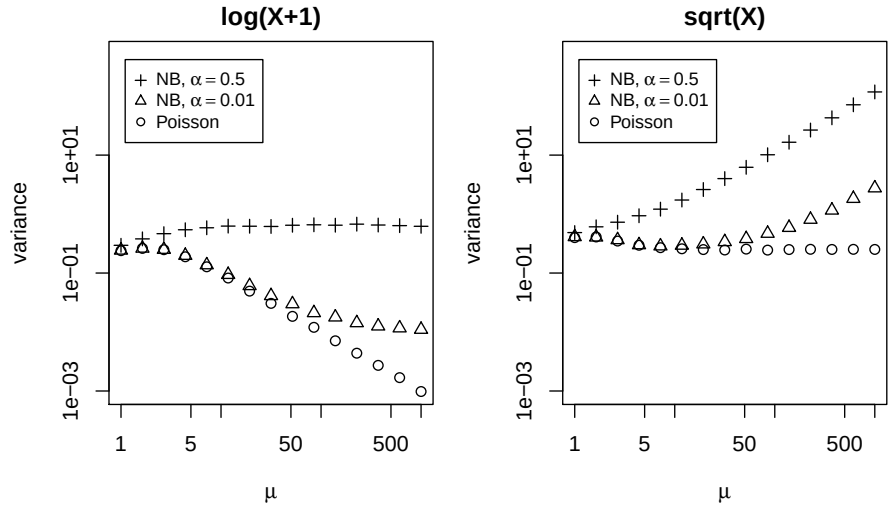

Figure 2.4: The empirical variance of transformed count data. On the left, the variance of log transformed counts plus a pseudocount of 1, and on the right, the variance of square-root-transformed counts. For each transformation, three different count distributions are plotted: a Poisson and a negative binomial with dispersion parameter $\alpha = .01$, and with $\alpha = .5$. The variance of the log transformed data generally decreases with increasing mean, while the variance of the square root transformed data generally increases with the mean. The square root transform apparently stabilizes the variance for Poisson distributed counts.

Even if a transformation is applied which effectively stabilizes the variance, the performance of the standard tests on transformed counts can fall behind those based on discrete distributions. For a demonstration of this, suppose Poisson distributed counts K for two groups $\{0, 1\}$, which have expected values μ_0 and μ_1, respectively. The $\log_2$ fold change between the two groups is then $\log_2(\mu_1) - \log_2(\mu_0)$. I compare two approaches: the likelihood ratio test for a negative binomial generalized linear model (abbreviated NB GLM) and a standard t-test on transformed data. The negative binomial generalized model is formally introduced in the next section, however for this demonstration it is only necessary to know that it models both the mean and the dispersion of count data. Alternatively, a t-test with equal variances is applied to counts which have been log transformed adding a pseudocount of 1, or counts which have been transformed by taking the square root. Both the NB GLM and the t-test attempt to estimate the group means and a parameter for the variance.

The results of this comparison demonstrate that for small counts or large counts with small fold changes, the discrete distribution approach is more sensitive to detect true differences than the t-test on transformed counts. Figure 2.5 shows the relative gain in statistical power of the NB GLM over the t-test on transformed counts. All cells in the figure represent data where a true difference in mean exists between the two groups. The negative binomial model outperforms the t-test for almost all cells, though the advantage disappears as the counts grow large and the fold change becomes small. Note that the NB GLM outperforms the t-test and square root on Poisson distributed data, even though

the square root was shown to effectively stabilize variances of Poisson distributed data in Figure 2.4. Supplementary Table B.1 shows that under simulated data with no difference between groups (a "null dataset"), all methods control Type I error, the probability to reject the null hypothesis when there is no true difference between groups.

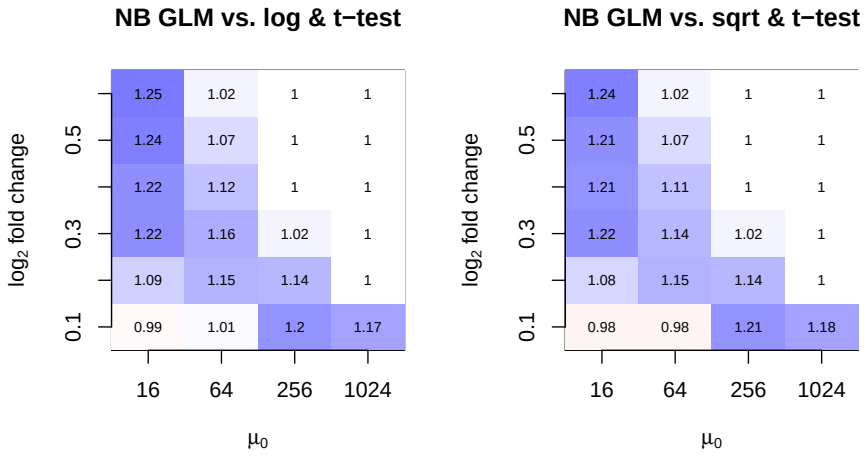

Figure 2.5: The relative power of testing with discrete distributions over testing with transformed counts, for difference in mean between two groups with three samples each. Plotted in each cell is the number of negative-binomial-based tests which rejected the null hypothesis divided by the same number for t-tests on log (left) or square root (right) transformed counts, over 1000 replications. The x-axis shows the true mean for one group, and the y-axis shows the $\log_2$ fold change for the other group. As the mean increases and the fold change decreases, there is decreasing gain in power from modeling with a discrete distribution.

2.4 Introduction to the generalized linear model

Models using discrete distributions benefit from interpretable coefficients, a unified approach regardless of sample size or experimental design, and increased power compared to a standard test of differences between two groups. I now introduce the "generalized linear model" (GLM) for count data, a form of which will be used in the following three chapters. They are "linear models" in that a column vector $\vec{\beta}$ of real numbers is used to create a linear combination of the columns of a data matrix X, written as a matrix multiplication $X\vec{\beta}$, which minimizes the error in approximating a target vector. "Generalized" refers to the fact that the target vector is constructed using a "link function" which can be applied to many different kinds of data, including continuous data on the real numbers, continuous non-negative data, non-negative integers (as with count data), and binomial or multinomial outcomes [49]. In the GLMs used in this thesis, a log link function is used, meaning that the variables in X have multiplicative effects on the counts, with the size of the effect specified by $\vec{\beta}$. Multiplicative effects are more appropriate than additive effects, as additive effects could lead to negative expected values which are impossible for counts.

The form of the GLM used in this thesis depends on the biological question and the type of data used for modeling. In Chapter 3 and Chapter 5, counts of reads in genomic ranges are modeled using a data matrix X, which contains covariates which also run along the length of the genome. The counts K_i of reads falling in genomic range i are given as:

$$K_i \sim \text{NB}(\mu_i, \alpha_i) \tag{2.3}$$

$$\log(\mu_i) = x_{i*}\vec{\beta} \tag{2.4}$$

where μ_i is the expected mean, α_i is a dispersion parameter, and x_{i*} is the i-th row of a matrix X containing the covariates. The coefficient vector $\vec{\beta}$ contains estimates of the expected log fold change in counts for each column in X, and is estimated once over all genomic ranges i of the genome.

In Chapter 4, an different model is used to account for multiple samples and to test for differences between these samples for every genomic range. The matrix X is called a "design matrix", and it has as many rows as the number of samples. The columns of X are variables containing information about the samples, e.g. control or treatment group, age, sex, etc. A separate coefficient vector $\vec{\beta_i}$ is fit for each genomic range i. The counts K_{ij} of reads falling in genomic range i and for sample j, are given by:

$$K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i) \tag{2.5}$$

$$\log(\mu_{ij}) = x_{j*}\vec{\beta_i} \tag{2.6}$$

where x_{j*} is the j-th row of the design matrix X. The vector $\vec{\beta_i}$ again contains estimates of expected log fold changes in counts for each column of X.

Each chapter has some variation on the basic GLM presented above: in Chapter 3, the mean parameter is a composite of a copy number ratio and a GLM-like term; in Chapter 4, an additional factor is incorporated in Eq. 2.6 to control for sequencing depth, and priors are used to estimate α_i and $\vec{\beta_i}$; in Chapter 5, the model is expanded to a hierarchical Bayes Poisson log normal model, though the form is similar and the interpretation of the parameter $\vec{\beta}$ as log fold changes remains the same.

2.5 Parametric model fit

For parametric models, the question remains how badly things can go wrong if the data does not conform to the specified distribution. This can occur for the negative binomial if there are mostly small non-zero counts with a small percentage of very high counts. As can be seen in Figure 2.3, the negative binomial distribution with high dispersion shifts

more of the probability density towards zero, so such counts will not fit well to a negative binomial.

In Chapter 3 (estimating copy number state from DNA-Seq read counts), a small percentage of high counts will not break the model, as the copy number state of neighboring genomic ranges will reduce the likelihood of calling a single range as copy number variant, despite an abnormally high count. In Chapter 4 (differential expression analysis of RNA-Seq data), outliers pose more of a problem, because a coefficient vector $\vec{\beta}_i$ is fit for each genomic range i. Here, Cook's distance is used to detect genomic ranges with individual counts which overly influence model parameters [50]. In Chapter 5 (hierarchical modeling of ChIP-Seq data), the parameters are estimated over hundreds of thousands of features and control experiments are used to accounts for technical artifacts on read counts.

Chapter 3

Modeling read counts for CNV detection in exome sequencing data

3.1 Introduction

Copy number variants (CNVs) are regions of a genome present in varying number in reference to another genome or population. CNVs are increasingly recognized as important components of genetic variation in the human genome and effective predictors of disease states. CNVs have been associated with a number of human diseases including cancer [13], autism [51, 52], schizophrenia [53], HIV (susceptibility) [54], and intellectual disability [55]. These variants produce phenotypic changes through gene dosage effects, when the number of copies of a gene leads to more or less of a gene product, through gene disruption, when a CNV breakpoint falls within a gene, or through regulatory effects, when a CNV affects regulatory sequences such as enhancers and insulators [56]. Recent studies report that $20 - 40$ megabases, around 1% of the genome, are copy number variant in individual human genomes, making CNVs a larger source of basepair variation than single nucleotide polymorphisms [57, 58].

Two primary technologies for genome-wide detection of CNVs are array comparative genomic hybridization (arrayCGH) and high-throughput sequencing (HTS). ArrayCGH measures the fluorescence of two labeled DNA samples, which competitively bind to many probe sequences printed on an array. When the values from the probes are lined up according to genomic location, regions with variant copy number ratio can be observed as consecutive probes with higher or lower fluorescence ratio. CNVs exhibit a number of different signatures in resequencing data, where HTS reads from a sample are mapped to a reference genome, as reviewed by Medvedev et al. [59]. One kind of HTS signature is given by aberrant distances between the mapped positions of a paired end fragment overlapping a CNV, or between the ends of an unmappable read overlapping a CNV breakpoint. Another HTS signature, which this paper will focus on, is the amount of HTS reads mapping to regions along the chromosome, or "read depth". The signature in this case is a region with higher or lower read depth compared to a control sequencing experiment, or compared to other regions within an experiment, assuming that HTS reads are distributed uniformly along the sample genome.

The read depth CNV signature is similar to the pattern seen in arrayCGH, so it is helpful to review the algorithms devised for this task. Popular algorithms for analyzing arrayCGH data include circular binary segmentation [60] and hidden Markov models [61, 62].

Hidden Markov models are useful for segmentation of many kinds of genomic data, as they represent linear sequences of observed data made up of homogeneous stretches associated with a hidden state. There are efficient algorithms for assessing the likelihood of an HMM with certain parameters given observed data and for estimating the most likely sequence of underlying states for a set of parameters [63]. The HMMs designed for arrayCGH data take as input log ratios of measured fluorescence, a continuous variable, while read depth data consists of discrete counts of reads. We will therefore consider how to adjust the HMM framework to model read counts.

The main obstacle for CNV detection from read depth is the variance due to technical factors rather than copy number changes. HTS reads are subject to differential rates of amplification before sequencing and differential levels of errors during sequencing and mapping. For any HTS experiment, read depth in a genomic region can be related to local GC-content, as well as sequence complexity and sequence repetitiveness in the genome [64]. In whole genome sequencing, it has been shown that normalizing read depth against GC-content can be sufficient to predict CNVs accurately [13–15, 65, 66]. In paired sequencing experiments, such as in tumor/normal samples, position-specific effects can be eliminated through direct comparison, similarly to the elimination of probe-specific effects in arrayCGH [67–71]. However, HTS experiments do not always cover the whole genome and do not always include a reference sample sequenced using the same experimental protocol.

In targeted sequencing, such as exome sequencing, DNA fragments from regions of interest are enriched over other fragments and sequenced. Ideally, the sequenced reads map only to the targeted regions. Targeted sequencing therefore results in fewer positions at which to observe a change in read depth attributable to a CNV. Most target enrichment platforms use the following steps:

1. DNA from a sample is fragmented and prepared for later sequencing.

2. Prepared DNA fragments are hybridized to biotinylated RNA oligonucleotides and captured with magnetic beads or hybridized to probes on an array.

3. The beads are washed, eluted and the RNA is digested or the array is washed and eluted.

4. The remaining DNA sequences are amplified and sequenced.

Within the targeted regions, the enrichment steps lead to less uniform read depth than in whole genome sequencing, but the read depth pattern is consistent among samples using the same sequencing technology and enrichment platform. Sequencing with three different technologies using the same enrichment platform, Harismendy et al. [72] find "a unique reproducible pattern of non-uniform sequence coverage" within each group and low correlation of read depth across different technologies. Testing three different target enrichment platforms with the same sequencing technology, Hedges et al. [73] report high correlation within samples from the same platform and low correlation across different platforms. Taking advantage of the reproducibility of read depth, Herman et al. [74] and

Nord et al. [75] are able to identify CNVs in targeted sequencing by normalizing read depth in individual samples against average depth over control samples, though thresholds must be set for calling a position as CNV.

We sought to extend the HMM framework for CNV detection in targeted sequencing data, modeling read counts in non-overlapping genomic ranges as the observed variable generated from a distribution depending on the hidden copy number state. Similar to the usage of covariates by Marioni et al. [62] in modulating transition probabilities, we outline a model which fits non-uniform read counts to positional covariates such as background read depth, GC-content and genomic range width. Background read depth is generated similarly to the methods of Herman et al. [74] and Nord et al. [75] by taking the median of normalized read depth per genomic range over a control set. By using a number of explanatory covariates, one can analyze samples which have positive but low correlation with background read depth and residual dependence on GC-content. Another benefit of the HMM framework is the forward algorithm, which allows for fitting the distributional parameters without knowing the underlying copy number state. The model formulation replaces preprocessing, thresholding, and genomic-range-merging steps with the optimization of a statistical likelihood over a parameter space.

We will present an HMM for predicting copy number state in exome and other targeted sequencing data using observed read counts and positional covariates. We show that this model can successfully detect private CNVs in an exome sequencing project using all samples to generate background read depth. We then evaluate the robustness of our method using a control set from publicly available exome sequencing data from an alternate enrichment platform. We simulate CNVs of various sizes and copy number in exome sequencing data and find that our model outperforms normalization and segmentation methods in recovering the simulated CNVs. Finally, we summarize the results and discuss possible extensions of the method.

3.2 Methods

3.2.1 Modeling resequencing read counts

As a measure of read depth, we count the number of start positions of reads with high mapping quality in non-overlapping genomic ranges along a chromosome. To examine the characteristics of targeted sequencing read depth, we will count reads from a whole exome sequencing project (Li et al. [76], discussed later) in genomic ranges covering only the consensus coding sequence (CCDS) [77]. CCDS regions larger than 200 base pairs (bp) are subdivided evenly into genomic ranges of around 100 bp. The distributions of counts per genomic range for one sample often have positive skewness (Figure 3.1). The maximal count in a genomic range can be up to 20 times the mean count.

Another method of setting genomic range locations, by covering the targeted regions with fixed-size genomic ranges, is comparable in terms of the qualitative signature of CNVs

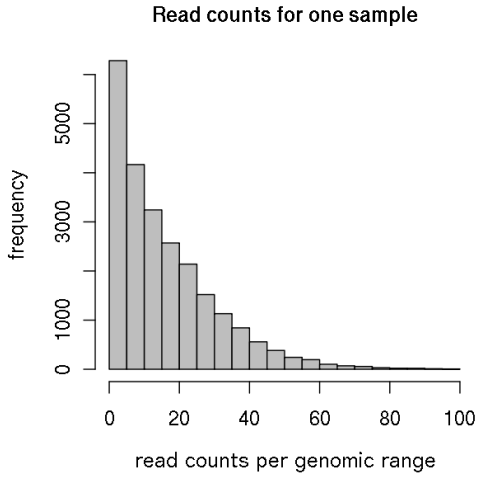

Figure 3.1: Distribution of read counts in genomic ranges covering the CCDS regions of chromosome 1 for one exome sequencing sample, cropped at 100 reads per genomic range.

in read depth and the resulting predicted CNV breakpoints. Setting genomic ranges within the CCDS regions has two advantages though. First, the CCDS regions are more likely to be covered equally across different enrichment platforms, enabling cross-platform comparison or control sets. Second, we find that the extremes of the targeted regions have more variability than the centers. By starting with the CCDS regions we can avoid these variable flanking regions. Genomic range width in the range of 1-200 bp was chosen based on the size distribution of annotated coding regions.

A suitable distribution for modeling the observed read counts in genomic ranges should have support on the non-negative integers. We could consider the Poisson distribution with a position-dependent mean parameter, representing the underlying rate of technical inflation of read counts. If the counts are distributed as a Poisson, then replicates should have equal mean and variance. We can check this assumption with read counts from a set of samples with similar amount of total sequencing. While these samples are not replicates, we expect that the private CNVs and SNPs which would alter read counts per sample should be rare in the coding regions. Plotting the variance over the mean for the read counts shows that most genomic ranges fall above the line $y = x$, and are therefore overdispersed for Poisson distributed data (Figure 3.2).

We use a negative binomial distribution to model the counts, and we use positional covariates to account for as much variance in read counts over genomic ranges as possible, allowing for the situation that unknown factors lead to overdispersed counts. We will first attempt to fit a single dispersion value α over all genomic ranges, then add model parameters to allow for α to vary over genomic ranges.

To obtain a measure of the positional non-uniformity in read depth, we calculate the median of sample-normalized read counts over a control set. Because samples vary in the total number of reads which map to the reference genome, we first need to normalize read counts per sample. Boxplots of read counts per genomic range for 5 samples are shown in

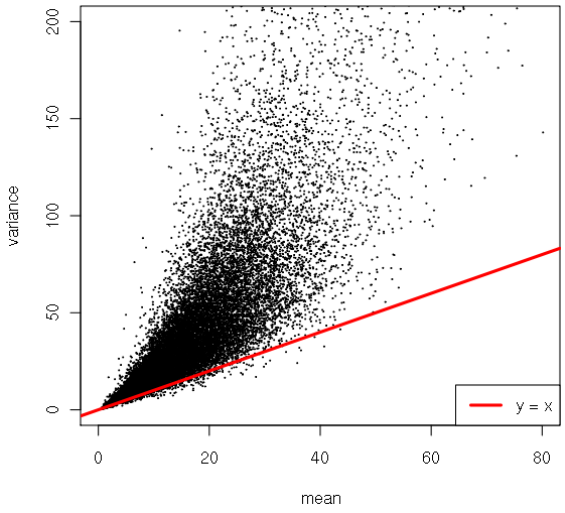

Figure 3.2: Mean and variance of read count for 23,619 genomic ranges over 40 samples with similar amount of total mapped reads.

Figure 3.3. The distributions all exhibit positive skewness but the median and quartiles are shifted. Given a matrix K of counts of reads in T genomic ranges on a chromosome (rows) across N samples (columns), K_{norm} is formed by dividing each column by its mean. Distributions of sample-normalized read counts per genomic range (rows of K_{norm}) indicate high variance in medians across consecutive genomic ranges (Figure 3.4). Some but not all of this variance of median read depth can be explained by GC-content (Figure 3.5). We calculate the background read depth by taking the median of the sample-normalized read count per genomic range (median of rows of K_{norm}), and the background standard deviation similarly.

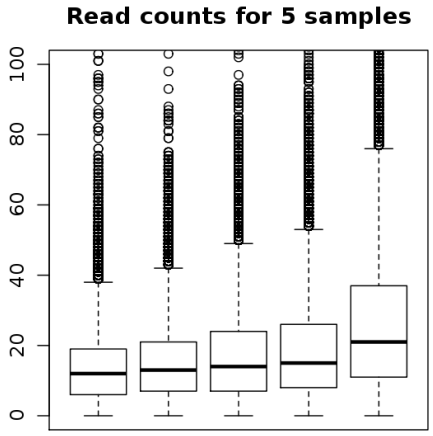

Figure 3.3: Boxplots of read counts for 5 samples over genomic ranges covering exons of chromosome 1.

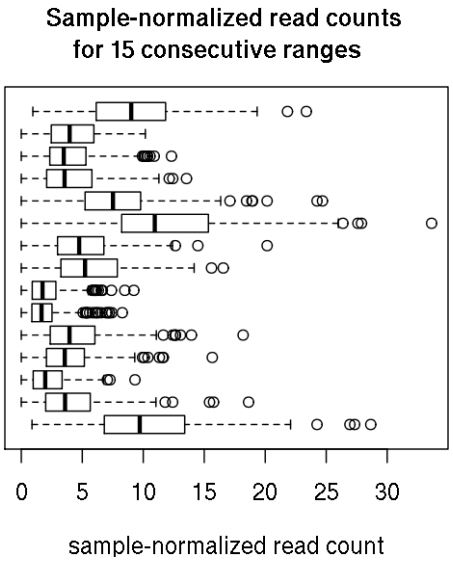

Figure 3.4: Sample-normalized read counts for 15 consecutive genomic ranges over 200 samples.

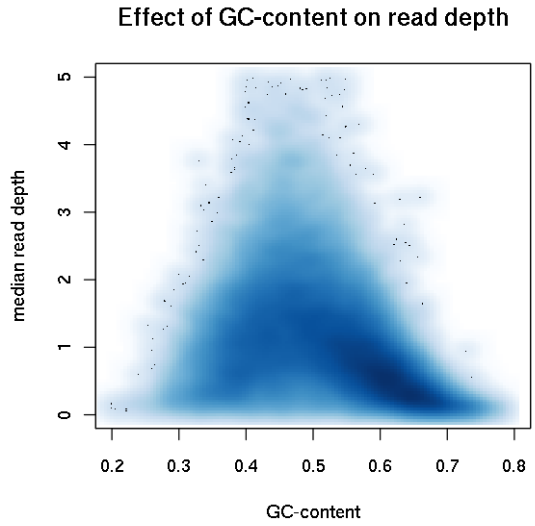

Figure 3.5: Smooth scatterplot of median read depth over GC-content. Median read depth is the median of sample-normalized read counts from 200 samples.

3.2.2 Hidden Markov model to predict sample CNVs

HMMs are a natural framework to segment genomic data with a discrete number of states, and we can take advantage of the algorithms that have been developed to evaluate these models. We observe K_{*j}, the j-th column of K, which represents the counts of HTS reads for sample j in T non-overlapping genomic ranges positioned linearly along a chromosome. These counts are the observed variables of our HMM, written as $\vec{O} = \{O_1, \ldots, O_T\}$, based on the notation of Rabiner [63] and Fridlyand [61]. We define a homogeneous discrete-time HMM, *exomeCopy*, to generate $\vec{O}$ by the following:

1. The number of states L. The set of states $\{S_1, \ldots, S_L\}$ represents the possible copy number states of the sample. $\vec{Q} = \{q_1, \ldots, q_T\}$ represents the vector of underlying copy number states over T genomic ranges. $q_t = S_i$ indicates that at genomic range t, the sample has copy number S_i.

2. The initial state distribution $\vec{\pi} = \{\pi_i\}$ where

$$\pi_i = P(q_1 = S_i), \quad 1 \leq i \leq L \tag{3.1}$$

3. The state transition probability distribution $A = \{a_{ij}\}$ where

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i), \quad 1 \leq i, j \leq L, \quad 1 \leq t \leq T - 1 \tag{3.2}$$

4. The emission distribution $B = \{b_i(\vec{O})\}$ where

$$b_i(\vec{O}) = \{f(O_t | q_t = S_i)\}, \quad 1 \leq i \leq L, \quad 1 \leq t \leq T \tag{3.3}$$

$$f \sim \text{NB}(O_t; \mu_{ti}, \alpha), \quad 1 \leq i \leq L, \quad 1 \leq t \leq T \tag{3.4}$$

   NB is the negative binomial distribution with mean and dispersion parameters $\mu, \alpha > 0$. Note that the mean of the emission distribution changes for different genomic ranges and states.

The choice of the number of underlying copy number states L must be fixed before fitting parameters, as well as the possible copy number values $\{S_i\}$ and expected copy number d. We tested the model for $\{S_i\} = \{0, 1, 2, 3, 4\}$ for the diploid genome ($d = 2$), and $\{S_i\} = \{0, 1, 2\}$ for the non-pseudoautosomal portion of the X chromosome in males ($d = 1$). Copy number ratios of sample to background higher than 2 can be modeled as well, but we expect reduced accuracy in differentiating between higher ratios, such as between 5/2 and 6/2.

Two transition probabilities are fit in the model: the probabilities of transitioning to a normal state and to a CNV state. These are depicted for a chromosome with expected copy count of 2 in Figure 3.6, with transitions going to the normal state as black lines and transitions going to a CNV state as gray dotted lines. The probability of staying in a state (grey solid lines) is set such that all transition probabilities from a state (rows of A) sum to 1. The initial distribution $\vec{\pi}$ is set equal to the transition probabilities from the normal state.

Consecutive genomic ranges in targeted sequencing can be adjacent on the chromosome if they subdivide a target region or very distant if they belong to different target regions. Therefore we might consider modifying the transition probabilities per genomic range, because two positions that are close together on the chromosome should have a higher chance of being in the same copy number state than those which are distant. This is reflected in the heterogeneous HMM of Marioni et al. [62] with transition probabilities that exponentially decay or grow to the stationary distributions as the distance grows. In

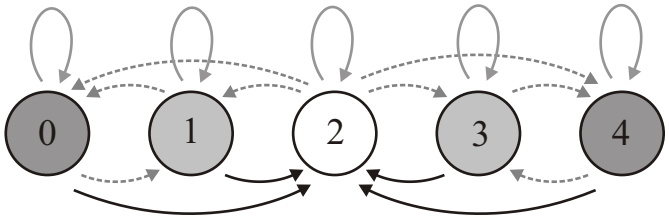

Figure 3.6: Transition probabilities for copy number states of the HMM with $\{S_i\} = \{0, 1, 2, 3, 4\}$ and expected copy number $d = 2$.

testing *exomeCopy*, we observed that a simple transition matrix results in similar CNV calls as the heterogeneous model without having to fit extra parameters.

While the HMMs of Fridlyand [61] and Marioni et al. [62] fit an unknown mean for the emission distribution of each hidden state, the emission distributions of *exomeCopy* for different states differ only by the discrete values $\{S_i\}$ associated with the hidden copy number state. Similar to the usage of positional covariates by Marioni et al. [62] to modulate the transition probabilities, we use positional covariates to adjust the means of the emission distribution, μ_{ti}. We introduce the following variables: X, a matrix with leftmost column a vector of 1's and remaining columns made up of positional covariates such as the log of the median background read depth, the genomic range width, and quadratic terms for GC-content; and $\vec{\beta}$ a vector of coefficients with length equal to the number of columns of X. The mean parameter μ_{ti} of the t-th genomic range and the i-th state is calculated by the product of the sample to background copy number ratio and an exponentiated linear combination of positional covariates x_{t*}, the t-th row of X. The mean parameter must be positive, hence the exponentiation of the term $x_{t*}\vec{\beta}$.

$$\mu_{ti} = \frac{S_i}{d} e^{(x_{t*}\vec{\beta})} \tag{3.5}$$

The parameters of the HMM can be written compactly as $\vec{\lambda} = (\vec{\pi}, A, B)$. The underlying parameters to fit in *exomeCopy* are the transition probability to normal state, the transition probability to CNV state, $\vec{\beta}$ and α. Parameters which are fixed are L, $\{S_i\}$ and d. The input data is $\vec{O}$ and X. The forward algorithm allows for efficient calculation of the likelihood of the parameters given the observed sequence of read counts, $\mathcal{L}(\vec{\lambda}|\vec{O})$ [63]. We use a slightly modified version of the likelihood function to deal with outlier positions. Some samples will occasionally have a very large count in genomic range t such that $b_i(O_t) < \epsilon$ for all states i and ϵ equal to the smallest positive number representable on the computer. In this case, the model likelihood is penalized and the previous column of normalized probabilities for the forward algorithm is duplicated.

To find an optimal $\vec{\lambda}$, we use Nelder-Mead optimization on the negative log likelihood function, with the `optim` function for the *stats* R package [78]. The optimization converges on a value of $\vec{\lambda}$ when changes to the negative log likelihood are less than a specified relative

tolerance. For this value of $\vec{\lambda}$, the Viterbi algorithm is used to evaluate the most likely sequence of copy number states at each genomic range,

$$\text{Viterbi path} = \underset{\vec{Q}}{\text{argmax}}(P(\vec{Q} \mid \vec{O},\ \vec{\lambda}))$$

This most likely path is then reported as ranges of predicted constant copy number. The ranges extend from the starting position of genomic range s with $\hat{q}_s \neq \hat{q}_{s-1}$ to the ending position of genomic range e, such that $\hat{q}_e = \hat{q}_t$ for all $t : s \leq t < e$. For targeted sequencing, the nearest genomic ranges are not necessarily adjacent, so the breakpoints could occur anywhere in between the end of genomic range $s-1$ and the start of genomic range s, for example. Ranges which correspond to CNVs can be intersected with gene annotations to build candidate lists of potentially pathogenic CNVs.

The optimization procedure requires that we set initial values for the various parameters to be fit. Initializing the probability to transition to a CNV state very low and the probability to transition to normal state high ensures that the Markov chain stays most often in the normal state. Initial probabilities for the first genomic range are set to the transition probabilities for the normal state. X is scaled to have non-intercept columns with zero mean and unit variance, as this was found to improve the results from numerical optimization. $\vec{\beta}$ is initialized to $\hat{\beta}$ using linear regression of the raw counts $\vec{O}$ on the scaled matrix of covariates X. α is initialized using the moment estimate for the dispersion parameter of a negative binomial random variable [37]. Although each genomic range is modeled with a different negative binomial distribution, we found a good initial estimate for α uses the sample mean $\bar{x}$ of $\vec{O}$ and the sample variance s^2 of $(\vec{O} - X\hat{\beta})$:

$$\hat{\alpha} = \max\left(\frac{(s^2 - \bar{x})}{\bar{x}^2}\ ,\ \epsilon\right), \quad \epsilon > 0 \tag{3.6}$$

We extend *exomeCopy* to an alternate model, *exomeCopyVar*, where α is replaced by $\vec{\alpha}$ which can vary across genomic ranges. The input data for modeling $\vec{\alpha}$ is the variance at each genomic range of sample-normalized read depth, which can be seen in Figure 3.4. This modification could potentially improve CNV detection by accounting for highly variable genomic ranges separately. We introduce Y, a matrix with leftmost column a vector of 1's and other columns of background standard deviation and background variance. The emission distributions are then defined by

$$f \sim \text{NB}(O_t; \mu_{ti}, \alpha_t), \quad 1 \leq i \leq L, \quad 1 \leq t \leq T \tag{3.7}$$

$$\alpha_t = e^{(y_{t*}\vec{\gamma})} \tag{3.8}$$

$\vec{\gamma}$ is a vector of coefficients fit similarly to $\vec{\beta}$ using numerical optimization of the likelihood. $\vec{\gamma}$ is initialized to $[\log(\hat{\alpha}), 0, 0, \ldots]$ with $\hat{\alpha}$ defined in Eq. 3.6.

3.3 Results

3.3.1 XLID project: chromosome X exome resequencing

The accuracy with which a model can predict CNVs from read depth depends on many experimental factors, so we try to recover both experimentally validated and simulated CNVs using backgrounds from different enrichment platforms. We use *exomeCopy* version 1.0.0, which enforces a positive mean parameter using $\max(x_{t*}\vec{\beta}, 0)$ rather the exponentiated term $e^{(x_{t*}\vec{\beta})}$ of Eq. 3.5 which is used in *exomeCopy* version 1.6.0 and higher. Both methods for enforcing a positive mean parameter produce nearly identical segmentations. First we run *exomeCopy* on data from a chromosome X exome sequencing project to find the potential genetic causes of disease in 248 male patients with X-linked Intellectual Disabilities (XLID) (manuscript submitted). As males are haploid for the non-pseudoautosomal portion of chromosome X, detection of CNVs is easier than in the case of heterozygous CNVs, where read depth drops or increases by approximately one half. The high coverage of the targeted region in this experiment also facilitates discovery of CNVs from changes in read depth. Each patient's chromosome X exons are targeted using a custom Agilent SureSelect platform and 76 bp single-end reads are generated using Illumina sequencing machines. Reads are mapped using RazerS software [79]. Total sequencing varies from 1 to 20 million reads per patient over 3.8 Mb of targeted region. Reads are counted in 100 bp genomic ranges covering the targeted region, and only genomic ranges with positive median read depth across all samples are retained. The positional covariates used are background read depth from all patients and quadratic terms for GC-content.

exomeCopy predicts on average 0.3% of genomic ranges per patient to be CNV. This represents 11,581 CNV segments from all patients combined, with 60% being single genomic ranges with outlying read counts. For candidate CNV validation we retain 640 predicted CNVs covering 5 or more genomic ranges. The larger segments are stronger causal candidates and we suspect are less enriched with artifacts. The majority of the 640 predicted CNVs are common across many patients. There are 66 predicted CNVs present in 1-2 patients, 14 in 3-10 patients, 8 in 11-20 patients, and 7 in 21-75 patients, described further in Table 3.1. We retain 16 predicted novel CNVs, which are present in 1-2 patients, not in the Database of Genomic Variants [80] and not already known to be associated with XLID.

As of writing, 10 predicted novel CNVs, 6 duplications and 4 deletions, have been tested and all were confirmed by arrayCGH or PCR. These CNVs are strong causal candidates based on segregation in the patients' families and the genes which are contained in the CNVs. This estimated lower bound of patients with causal candidate CNVs, about 4%, is in agreement with results from a previous study suggesting that 5-10% of cases of XLID can be attributed to CNVs [55]. Plots of experimentally validated CNVs found by our method are shown in Figure 3.7, with each point corresponding to the raw read count from a genomic range covering the targeted region.

		Genomic size							
		[600bp-10kb]		(10-20kb]		(20-100kb]		(100kb-4Mb]	
Type	Freq.	DGV+	DGV-	DGV+	DGV-	DGV+	DGV-	DGV+	DGV-
Dup.	1-2	10	10	2	3	2	3	2	16
	3-10	9	2	0	0	0	1	1	0
	11-20	2	1	1	0	2	0	0	0
	21-75	2	3	2	0	0	0	0	0
Del.	1-2	6	6	0	1	1	2	0	2
	3-10	1	0	0	0	0	0	0	0
	11-20	2	0	0	0	0	0	0	0
	21-75	0	0	0	0	0	0	0	0

Table 3.1: Predicted XLID CNVs by type, frequency, genomic size and inclusion in the Database of Genomic Variants (DGV)

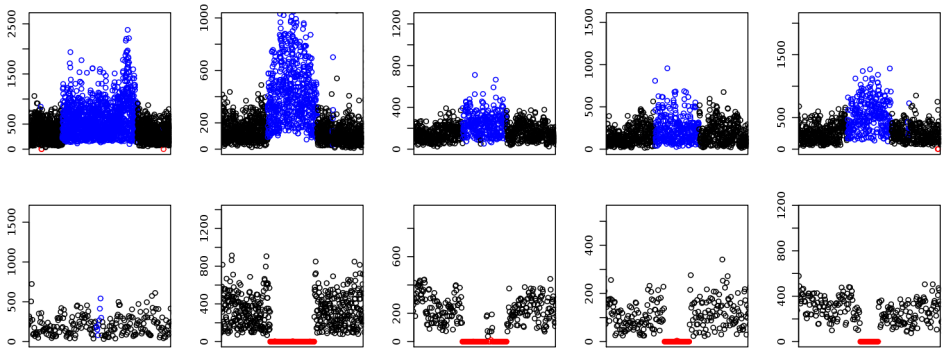

Figure 3.7: Experimentally validated CNVs identified in the XLID read depth data. The y-axis corresponds to the raw read counts for genomic ranges along the targeted region. The x-axis corresponds to the index of the genomic ranges. The color is the predicted copy number from *exomeCopy*, with blue indicating a hemizygous duplication and red indicating a hemizygous deletion.

3.3.2 Recovering XLID CNVs with a cross-platform control set

To investigate the effect of background read depth on CNV detection, we attempt to recover the experimentally validated CNVs in the XLID patients, substituting the XLID read depth background used in the previous section with a read depth background from a whole exome sequencing project of 200 Danish male and female individuals published by Li et al. [76] (referred to afterward as "Danish" or "Danish exomes"). We also run *exomeCopy* on nine XLID patients using no background read depth, but only GC-content and genomic range width information. In contrast to the custom Agilent platform used in the XLID project, the Danish samples were enriched for exons using a NimbleGen array and the coverage is substantially lower, with a median of 15 reads per genomic range compared to 326 per genomic range in the XLID project. For comparison of background read depth between the XLID samples and the Danish samples, we restrict the analysis to 9,710 CCDS-based genomic ranges on chromosome X, excluding the pseudoautosomal regions and regions not covered by both enrichment platforms. The CCDS regions are split evenly into genomic ranges no larger than 200 bp.

Comparing median read depth for XLID samples with median read depth for Danish samples shows positive but not strong correlation across the different platforms (Figure 3.8). Comparing within groups shows that two randomly selected subsets of a group are highly correlated in both datasets. This is in agreement with the observations of Hedges et al. [73] that read depth is highly correlated within enrichment platforms but only partially correlated across platforms.

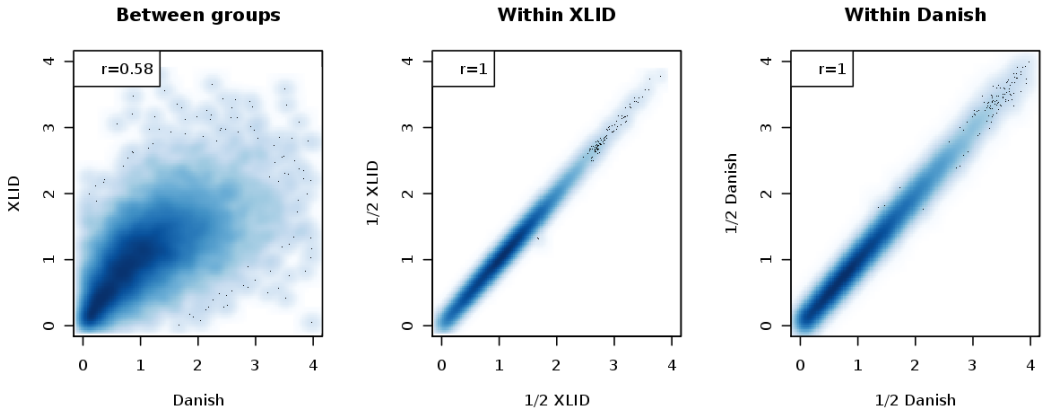

Figure 3.8: XLID median read depth and Danish exome median read depth. Between groups there is positive but not strong Pearson correlation, while randomly dividing groups and comparing median read depth within groups gives very high correlation.

As a robust measure of signal to noise, we calculate the median read depth divided by the median absolute deviation of read depth across genomic ranges on chromosome X covered by both the custom Agilent and NimbleGen enrichment platforms. For comparison of targeted sequencing with whole genome sequencing, we also provide read depth statistics from one sample of the 1000 Genomes project[1]. The decreased signal to noise ratio displayed in Table 3.2 for the targeted sequencing projects supports our assumption that target enrichment leads to non-uniformity in read depth.

study	submitted	Li et al.	1000 Genomes	1000 Genomes
population	XLID	Danish	PUR	NA12878
sequencing target	chrX exons	exome	exome	whole genome
# samples used	248	200	16	1
median read count (mean ± sd)	326 ± 96	15 ± 6	200 ± 102	105
signal to noise ratio (mean ± sd)	2.0 ± 0.2	1.3 ± .1	1.1 ± .03	2.7
mean pairwise correlation	.87	.77	.97	–

Table 3.2: Read depth statistics for four experiments in CCDS-based genomic regions on chr X.

We run *exomeCopy* on nine of the XLID patients with experimentally validated CNVs, once while substituting the XLID background with the Danish background, and again

[1]http://www.1000genomes.org/.

using no background read depth, only GC-content and genomic range width as covariates. One experimentally validated duplication is removed from analysis, as it spans genomic ranges not targeted by the NimbleGen platform. The median read depth from the XLID dataset and the Danish exome dataset is only partially correlated (r = 0.58), so dividing one by the other would not necessarily help to recover CNV signal. However, *exomeCopy* is able to adapt to less correlated backgrounds by reducing the contribution of the background term and increasing the contribution of the other covariates, genomic range width and quadratic terms for GC-content. The results in Table 3.3 demonstrate that with an independent control set for generating background, *exomeCopy* is frequently able to recover most of the genomic ranges contained within the experimentally validated CNVs. The sensitivity is measured as the percent of genomic ranges which are predicted as CNV out of the total number of genomic ranges contained within the validated CNV region, as the HMM does not always fit the entire span with the correct copy number state. The use of Danish exome background is always more sensitive in recovering CNVs than when *exomeCopy* is run without any read depth background. The average percent of genomic ranges predicted to be CNV is 5.4% and 1.9%, using Danish background and without background respectively. Also noteworthy in Table 3.3 is that CNVs with comparable genomic size can cover different numbers of genomic ranges, so methods for CNV detection in exome data should be sensitive to events covering only a few genomic ranges.

CNV type	# ranges	genomic size in kb	% CNV ranges recovered	
			Danish bg	without bg
duplication	488	899	80	31
duplication	218	291	96	94
duplication	90	541	100	34
duplication	90	541	100	1
duplication	74	329	87	83
deletion	51	237	100	100
deletion	21	169	77	77
deletion	17	27	100	100
deletion	4	49	100	100

Table 3.3: Recovery of experimentally validated XLID CNVs

3.3.3 Sensitivity analysis on simulated autosomal CNVs

In order to further evaluate the performance of the model on CNVs in autosomes and in low coverage samples, we simulate CNVs of various size and copy number on chromosome 1 in the Danish exome data. Simulated heterozygous deletions and duplications are generated in the Danish exome data by randomly sampling 50% of reads in a specified region and either removing or doubling the counts respectively. Simulated homozygous deletions and duplications are generated by removing 95% of the reads or doubling the reads respectively.

For sensitivity analysis, we simulate CNVs overlapping varying numbers of CCDS-based genomic ranges on chromosome 1, and report the percent of genomic ranges within the simulated CNV with accurate predicted copy number, averaging over a number of simu-

lation runs. We report the sensitivity in terms of genomic ranges rather than base pairs, as the major factor influencing sensitivity is the amount of exonic (targeted) base pairs contained within the CNV. The number of genomic ranges is approximately the amount of targeted base pairs contained within the CNV divided by the average genomic range size (112 bp for CCDS regions on chromosome 1). For reference, we include Table 3.4 which gives the estimated quartiles of genomic sizes in kilobases for varying number of CCDS-based genomic ranges on chromosome 1.

# CCDS-based genomic ranges	1Q	2Q	3Q
10	10	23	58
20	35	72	160
50	125	238	460
100	324	566	1043
200	684	1145	2037
400	1640	2656	4400

Table 3.4: Quartiles of genomic size (kb) by number of CCDS-based genomic ranges

We test the recovery of simulated CNVs with or without background variance information using *exomeCopy* and *exomeCopyVar* respectively. The model incorporating background variance performs nearly the same, although it has increased calling outside of the simulated CNVs and longer running time (Figure 3.9). For both models we can calculate the fitted ratio of variance to normal state mean, $(1 + \alpha\mu_{norm})$, averaging over all genomic ranges. *exomeCopy* fits the dispersion parameter α such that genomic range variance is on average 1.51 times the normal state mean. This supports the earlier analysis that read counts are overdispersed for Poisson. *exomeCopyVar* fits $\vec{\alpha}$ with a linear combination of Y, (Eq. 3.8) such that genomic range variance is on average 1.32 times the normal state mean. α_t is set to nearly zero for some genomic ranges, reducing the emission distributions to Poisson, but has higher α_t than used by *exomeCopy* for genomic ranges with high background variance.

We further compare the sensitivity of *exomeCopy* against normalization of log ratios followed by segmentation. We leave out *exomeCopyVar* as it uses background variance information in predicting copy number state which cannot be incorporated into normalization methods. For segmentation we use the circular binary segmentation algorithm of Venkatraman and Olshen [60] and the hidden Markov model of Marioni et al. [62], implemented in the R packages *DNAcopy* and *BioHMM* respectively. For comparing against normalization methods, we calculate the log ratio of sample counts plus a pseudocount of 0.1 over the median background. Log ratios are regressed on the remaining covariates (genomic range width and quadratic terms for GC-content), and the residuals are used as inputs to the segmentation algorithms.

Segmentation algorithms on the normalized data are preferable to the many false positives that would result from using thresholds. *DNAcopy* and *BioHMM* are run using default settings, except the `epsilon` parameter was lowered for *BioHMM* to 1e-4 to allow for sufficient number of simulations. Predicted segment means are translated into estimates of discrete copy number by thresholding at intermediate values. For diploid genome

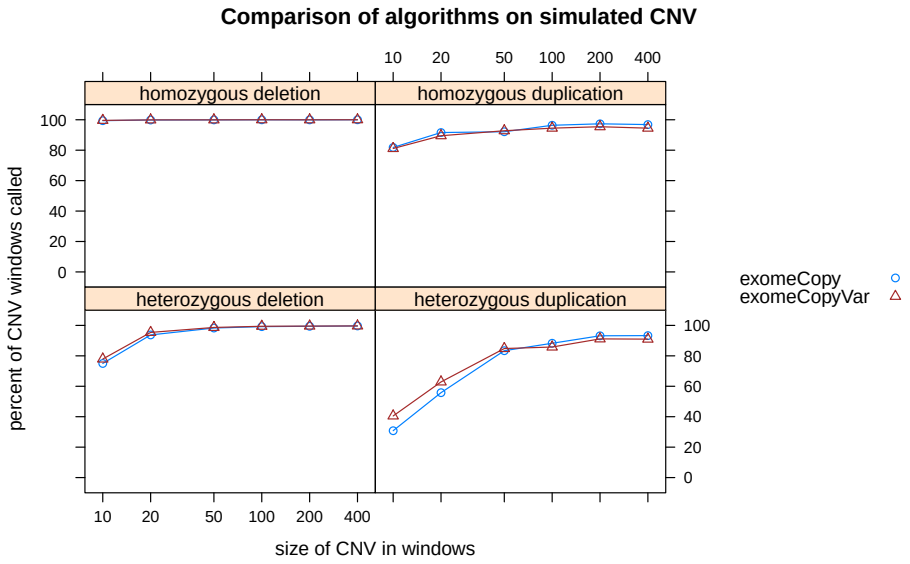

Figure 3.9: *exomeCopy* and *exomeCopyVar* perform similarly in recovering simulated CNVs of different type and size. Average percent of genomic ranges called CNV outside of the simulated CNVs is 0.5% and 0.8% and average run time is 7.6 s and 10.3 s for *exomeCopy*, *exomeCopyVar* respectively. Each point is the average over 100 simulations. The CNV sizes are described in terms of the genomic ranges, or "windows".

sequences, normalized log ratio in $(-\infty, \log(0.25)]$ is recorded as homozygous deletion, normalized log ratio in $(\log(0.25), \log(0.75)]$ is recorded as hemizygous deletion, etc. Relaxed evaluation allows any predicted value in $(-\infty, \log(0.75)]$ to be accepted for deletions and any predicted value in $(\log(1.25), \infty)$ to be accepted for duplications.

exomeCopy has equal or superior sensitivity to normalization and both segmentation methods for almost all types of CNVs (Figure 3.10). *exomeCopy* is often more sensitive for CNVs overlapping less than 100 genomic ranges, which is important as many of the experimentally validated CNVs from the XLID project overlapped 100 or fewer genomic ranges (Table 3.3). In the case of homozygous deletions, all methods can recover almost all genomic ranges of the simulated CNVs. In the relaxed evaluation, the results are very similar, with improved recovery for *BioHMM* in homozygous duplications and heterozygous deletions.

As our method relies on the sample having increased read depth relative to the background, it can be expected that the presence of the identical CNV in the control set would reduce sensitivity. To estimate this effect on sensitivity, we simulate CNVs both in the test sample and at different minor allele frequencies (MAF) in the control population. 400 simulations are performed for both homozygous and heterozygous deletions/duplications covering 100 genomic ranges on chromosome 1 in the Danish exome data. We vary the MAF and the number of control samples used to make the background. The simulated CNV is inserted into control sample chromosomes with probability equal to the MAF. At MAF levels less than 10%, we find that *exomeCopy* has 86% sensitivity or greater, nearly equal to the sensitivity with an MAF of 0% (Table 3.5). The number of controls used does not seem

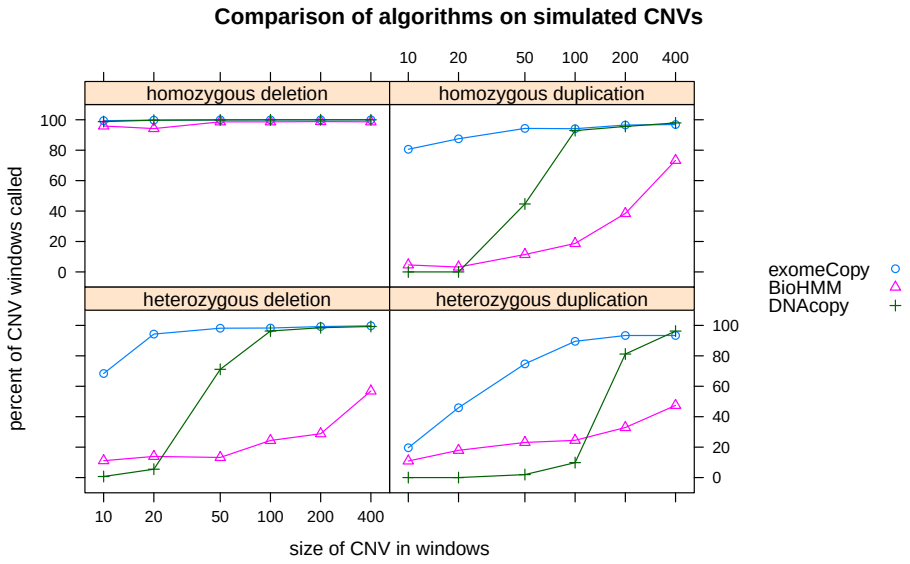

Figure 3.10: Performance of algorithms in recovering simulated CNVs on chr 1 of the Danish exome samples. *exomeCopy* is equally or more sensitive for almost all types and sizes of CNVs. Average percent of genomic ranges called CNV outside of the simulated CNVs is 0.4%, 5.2%, 0.2% and average run time is 7.4 s, 111.9 s, 3.7 s for *exomeCopy*, *BioHMM*, and *DNAcopy* respectively. Each point is the average over 100 simulations.

to have a large effect on the sensitivity, however individual samples in small control sets might bias results. The average percent of genomic ranges called CNV outside of the simulated CNVs is less than 0.9% for all combinations.

		MAF					
CNV type	# controls	0%	1%	5%	10%	25%	50%
homozygous deletion	10	100	100	100	100	98	68
	20	100	100	100	100	99	62
	100	100	100	100	100	100	56
homozygous duplication	10	97	96	92	88	59	16
	20	96	95	94	91	55	9
	100	95	97	94	90	59	3
heterozygous deletion	10	99	99	98	96	51	0
	20	99	99	98	96	48	0
	100	99	98	99	97	42	0
heterozygous duplication	10	89	89	87	75	38	0
	20	90	90	86	83	35	1
	100	91	88	88	82	36	0

Table 3.5: Percent of simulated CNV genomic ranges recovered by minor allele frequency and number of controls

We demonstrate *exomeCopy* adjusting to less correlated or uncorrelated backgrounds in Figure 3.11. After adding increasing amounts of noise to the original Danish background, the absolute value of the coefficients for genomic range width and quadratic terms for GC-content rise to replace the coefficient for noisy background. In the case that the sample is

entirely uncorrelated with the background, the model will remove all contribution of the background in modeling the read counts.

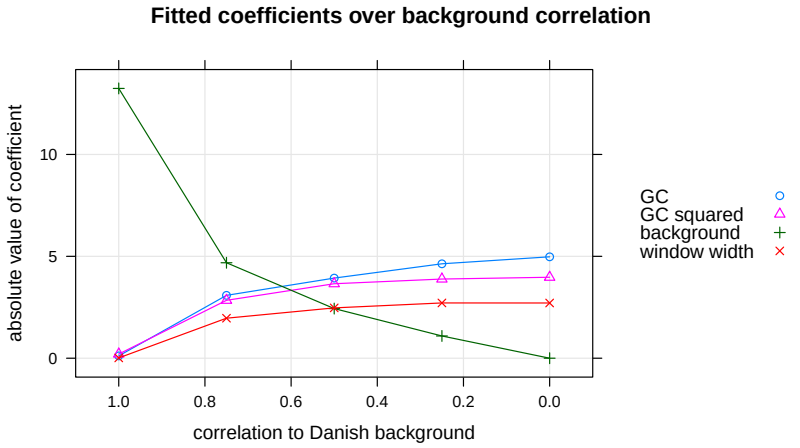

Figure 3.11: Effect of background correlation on the absolute value of fitted coefficients. The x-axis shows the correlation of the simulated background with the original background. Each point is the average over 100 simulations.

Simulations on the Danish exome data demonstrate that *exomeCopy* can often recover CNVs in low coverage data if they overlap sufficient amount of targeted sequence. However, we expect that *exomeCopy* will have improved performance with higher coverage autosomal datasets. To assess the influence of total sequencing depth on recovery of different kinds of CNVs, we performed further simulations on 16 high coverage exome sequencing samples from the PUR population of the 1000 Genomes Project. [81] The library format is paired-end data, and we count both ends in their respective genomic ranges. Although this decision introduces dependency between the counts in nearby genomic ranges, it avoids the loss of sample coverage information at either or both positions.

To simulate experiments with different amounts of total sequencing, we subsample reads from the original PUR samples to achieve 10, 20, 50 and 100 average read counts in genomic ranges subdividing the CCDS regions of chromosome 1. At each level of read depth, we create a background across all 16 PUR samples, then simulate CNVs of varying length and type as before. As expected, increasing the read depth increases the sensitivity of *exomeCopy*, especially for the detection of the smallest heterozygous duplications, with 78% or more genomic ranges recovered at an average read count of 50. (Figure 3.12). This simulation suggests that average read counts of at least 50 per genomic range will result in high sensitivity to detect both heterozygous and homozygous CNVs.

3.4 Discussion

Targeted sequencing is desirable for achieving high read coverage over regions of interest, while keeping costs and the size of generated data to manageable amounts. Exome sequencing prioritizes the discovery of variants in exons, as we expect these variants are more likely to be associated with a distinct phenotype than those which do not overlap

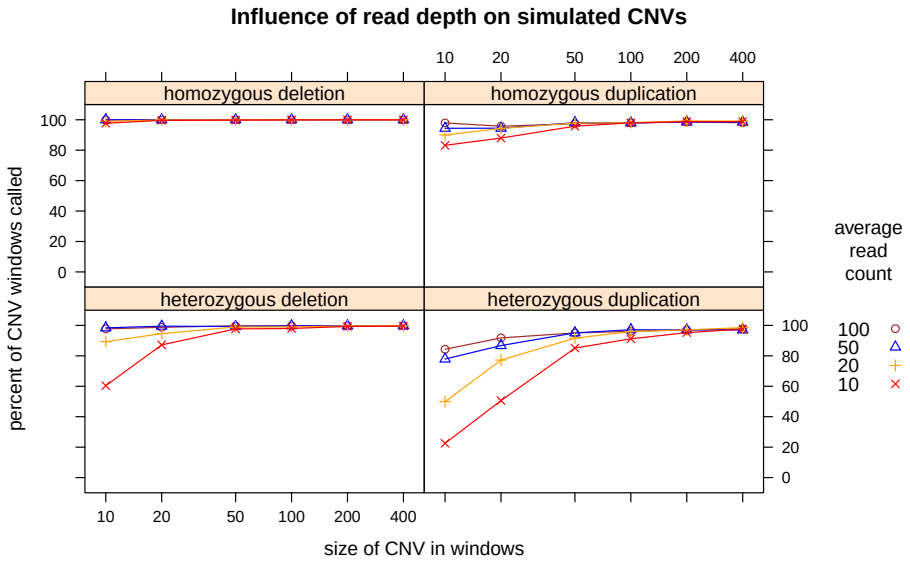

Figure 3.12: Performance of *exomeCopy* in recovering simulated CNVs on chr 1 after subsampling reads from the high coverage 1000 Genomes exome sequencing data. *exomeCopy* is increasingly sensitive with increasing average read counts. Average percent of genomic ranges called CNV outside of the simulated CNVs is always less than 0.7%. Each point is the average over 100 simulations.

exons. Nevertheless, methods for finding CNVs in targeted sequencing read depth data must overcome non-uniform patterns in read depth introduced by enrichment steps and a reduced number of genomic loci at which to observe changes.

We introduce a statistical model, *exomeCopy*, for detecting CNVs in targeted sequencing data which is robust across various enrichment platforms and different types and sizes of CNVs. In testing on exome sequencing data, our approach is more sensitive than normalization and state-of-the-art segmentation methods in finding duplications and heterozygous deletions which overlap few exons [60, 62]. *exomeCopy* formulates the CNV detection problem as the optimization of a likelihood function over few parameters, and therefore requires no thresholds or preprocessing decisions which might affect downstream results. In modeling sample read count using a number of covariates in addition to background read depth, our method can find CNVs in samples which show low correlation with the background. This allows for targeted sequencing projects with few samples to use median read depth from another project as background. While intuitively *exomeCopy* could also be applied to detect amplifications in cancer sequencing using the healthy tissue read depth as background, we believe the paired tumor/normal sequencing setup deserves a different statistical treatment. We therefore recommend the use of methods specifically designed for segmentation of paired tumor/normal exome sequencing experiments. [71]

Two limitations of CNV detection with targeted sequencing read depth are the effect of polymorphic CNVs in the control set and the inability to precisely localize CNV breakpoints. Although the median read depth method works well for finding CNVs which are rare in the control set, it might miss CNVs which are polymorphic. We formulate an HMM where the expected copy number d of the control set is constant over all genomic ranges.

For genotyping polymorphic CNVs, one could locally cluster samples in the control set by read depth and attempt to assign absolute copy numbers to the samples in a given region [15]. Then the read depth for $d = 2$ could be extrapolated from the clusters using their assigned copy numbers. Addressing the problem of localization, CNVs predicted from read depth in genomic ranges will not include exact breakpoints, and in the case of exome sequencing, the predicted breakpoints could fall anywhere between the outermost affected exons and the closest unaffected exons. Other sequencing based methods, such as partial mapping or anchored split mapping can be employed to recover breakpoints which fall within continuous targeted regions [75, 82].

As sequence read counts are increasingly taken as quantitative measurements, statisticians and bioinformaticians must adapt methods to separate technical bias from biologically meaningful signal. From our investigations, we find increased sensitivity to the underlying CNV signal in statistical modeling of the raw count data compared to converting counts to normalized log ratios. We expect that similar methods of contrasting individual samples against a background capturing technical bias will be useful in other sequencing protocols such as RNA-Seq and ChIP-Seq.

Chapter 4

Differential expression analysis for RNA-Seq using empirical Bayes priors for dispersion and fold change

4.1 Introduction

4.1.1 Detecting differences between samples over many genes

The widespread adoption of high-throughput sequencing technologies to assay various biological characteristics of the cell has resulted in a flourishing of statistical methods for detecting differences between samples. These methods often take advantage of the parallelized nature of the experiments; by integrating information about the samples over thousands of genomic ranges, these statistical methods can deliver more robust estimates and more sensitive tests of differences than methods which consider each genomic range in isolation. Recent methods for differential expression analysis of RNA-Seq data integrate information about the dispersion of counts across genes. For example, Robinson and Smyth [83] balance the estimate of dispersions for each gene with a common estimate across all genes using a weighted conditional likelihood. Anders and Huber [22] improve noisy dispersion estimates through modeling the dependence of the dispersion on the mean of counts over all samples. Hardcastle and Kelly [84] and Van De Wiel et al. [85] estimate priors for a Bayesian model over all the genes, and then provide posterior probabilities or false discovery rates for the case of differential expression. This chapter presents *DESeq2*, an update to the *DESeq* methodology of Anders and Huber [22], which incorporates both shrinkage of dispersion estimates and log fold changes using global information from all genes.

It helps to begin with a general mathematical description for the question of differential expression of count data. This chapter refers to counts of reads in genes for simplicity, however the methods presented here can also be used to perform differential analysis on any kind of count data generated from a parallelized experiment, including ChIP-Seq, 4C, Hi-C and spectral counts from mass spectrometry. Suppose K_{ij} is the count of sequencing reads aligning to gene i and sample j. For a read to align to a gene, it must overlap one or more of the exons of the gene, and not overlap any exons of other genes[1]. Suppose

[1]This is represented by the "union" mode depicted in the `htseq-count` documentation: http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

q_{ij} stands for an underlying rate which is proportional to the expected concentration of cDNA fragments for gene i across samples, and a function f_{ij} stands for the effect of technical artifacts on the underlying rate q_{ij}. Note that the rates q_{ij} are described as proportional across samples; the rates are not necessarily proportional to concentrations of cDNA fragments across different genes unless the estimate of f_{ij} takes into account all possible technical factors influencing abundance from cDNA fragments to read counts. Read counts falling in genes for technical replicates of RNA-Seq have been shown to fit a Poisson distribution for most genes [18], so then let K_{ij} follow a distribution given by:

$$K_{ij} \sim \text{Pois}(f_{ij}(q_{ij})) \tag{4.1}$$

In order to test, for a given gene i, whether the q_{ij} are different across groups, some estimate of the function f_{ij} must be made. One approach is to assume that the technical artifacts can be reduced to a single multiplicative factor for each sample: $f_{ij}(z) = s_j z$, for a "size factor", s_j. This controls for differences in read counts across samples due to the total number of sequenced read per sample, or "sequencing depth". For example, if sample A has two times the total sequenced reads of sample B, then one would expect two times the reads mapping to each genomic range, e.g. $s_A = 2s_B$. Another approach is to use known technical covariates in order to estimate multiplicative size factors for each sample and each gene: $f_{ij}(z) = NF_{ij}z$, for a "normalization factor", NF_{ij}. For samples sequenced at different times and across different labs, using sample- and gene-specific normalization factors can eliminate false positive calls which are due only to technical artifacts. Taking advantage of the parallelized nature of the data, Hansen et al. [86] and Risso et al. [87] model read counts for each sample on the GC-content of the genes and the gene length, in order to estimate f_{ij}. The methods presented in this chapter can employ either size factors or a matrix of normalization factors.

4.1.2 Generalized linear model for RNA-Seq

The differences between groups affecting the underlying rates q_{ij} can be conveniently represented using the generalized linear model notation of a design matrix X and a column vector $\vec{\beta}$ containing coefficients. For the simple two group model, with two samples per group, X can be written as:

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$$

As the model of Eq. 4.1 can only be used to model counts across technical replicates, the model is extended to a negative binomial distribution with gene-specific dispersion parameter α_i:

$$K_{ij} \sim \mathrm{NB}(f_{ij}(q_{ij}), \alpha_i) \tag{4.2}$$
$$\log_2(q_{ij}) = x_{j*}\vec{\beta}_i \tag{4.3}$$

where x_{j*} is used to denote the j-th row of the design matrix X which is multiplied by a column vector $\vec{\beta}_i$. The link function is switched from log to $\log_2$, so the coefficients can be more easily interpreted. The column vector $\vec{\beta}_i$ is indexed by i, as a different $\vec{\beta}_i$ is fit for each gene i. For the two group design matrix written above, $\vec{\beta}_i$ would be written as:

$$\vec{\beta}_i = \begin{pmatrix} \beta_{i0} \\ \beta_{i1} \end{pmatrix}$$

β_{i0} represents the intercept term, $\log_2$ of the mean of normalized counts for gene i for the first two samples. Normalization refers to the inverse of the function standing for technical artifacts, f_{ij}^{-1}, which is simple division in the case of multiplicative size factors. β_{i1} represent $\log_2$ fold changes between normalized counts of the two groups. For example, supposing $\beta_{i0} = 7, \beta_{i1} = 1$, the underlying proportions q_{ij} for samples 1, 2, 3 and 4 are:

$$q_{i1} = q_{i2} = 2^{\beta_{i0}} = 2^7 = 128$$
$$q_{i3} = q_{i4} = 2^{\beta_{i0}+\beta_{i1}} = 2^8 = 256$$

The true $\log_2$ fold change between the two groups is then β_{i1}:

$$\log_2\left(\frac{q_{i3}}{q_{i1}}\right) = (\beta_{0i} + \beta_{1i}) - (\beta_{0i}) = \beta_{i1}$$

4.1.3 Shrunken fold change estimates

One of the improvements offered in *DESeq2* over the previous implementation is the shrinking of highly variable log fold changes for genes with low counts or high dispersion to 0. As was observed with differential expression analysis using microarrays, genes with low intensity values might suffer from a decreased signal to noise ratio. Therefore the standard estimate of log fold change might not be the best estimator of the true log fold change. Newton et al. [88] propose using a prior distribution for the intensities, R and G, of two microarrays, and derive a Bayesian posterior estimate of the fold change, using a parameter $\nu > 0$:

$$\hat{\rho}_B = \frac{R + \nu}{G + \nu}$$

Without delving into the details of the method of Newton et al. [88], the advantage of such an estimator can be readily seen. For low intensity genes, ν will pull the ratio $\hat{\rho}_B$ toward 1 (and pull the log ratio toward 0), while barely changing the ratio for genes with $R, G \gg \nu$. This is desirable if the low intensity values are highly variable, and would otherwise result in high variance of $\hat{\rho}_B$. Low variance shrunken (or "moderated") log fold changes were also developed for microarray data using variance stabilizing normalization (VSN) [46]. Lowered variance comes at the cost of biasing log fold changes to 0, which is described in statistical literature as a "bias-variance trade-off" [89, 90].

An advantage of moderated log fold changes is that they can be used in downstream analysis, as in "continuous gene set enrichment analysis" where gene sets are identified with unexpectedly high or unexpectedly low log fold changes. Downstream analysis on unmoderated log fold changes is problematic, as the truly differential genes are mixed with genes with high variance estimates. Efforts toward moderating log fold changes for RNA-Seq include fully Bayesian hierarchical models [85] and a "generalized fold change" using a posterior distribution of log fold changes [91]. *DESeq2* assumes a zero-centered normal prior for log fold changes, using an "empirical Bayes" approach, where the variance of the prior is estimated using the distribution of maximum likelihood estimates for $\vec{\beta}_i$. The maximum *a posteriori* estimates of $\vec{\beta}_i$ and their standard errors are then used in a Wald test [92] of the null hypothesis $\mathcal{H}_0 : \beta_{ij} = 0$.

4.1.4 Shrinkage estimators for dispersion

The estimation of the dispersion parameter α_i is critical for any inference on differential expression, as the biological replicates are not expected to follow a Poisson distribution. Figure 4.1 shows simulated negative binomial counts using different dispersion parameters α_i. Typical dispersion estimates for RNA-Seq data can vary from around $0.01 - 0.1$ for genetically identical organisms or cell cultures to around 0.5 and higher for genetical heterogeneous populations. The levels of dispersion for RNA-Seq also typically exhibit a dependence on the mean counts, which is modeled in *DESeq2* using the parametric model of Anders and Huber [22].

Given the impact of different dispersion values on observed counts shown in Figure 4.1, an accurate estimate of dispersion is clearly needed for statistical inference. As estimates of dispersion are functions of the data, the estimates themselves are random variables which have a variance. Unfortunately typical sample sizes for RNA-Seq result in highly variable estimates of dispersion. Figure 4.2 shows the distribution of dispersion estimates for simulated data at various sample sizes. While the mean of the estimates is close to the true value for five samples and higher, the range between the 1st and 3rd quartiles is large even up until 20 samples. One sensible solution is to shrink the gene-wise estimates of dispersion toward a common value of dispersion for all genes [83, 93]. This moderation greatly reduces the variance of the gene-wise estimators, and therefore reduces many false positive calls of differential expression. *DESeq2* accomplishes moderation of dispersion estimates by assuming a log normal prior on dispersions. This is similar to the dispersion

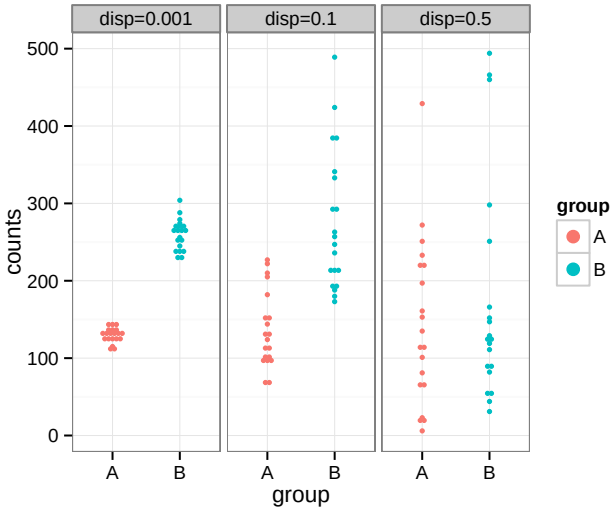

Figure 4.1: Example of simulated negative binomial counts with different dispersion parameters. For each plot, the true mean for group A is 128, and the true mean for group B is 256. The dispersion is changed between three values: 0.001, 0.1, 0.5.

estimation method of Wu et al. [93], though in the case of *DESeq2*, individual gene-wise estimates are shrunk towards a fitted value depending on the mean of counts for the gene.

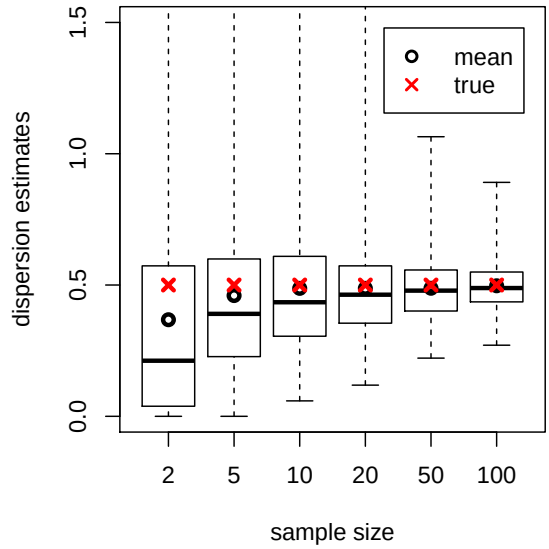

Figure 4.2: "Method of moments" [37] estimates of dispersion from simulated negative binomial data for different sample sizes. The true mean is 100, the true dispersion is 0.5 and shown as a red cross, and 1000 replicates are used for each sample size. The mean value for each sample size is shown as a circle. The Cox-Reid estimates of dispersion also show similar variance, though reduced bias.

4.1.5 Robust estimation and inference with *DESeq2*

This chapter introduces *DESeq2*, a statistical method for differential analysis of sequence count data. *DESeq2* extends the model of its predecessor *DESeq* [22], offering improvements in the estimation of the dispersion and fold changes. These changes result in increased sensitivity and more robust fold change estimates across experiments. Applications which are possible within the *DESeq2* framework are then discussed: the "regularized log" transformation, continuous gene set enrichment analysis using log fold changes, and alternate tests of log fold changes above or below a threshold.

4.2 Methods

4.2.1 GLM definition

Define the following GLM:

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i) \tag{4.4}$$

$$\mu_{ij} = s_j q_{ij} \tag{4.5}$$

$$\log_2(q_{ij}) = x_{j*}\vec{\beta}_i \tag{4.6}$$

K_{ij}	counts of reads for gene i, sample j
μ_{ij}	fitted mean
α_i	gene-specific dispersion
s_j	sample-specific size factor
q_{ij}	parameter proportional to the expected true concentration of fragments
x_{j*}	the j-th row of the design matrix X
$\vec{\beta}_i$	the $\log_2$ fold changes for gene i for each column of X
m	the number of samples
p	the number of coefficients to estimate, i.e. columns of X

The size factors s_j are calculated through the median ratio method [22]. Though the coefficients are defined here on the $\log_2$ scale, for simplicity the *DESeq2* fitting code and the following formulas are for a generalized linear model on the log scale. Recall the variance as a function of the mean for the negative binomial is given by:

$$V(\mu) = \mu + \alpha\mu^2 \tag{4.7}$$

4.2.2 Dispersion estimates and prior

Gene-wise estimates of dispersion $\hat{\alpha}_{\text{gene-est}}$, are calculated by optimizing the dispersion over the Cox-Reid adjusted log likelihood [94, 95]. The Cox-Reid adjustment corrects for a negative bias on dispersion estimates from using the maximum likelihood estimates (MLE) for μ (analogous to the negative bias of the MLE sample variance). A parametric curve of the form $y = a/x + b$ is fit to the gene-wise dispersion estimates over the mean of the normalized counts for a gene [22]. The fitted values α_{fit} are used as the mean of a log-normal prior on dispersions.

In order to calculate the prior variance, first the sample variance of the log dispersion estimates and the expected variance of log dispersion estimates must be calculated. A robust estimate of variance of the log gene-wise dispersion estimates $s^2_{\text{rob}}(\log(\hat{\alpha}_{\text{gene-est}}))$ is calculated using the residuals of the log gene-wise estimates $\log(\hat{\alpha}_{\text{gene-est}})$ to the fitted values $\log(\alpha_{\text{fit}})$. The robust sample variance is then the square of the scaled median absolute deviation of these residuals (using the scaling factor for the normal distribution provided by the R function `mad`).

A rough estimate for the expected variance for the log dispersion estimates, $\text{Var}(\log(\hat{\alpha}))$, is calculated as follows. First, the distribution of the dispersion estimate for a single gene is considered. The expected variance of the method of moments estimator is calculated instead of the maximum likelihood estimate or Cox-Reid estimate. Furthermore, a simplification is made assuming that μ is large; in this case, the variance of negative binomial counts is dominated by the $\alpha\mu^2$ term:

$$\text{Var}(K) \approx \alpha\mu^2$$

The dispersion α can then be estimated as the sample variance divided by the sample mean squared:

$$\hat{\alpha} = \sum_j \frac{(K_j - \hat{\mu}_j)^2}{(m-p)\hat{\mu}_j^2} \tag{4.8}$$

where $\hat{\mu}$ is the sample mean, m the number of samples and p the number of coefficients to estimate. The $(m-p)$ comes from the unbiased pooled sample variance of K.

The sum of squared Pearson residuals, $(K_j - \hat{\mu}_j)^2/V(\hat{\mu}_j)$, for a generalized linear model is approximately distributed as a chi-squared random variable with $(m-p)$ degrees of freedom [96]. This can then be used to obtain a rough estimate of the distribution of the dispersion estimate:

$$\sum_j \frac{(K_j - \hat{\mu}_j)^2}{\alpha \hat{\mu}_j^2} \sim \chi^2_{m-p} \tag{4.9}$$

$$\frac{1}{\alpha} \sum_j \frac{(K_j - \hat{\mu}_j)^2}{\hat{\mu}_j^2} \sim \chi^2_{m-p} \tag{4.10}$$

$$\frac{(m-p)}{\alpha} \sum_j \frac{(K_j - \hat{\mu}_j)^2}{(m-p)\hat{\mu}_j^2} \sim \chi^2_{m-p} \tag{4.11}$$

$$\frac{(m-p)}{\alpha} \hat{\alpha} \sim \chi^2_{m-p} \tag{4.12}$$

Fortunately, the true dispersion value α is not needed in order to estimate the sampling variance of the log dispersion, as the variance of the log of a product of independent variables can be separated into a sum of the variances of the log of each variable. As α and $(m-p)$ are constants, they have zero variance.

$$\begin{aligned} \mathrm{Var}(\log(((m-p)/\alpha)\hat{\alpha})) &= \mathrm{Var}(\log((m-p)/\alpha)) + \mathrm{Var}(\log(\hat{\alpha})) \\ &= \mathrm{Var}(\log(\hat{\alpha})) \end{aligned}$$

Abramowitz and Stegun [97] provide a formula for variance of the log of a chi-squared random variable, using the trigamma function ψ_1:

$$\begin{aligned} A &\sim \chi^2_{m-p} \\ \mathrm{Var}(\log(A)) &= \psi_1((m-p)/2) \end{aligned} \tag{4.13}$$

The expected variance of the log dispersion estimate is then estimated by:

$$\mathrm{Var}(\log(\hat{\alpha})) \approx \psi_1((m-p)/2) \tag{4.14}$$

Supplementary Table B.2 provides examples of this theoretical estimate for the variance of log dispersion compared with the sample variance of log dispersion estimates for simulated data, over a combination of different sample sizes, number of parameters and true dispersions.

The variance of the prior of the log dispersion is then calculated by subtracting the expected variance from the sample variance:

$$\sigma^2_{\alpha\text{-prior}} = \max(s^2_{\mathrm{rob}}(\log(\hat{\alpha}_{\text{gene-est}})) - \mathrm{Var}(\log(\hat{\alpha})), \epsilon) \tag{4.15}$$

$\epsilon = 0.25$ is used so that the dispersion estimates are not shrunk entirely to the fitted values in the case that the sample variance is smaller than the expected variance.

4.2.3 Dispersion outliers

The counts for some genes might not fit the negative binomial model or the dispersions estimates might not seem to belong to the common distribution of dispersion estimates. If these gene-wise estimates were moderated to the fitted line, this could produce spurious results: low p-values for genes with high variance of counts. A good heuristic for identifying dispersion outliers is those genes where $\log(\hat{\alpha}_{\text{gene-est}})$ is more than 2 times $s_{\text{rob}}(\log(\hat{\alpha}_{\text{gene-est}}))$ from $\log(\alpha_{\text{fit}})$. In this case, the genes are flagged and the gene-wise estimates are not shrunk towards the common value.

4.2.4 Final dispersion estimates

The final, maximum *a posteriori* (MAP) dispersion estimate is given by:

$$\text{CR}(\alpha) = -\frac{1}{2}\log(\det(X^t W X)) \tag{4.16}$$

$$\text{prior}(\alpha) = f_{\mathcal{N}}(\log(\alpha); \log(\alpha_{\text{fit}}), \sigma^2_{\alpha\text{-prior}}) \tag{4.17}$$

$$\hat{\alpha} = \underset{\alpha}{\text{argmax}}\left(\ell(\alpha|k, \hat{\mu}) + \text{CR}(\alpha) + \log(\text{prior}(\alpha))\right) \tag{4.18}$$

Where $\text{CR}(\alpha)$ is the Cox-Reid adjustment to the log likelihood [95], and W is the diagonal weight matrix from the standard iteratively re-weighted least squares (IRLS) algorithm. The link function is $g(\mu) = \log(\mu)$, so the elements of W are given by:

$$w_{jj} = 1/(g'(\mu_j)^2 V(\mu_j)) \tag{4.19}$$

$$w_{jj} = 1/(\mu_j^{-2}(\mu_j + \alpha\mu_j^2)) \tag{4.20}$$

$$w_{jj} = 1/(\mu_j^{-1} + \alpha) \tag{4.21}$$

Optimization of the log dispersion, $\log(\hat{\alpha})$, is performed using a backtracking line search with proposals accepted which satisfy Armijo conditions [98].

4.2.5 Beta prior

A zero-centered normal prior is introduced for the non-intercept betas. First, maximum likelihood estimates must be calculated for $\vec{\beta}_i$ using the standard IRLS algorithm [96]. The prior variance for the coefficient corresponding to the k-th column of X, $\sigma^2_{\beta_k\text{-prior}}$ is

then calculated as the mean of the squared values of the MLE β_k over all genes, excluding those genes where the MLE is undefined (e.g. zero counts for all samples of one condition).

4.2.6 Final beta estimates

The final, MAP beta estimate is given by:

$$\text{prior}(\vec{\beta}) = f_{\mathcal{N}}(\vec{\beta}; 0, \sigma^2_{\vec{\beta}\text{-prior}}) \tag{4.22}$$

$$\hat{\beta} = \underset{\vec{\beta}}{\text{argmax}} \Big(\ell(\vec{\beta}|\vec{K}, X, \hat{\alpha}) + \log(\text{prior}(\vec{\beta})) \Big) \tag{4.23}$$

where $\vec{K}$ represents the counts for a single gene over the samples. This is optimized using the IRLS algorithm with a ridge penalty. Updates are of the form:

$$\hat{\beta} \leftarrow (X^t W X + \vec{\lambda} I)^{-1} X^t W \vec{z} \tag{4.24}$$

$$\lambda_k = 1/\sigma^2_{\beta_k\text{-prior}} \tag{4.25}$$

$$\vec{z} = \log(\vec{\mu}/\vec{s}) + (\vec{K} - \vec{\mu})/\vec{\mu} \tag{4.26}$$

for size factors $\vec{s}$. This algorithm for regularized generalized linear models is described as the "iteratively reweighted ridge regressions algorithm" by Park [99] and as "weighted updates" by Friedman et al. [100].

4.2.7 Wald test

The Wald test compares the beta estimate $\hat{\beta}$ divided by its estimated standard error $\text{SE}(\hat{\beta})$ to a standard normal distribution or t-distribution. The estimated standard error is calculated from the following formula for a generalized linear model with normal prior on betas [99, 101].

$$\text{Var}(\vec{\beta}) = (X^t W X + \vec{\lambda} I)^{-1} (X^t W X) (X^t W X + \vec{\lambda} I)^{-1} \tag{4.27}$$

DESeq2 uses a normal distribution for Wald test p-values, as tests performed on simulated data suggest that the dispersion shrinkage results in a large gain of degrees of freedom (Supplementary Figure A.1). The tail integrals are then multiplied by 2 in order to achieve a "two-tailed" test.

4.2.8 Cook's distance for outlier detection

Cook's distance [50] is used to identify genes with counts which might not fit to the specified model. Cook's distance for a sample is defined as the distance that the coefficients of a model would move if the sample were removed and the model refit. If this distance is large, it might indicate that a single sample is overly influencing the coefficients, and so the gene should be flagged for further inspection. Considering a single gene and sample j, Cook's distance for generalized linear models [102] is given by:

$$D_j = \frac{\text{Pearson-res}_j^2}{\tau p} \frac{h_{jj}}{(1-h_{jj})^2} \tag{4.28}$$

for the generalized linear model dispersion parameter τ (in the case of negative binomial GLM $\tau = 1$, and dispersion is modeled instead using α), p the number of parameters including the intercept, elements h_{jj} of the hat matrix H:

$$H = W^{1/2} X (X^t W X)^{-1} X^t W^{1/2} \tag{4.29}$$

Pearson residuals are calculated as:

$$\text{Pearson-res}_j = \frac{(K_j - \hat{\mu}_j)}{\sqrt{V(\hat{\mu}_j)}} \tag{4.30}$$

using α_{fit} in the variance function $V(\mu) = \mu + \alpha\mu^2$. *DESeq2* flags genes which have a sample with Cook's distance greater than the 0.75 quantile of the $F(p, m-p)$ distribution, where p is the number of model parameters including the intercept, and m is the number of samples. The use of the F distribution is motivated by Cook [50] as such: removing a single sample should not move the vector $\vec{\beta}$ outside of a 75% confidence region around $\vec{\beta}$ fit using all the samples.

4.2.9 Regularized log transformation

Similar to the MAP estimates given by Eq. 4.23, the following "regularized log" transformation (rlog) involves the shrinkage of fold changes between samples, where more shrinkage occurs for genes with higher estimates of dispersion. In practice, this transformation helps to stabilize the variance of counts across samples (which would otherwise depend on the mean counts), allowing for improved visualization and clustering of samples or genes.

The rlog transformation is calculated as follows: the normal design matrix is substituted by one in which an indicator variable is included for every sample, in addition to a column specifying an intercept term. While this design matrix is of less than full rank and would lead to an indeterminate solution for maximum likelihood estimation, the maximum *a posteriori* coefficient vector is still determinable using a prior on the non-intercept terms.

The variance of the prior is taken from the variance of all non-intercept MAP coefficients using a very wide prior (variance of $\log_2$ fold changes of 10^4).

The rlog transformed values are then given by $\log_2(q_{ij})$:

$$\log_2(q_{ij}) = x_{j*}\vec{\beta}_i \tag{4.31}$$

The transformation accounts for variation in sequencing depth across samples by dividing out the size factors s_j or normalization factors NF_{ij}. The rlog transformation is offered in *DESeq2* alongside the variance stabilizing transformation (VST) [22] derived from the methods of Huber et al. [46]. Large variation in sequencing depth across samples was observed to have a detrimental effect on the variance stabilizing transformation (VST), while this does not adversely affect the rlog transformation.

4.3 Results

4.3.1 Accuracy of MAP dispersions for simulated data

The accuracy of *DESeq2*'s maximum *a posteriori* (MAP) estimates of dispersion are assessed through simulation where the true value of the dispersion is known. As noted in Figure 4.2, gene-wise estimates of dispersion have high sampling variance up to medium sample sizes for RNA-Seq experiments ($m \sim 20$). However, the variable gene-wise estimates can be improved if the population dispersions are similarly distributed for genes with similar mean counts. Simulated negative binomial counts are generated for many genes using the same mean μ and random log-normal-distributed dispersions α_i. The *DESeq2* methods are performed resulting in gene-wise dispersion estimates, a fitted value (in this case just the mean of the gene-wise estimates), and the MAP dispersion estimates.

While the gene-wise estimates are unbiased for an individual gene, assuming the dispersions share a common distribution, an estimate which incorporates knowledge of the common distribution can be more accurate than the gene-wise estimate. This phenomenon is demonstrated in the boxplots of Figure 4.3. The simulated genes are broken into two groups: those whose gene-wise dispersion estimates are below the fitted value and those whose gene-wise estimates are above the fitted value. As would be expected, the gene-wise estimates below the fitted line are more likely to be underestimates of their true value, while the gene-wise dispersion estimates above the fitted line are more likely to be overestimates. Taking the fitted value reverses this problem: the true value α_i is passed over in moving from the gene-wise estimate to the fitted value. The MAP estimates find a balance such that the distribution of $\hat{\alpha}_i/\alpha_i$ is centered on 1.

The previous version *DESeq* adjusted gene-wise dispersion estimates by applying a "maximum rule": the final estimate of dispersion was taken as the maximum of the gene-wise estimate and the fitted value. This is a conservative rule, helping to control the number of false positives by raising the estimates below the fitted line (likely underestimates) up

to the fitted line. However, the control of false positives comes at the cost of sensitivity in two ways: the initially underestimated dispersions might be brought up past their true value, and the likely overestimates are kept at their original value. *DESeq2* increases sensitivity while holding specificity, by raising the underestimates and lowering the overestimates. This is accomplished through multiplying the gene-wise likelihood with a prior which incorporates global properties of the dataset.

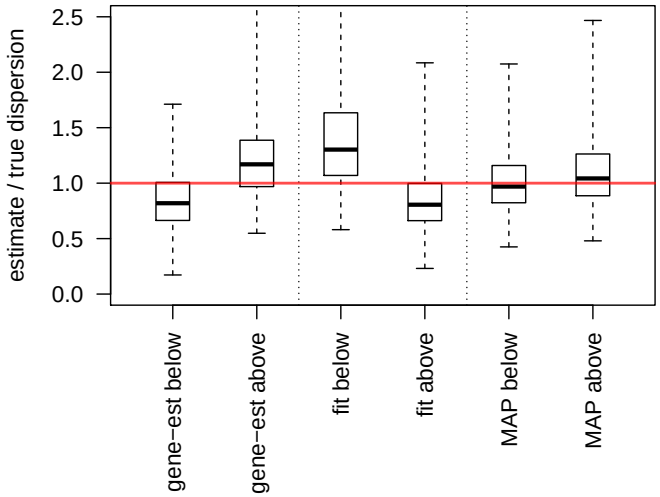

Figure 4.3: The ratio of various dispersion estimates over the true dispersion for 1000 genes with simulated negative binomial counts of 20 samples. The ratios are grouped by whether the gene-wise estimate is below or above the fitted value. The gene-wise estimates tend to be either below or above the true value depending on whether they are below or above the fitted trend line. The fitted line tends to be above the true value if the gene-wise estimate is below the fitted line, and below the true value if the gene-wise estimate is above the fitted line. The distribution of ratios for MAP estimates over the true value is centered on 1.

4.3.2 Effect of prior on log fold changes

The effect of using a zero-mean normal prior on log fold changes for real data is demonstrated in the following section, but first it is informative to consider how the prior affects the log fold changes for a toy example. Suppose two genes with counts split between two groups, of five samples each. The two genes have the same mean count and the same MLE $\log_2$ fold change – which is simply the difference between the $\log_2$ of mean counts for each group. However, one gene has a low dispersion and the other gene has a high dispersion. Though the genes have the same MLE $\log_2$ fold change, the likelihood will be more peaked for the gene with the low dispersion. This can be seen in Figure 4.4, where the low dispersion gene is in black and the high dispersion gene is in blue. A prior distribution on $\log_2$ fold change of $\mathcal{N}(0, 1)$ is applied to both genes. The posterior (dotted line) is very similar to the likelihood for the low dispersion gene, so the MAP estimate is close to the MLE. However, for the high dispersion gene, the posterior and MAP are pulled closer toward 0.

This provides some intuition as to how the MAP estimates can be more robust across experiments than the MLE. It is convenient to describe this effect in terms of the "observed Fisher information", which is the negative second derivative of the log likelihood at the MLE. If the likelihood is very peaked, as in the case of the low dispersion gene, then the log likelihood will also have a very negative second derivative, resulting in high Fisher information. The prior has less of an effect on the posterior if the Fisher information is high. If the likelihood surface is very flat at the MLE, then the Fisher information is low; estimates to the left or right are almost as likely as the MLE. The prior then provides a statistically principalled approach for moderating low information estimates while barely affecting the high information estimates.

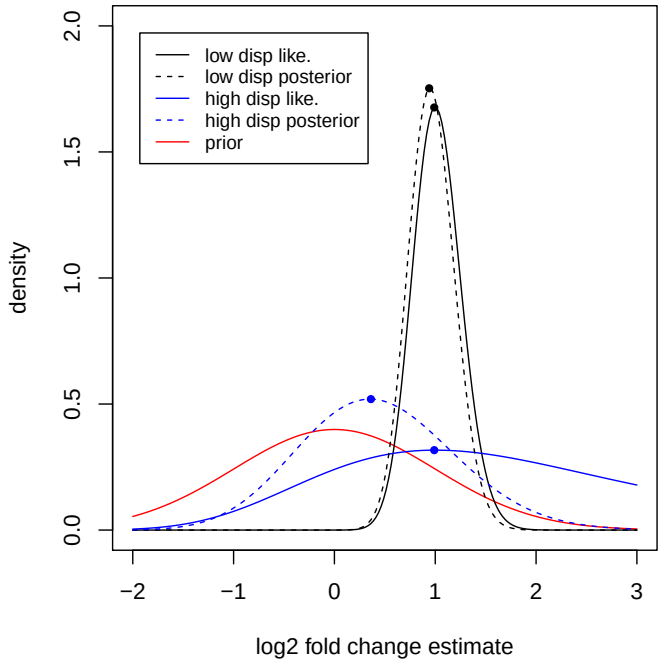

Figure 4.4: Effect of a prior on $\log_2$ fold changes for two toy genes. Counts are constructed for two groups of five samples each, such that both genes have the same mean count (192) and both have an MLE $\log_2$ fold change of 1. One gene has dispersion of 0.13 (black line, low dispersion), and the other gene has 5.55 (blue line, high dispersion). The solid lines indicate the likelihood (normalized to integrate to 1), the dotted lines indicate the posterior, and the points indicate the maximum values. A prior of $\mathcal{N}(0, 1)$ is shown as a solid red line. The MAP $\log_2$ fold change is barely changed for the low dispersion gene because the likelihood is more peaked; for the high dispersion gene, the prior pulls the MAP estimate closer to 0.

4.3.3 Differential expression analysis on RNA-Seq data

In order to demonstrate the effects of dispersion and fold change priors on real data, *DESeq2* was used to analyze RNA-Seq samples of primary cultures extracted from parathyroid tumors of 4 patients, each with control samples and samples treated with diarylpropionitrile (DPN), a selective estrogen $\beta 1$ agonist [103]. Differential expression analysis was performed on the samples cultured at 48 hours. The variation in the data due to the pa-

tient type contributes more than the variation due to the treatment type, as can be seen in a principal component plot of the samples (Supplementary Figure A.2). For this reason, it is necessary to employ a design matrix X which accounts for the patient variable and the treatment variable. The ability to control for covariates in estimation of fold changes and statistical inference is possible using generalized linear models.

The shrinkage of MAP dispersion estimates toward the fitted values is shown in Figure 4.5. This shrinkage includes the raising of dispersion estimates for many genes which could otherwise tend to 0 (these gene-wise estimates are given a minimal value of 10^{-8}), which would otherwise produce many false positive calls for genes with very low mean count. The shrinkage of MAP $\log_2$ fold changes towards zero is shown in Figure 4.6, compared to a similar plot of the MLE $\log_2$ fold changes. Again, the genes with very low mean counts would otherwise pose a problem, leading to large, highly variable estimates of $\log_2$ fold change. With the introduction of a zero-centered prior, these $\log_2$ fold change estimates are moderated to zero. Figure 4.7 also demonstrates the effect of shrinkage on fold changes, by plotting the MAP directly over the MLE, coloring genes by the mean normalized read count. For genes with mean count above 20, the prior does not have a large effect on the MAP $\log_2$ fold change.

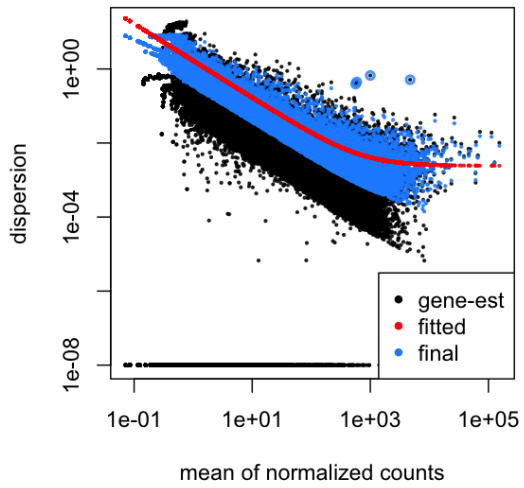

Figure 4.5: Dispersion estimates for the dataset of Haglund et al. [103]. The gene-wise estimates (black points) are shrunk towards the fitted value (red points), resulting in the final MAP estimates (blue points). The larger blue circles indicate dispersion outliers which are not shrunken towards the fitted values. The dispersion estimates for very high counts are moderated less toward the fitted line than those for the very low counts.

4.3.4 Regularized log transformation

The effect of the regularized log transformation is to shrink the $\log_2$ counts across samples towards a common mean, with higher shrinkage for those genes with high dispersion. The

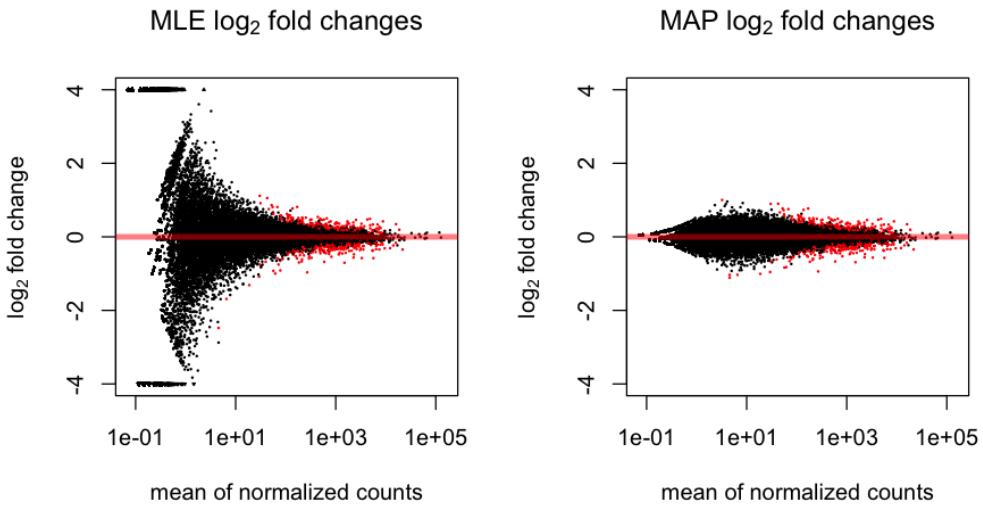

Figure 4.6: Plot of the maximum likelihood estimate (MLE) and maximum *a posteriori* (MAP) $\log_2$ fold changes, both over the mean of normalized counts. While the genes with mean normalized count less than 10 are shrunk towards zero, the estimates for genes with greater mean normalized count are nearly unchanged. The red points are those genes with Benjamini-Hochberg adjusted p-values less than 0.1. The points in the left plot at -4 and 4 are those genes with undefined beta estimates due to zeros in one condition. These are moderated toward zero through the use of a prior.

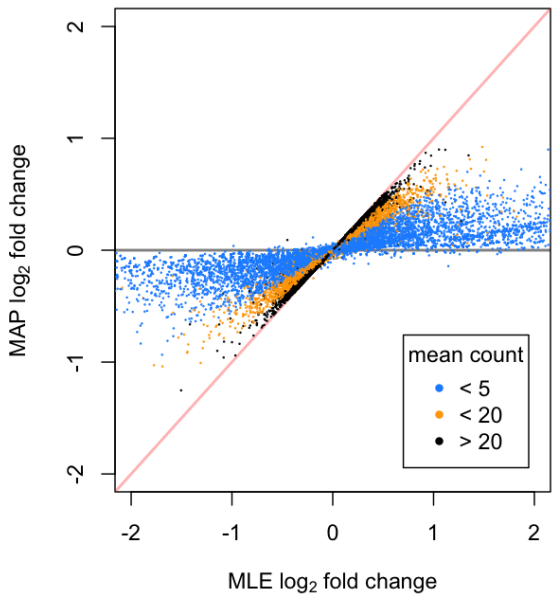

Figure 4.7: MAP $\log_2$ fold changes for each gene plotted over MLE $\log_2$ fold changes, colored by the mean normalized read count. For genes with mean normalized read count over 20 (the black points), the estimates are nearly equal, falling on the line $y = x$. Some MAP $\log_2$ fold changes for low count genes cross the horizontal axis, which is possible because shrinkage is also occurring on each of the patient coefficients simultaneously.

effect of the rlog transformation in comparison to a simple $\log_2$ transform for two genes is shown in Figure 4.8. One gene has high counts with a low dispersion estimate, and the other gene has low counts with a high dispersion estimate. The rlog transformed data for the high count, low dispersion gene is almost equal to the $\log_2$ of the counts, while the rlog transformation moderates the spread of values for the low count, high dispersion gene. For the Haglund et al. [103] dataset, the rlog transformation helps to stabilize the variance across the range of mean counts, though not as well as the VST discussed in Anders and Huber [22]. This can be seen by plotting the standard deviation over the rank of the mean for all genes, as shown in Supplementary Figure A.3. Using the rlog transformation allows for a cleaner picture of the sample-to-sample relationships, as seen in the PCA plot of Supplementary Figure A.2.

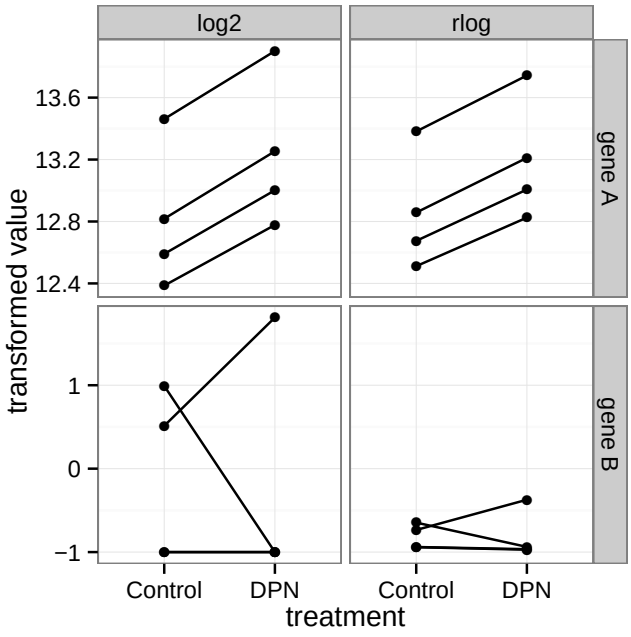

Figure 4.8: Example of two genes with $\log_2$ and rlog transformation applied to the counts. The lines connect samples from the same patient, which are either control cell lines or DPN treated. The rlog transformation for the gene with high counts (gene A) is almost equal to the $\log_2$ value, while the rlog transformed values for the gene with low counts (gene B) are shrunk towards each other. In order to take the $\log_2$, a pseudocount of 0.5 is added to the normalized counts, showing that the log counts can become negative if a pseudocount below 1 is chosen.

4.3.5 Cook's distance for detection of outliers

Cook's distance is used to detect genes which contain a sample which has large influence on the coefficient vector $\vec{\beta}$. The gene with the highest Cook's distance for the Haglund et al. [103] dataset has counts of zero for all samples except one treatment sample, which has a count of 66. The Cook's distance is therefore very high, because without this single sample, the coefficient vector would tend to negative infinity for the intercept (algorithm convergence at a very low value) and 0 for the patient and treatment effects. The gene-

wise dispersion estimate is relatively high (0.63), but the final MAP dispersion estimate is moderated close to the fitted value (0.30). The Wald statistic for the treatment variable is still large enough (3.25), such that an uncorrected p-value for this gene would be $\sim 10^{-3}$. While it could be that this gene is truly differentially expressed due to treatment and the other samples would show a fold change if the sequencing depth was higher, it is preferable to flag this gene as containing a sample with an outlier count.

4.3.6 Comparison of *DESeq2* against other methods

The performance of *DESeq2* was benchmarked against other packages: *DESeq* [22], *edgeR* [21], *DSS* [93] and *baySeq* [84] [2]. All packages make use of the negative binomial distribution for modeling read counts, though the implementation details differ. Comparisons are made across five real datasets[3] and two simulated datasets. Short descriptions of the datasets follow:

1. Bottomly et al. [105]: two inbred strains of mice
2. Hammer et al. [106]: rats with spinal nerve ligation or control
3. ModENCODE fly [107]: developmental time course of drosophila
4. Pasilla [108]: drosophila with knock-down of the pasilla gene
5. Wang et al. [109]: human tissue comparison
6. s0: simulated null dataset
7. s1: simulated dataset with true $\log_2$ fold changes $\sim \mathcal{N}(0, 1)$

The main condition for each experiment is given in Supplementary Table B.3. The number of genes and samples for each dataset is given in Supplementary Table B.4. Three of the experiments have a grouping factor, which was used for balancing when constructing subsets of the samples and provided to the packages which can accomodate complex designs[4].

All of the packages were run across all of the datasets using all the samples in order to get a sense of the total amount of calling. Table 4.1 provides the number of genes with Benjamini-Hochberg adjusted p-value [110], referred to here as FDR, less than 0.1. Table 4.2 provides the number of unique calls: genes which were called in only one package. Generally, *DESeq2* and *edgeR* call more genes differentially expressed compared to the other two packages. None of the packages call any of the genes differentially expressed in

[2] All packages were downloaded from the Bioconductor release branch 2.12. Version numbers: *DESeq* 1.12.0, *DESeq2* 1.0.17, *edgeR* 3.2.3, *DSS* 1.4.0 and *baySeq* 1.14.1. The GLM test of *DESeq* was used for easier comparison with *DESeq2*. The trended dispersion estimation of *edgeR* was used. *baySeq* was run with `getPriors.NB()` `samplesize` argument set to 10^4 for speed concerns.

[3] Four RNA-Seq datasets processed by Frazee et al. [104] and available at http://bowtie-bio.sourceforge.net/recount/ and one available in the Bioconductor package *pasilla*.

[4] Grouping factors which were supplied to *DESeq*, *DESeq2* and *edgeR*: Bottomly: three batches of samples; Hammer: two time points; pasilla: single-end and paired-end batches.

the simulated null dataset `s0`. *DESeq2* and *edgeR* have about equal sensitivity uncovering the truly differentially expressed genes in the simulated dataset `s1`. *baySeq* has much longer running times than the other four packages, despite using one tenth of the recommend sample size for estimating the priors (timing for each dataset provided in Supplementary Table B.5). *baySeq* was excluded from further analyses as it does not offer estimates of fold change which were used later for comparisons.

	DESeq	DESeq2	edgeR	DSS	baySeq
bottomly	1815	2420	2557	180	574
hammer	8363	8870	9870	1829	5473
modencodefly	2449	4794	2963	155	3462
pasilla	723	1241	1217	216	252
wang	2685	5255	5959	11	7779
s0	0	0	0	0	0
s1	914	1175	1241	0	748

Table 4.1: Number of total differential expression calls, FDR < 0.1.

	DESeq	DESeq2	edgeR	DSS	baySeq
bottomly	12	155	233	0	1
hammer	12	223	808	0	0
modencodefly	386	1330	18	0	434
pasilla	0	212	97	0	0
wang	0	321	28	0	1175
s0	0	0	0	0	0
s1	2	4	52	0	4

Table 4.2: Number of unique differential expression calls (only called by one package), FDR < 0.1.

As the ground truth of which genes are differentially expressed in these datasets is not known, the datasets were randomly split into two subsets, one small and one large, to determine how reproducible is the set of genes below an FDR cutoff. All the packages were run on both subsets, and the datasets were randomly split 10 times in total. The smaller subset was of size three (or less if there are insufficient samples to split), and the larger subset contained the remaining samples. The sets were balanced with respect to the condition of interest and any potential grouping factors. Figure 4.9 shows the number of true positive (TP) and false positive (FP) calls, where a false positive is defined as a gene with FDR < 0.1 in the small subset with an estimated log fold change of the opposite sign in the larger subset, and a true positive is defined as a gene with FDR < 0.1 in the small subset with the same sign log fold change in the large subset. The trend across experiments is consistent, with the packages typically ordered from least to most true positives: *DSS*, *DESeq*, *DESeq2*, *edgeR*. Figure 4.10 shows the ratio of false positive calls divided by all positive calls, which should be comparable to the false discovery rate. Almost all repetitions for the packages result in less than 10% false discoveries in terms of wrong sign of log fold change in the held-out set. Figure 4.11 displays the percent of false positives for the top n genes with lowest FDR in the small subset, and displays the reported FDR for the n-th gene. The four packages all provide a similar false discovery

rate when ranking genes by the reported FDR, though the reported FDR varies across packages.

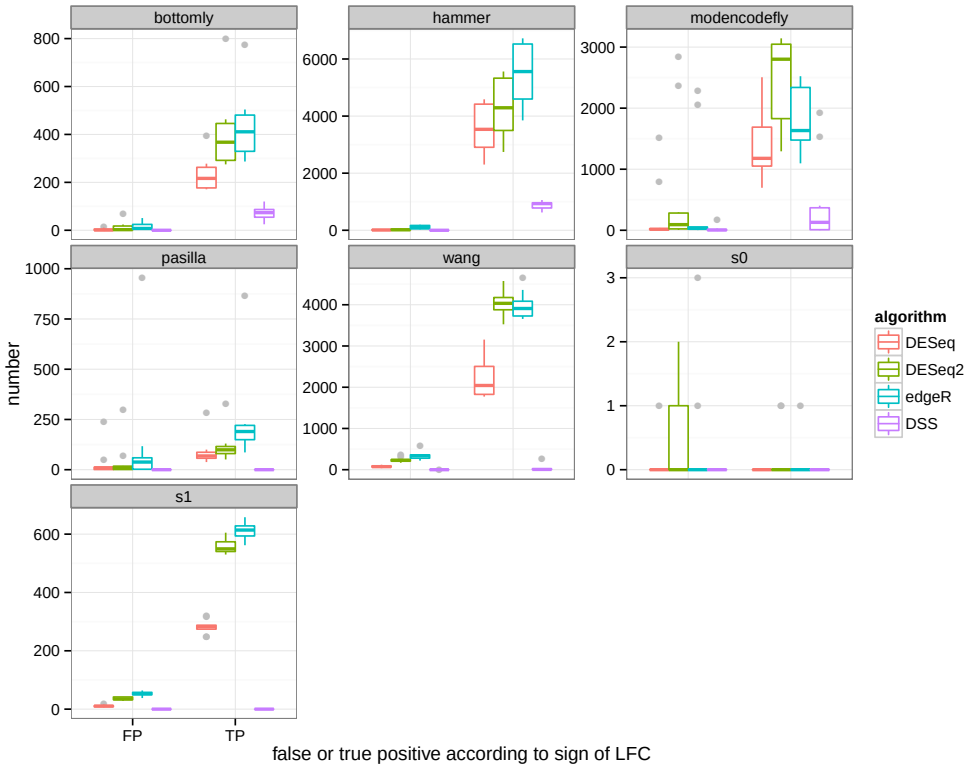

Figure 4.9: The true positive and false positive calls for various packages. False positive is defined as a gene with reported FDR < 0.1 in a small subset which has an estimated log fold change of opposite sign in the large subset. As was the case with calling on the full datasets, the *DESeq2* and *edgeR* packages have the highest number of calls, which are often true positives according to the sign of the log fold change in the held-out set.

DESeq2 provides more reproducible rankings by log fold change in five of the six non-null datasets. Figure 4.12 provides the ratio of top-ranked genes which are shared across the subsets, ranking genes by the estimated log fold change. Only in the ModENCODE fly dataset do the *DESeq2* fold changes have slightly reduced reproducibility compared to the other datasets, in which case the bias of the shrunken log fold changes towards 0 might be masking some low count genes with reproducible differential expression. Note that the ratio of shared top-ranked genes by log fold change is elevated for the other software packages for the null dataset, in which all log fold changes are equal to 0. This is most likely due to genes with low count, which are likely to have highly variable fold changes; by chance, half of the time these signs will be concordant across the subsets.

In summary, *DESeq2* incorporates moderation of dispersion estimates, increasing sensitivity while controlling the false discovery rate. The use of a prior on dispersion estimates is shown to produce more accurate estimates in simulation, when compared to the maximum-rule of the previous implementation, *DESeq*. The use of a prior on log fold changes moderates the spread of highly variable estimates for low count genes typical in

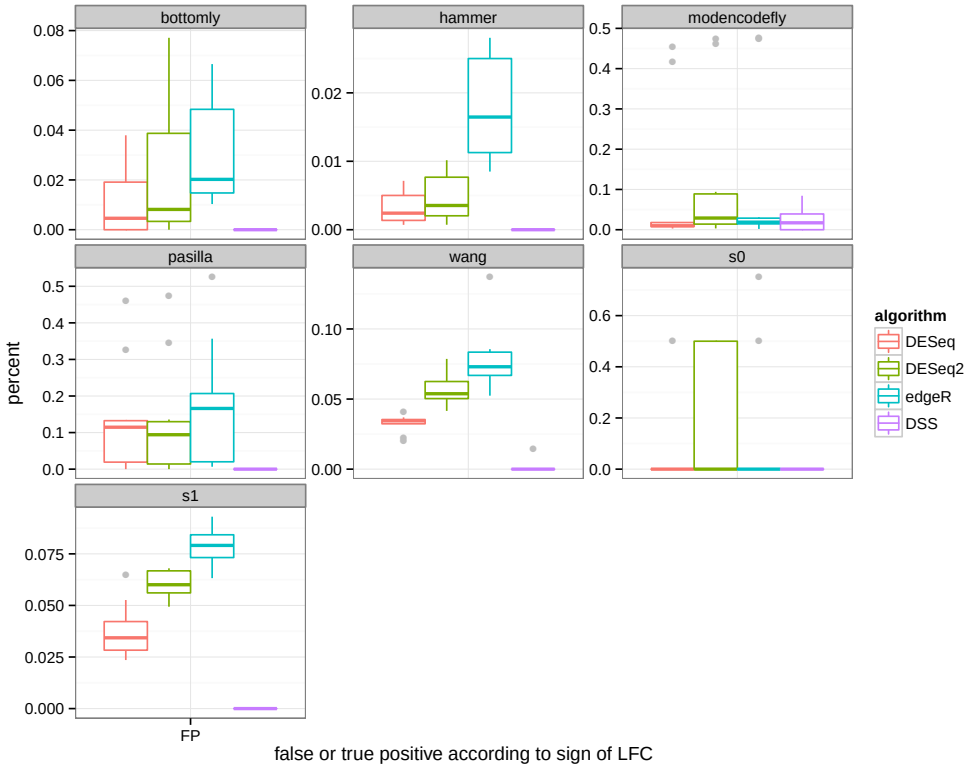

Figure 4.10: False discovery rate at reported FDR < 0.1. Shown is the ratio of false positive calls divided by the total number of positive calls plus 1. The packages almost always control for a 10% false discovery rate. The range of 0-50% FDR for *DESeq2* for the simulated null dataset `s0` comes from 1-2 genes called, as shown in Figure 4.9.

RNA-Seq data. The log fold change prior can be justified by the observation that any controlled experiment or comparison of groups will likely uncover many very small changes and a few large changes. The practical consequence of the log fold change prior is that genes can then be reliably ranked by fold change, arguably as meaningful as ranking genes by p-value.

4.4 Discussion

DESeq2 is based on the generalized linear model, and therefore can be applied easily to small and large datasets, and to datasets with paired design (e.g. tumor/normal comparisons across patients). Global patterns in the distribution of dispersion estimates are used to improve the gene-wise estimates, while genes which do not appropriately fit this common distribution are flagged as dispersion outliers. Genes with low mean counts and high dispersion have log fold changes shrunk toward zero, while genes with high mean counts and low dispersion have nearly the same log fold changes as those calculated from group averages of normalized counts. Genes with individual samples which have a large influence on the log fold change estimates are flagged using Cook's distance, a standard diagnostic for measuring influence in linear and generalized linear models.

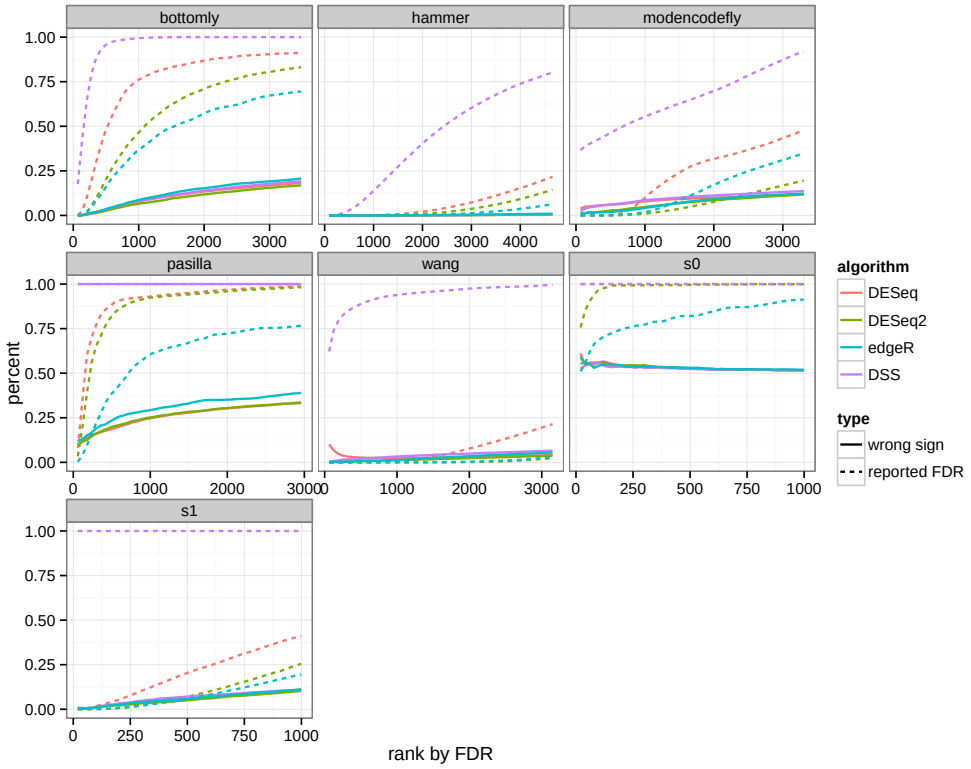

Figure 4.11: Percent of false positives ("wrong sign") and reported FDR for n genes with lowest FDR. For a given number n of genes with the lowest reported FDR in the small subset (x-axis), the percent of genes with the opposite log fold change in the larger subset is plotted (y-axis). Also plotted is the reported FDR for the n-th gene. The four packages have very similar ranking of genes by FDR, so the "wrong sign" curves are nearly identical.

Future improvements of the *DESeq2* methodology include the adjustment of the Wald statistic null distribution, using a t-distribution to replace the normal distribution. The variance of the dispersion prior can be used to inform the degrees of freedom of the t-distribution for the Wald statistics. This can be most easily explained by considering the extremes: as the dispersion prior variance goes to zero, then the MAP dispersions shrink to the fitted line. In this case, the Wald statistics should be normally distributed, assuming the fitted line accurately describes the true dispersion-mean relationship. As the dispersion prior variance grows to infinity, the MAP estimates are equal to the gene-wise estimates, and in this case the Wald statistics should follow a t-distribution with $(m - p)$ degrees of freedom (Supplementary Figure A.1).

The robust log fold change estimates allow for a number of new possibilities for downstream analysis. One such possibility is the use of fold changes for gene set enrichment analysis (GSEA). Gene sets, such as those provided by the Gene Ontology [111], KEGG [112], or Reactome [113] projects, can offer biological interpretations to the results at the end of a differential expression analysis pipeline. It is often of interest to the investigator to know that certain gene sets are overrepresented in the list of top up- or down-regulated genes. One method to determine these overrepresented sets is to perform hypergeometric tests

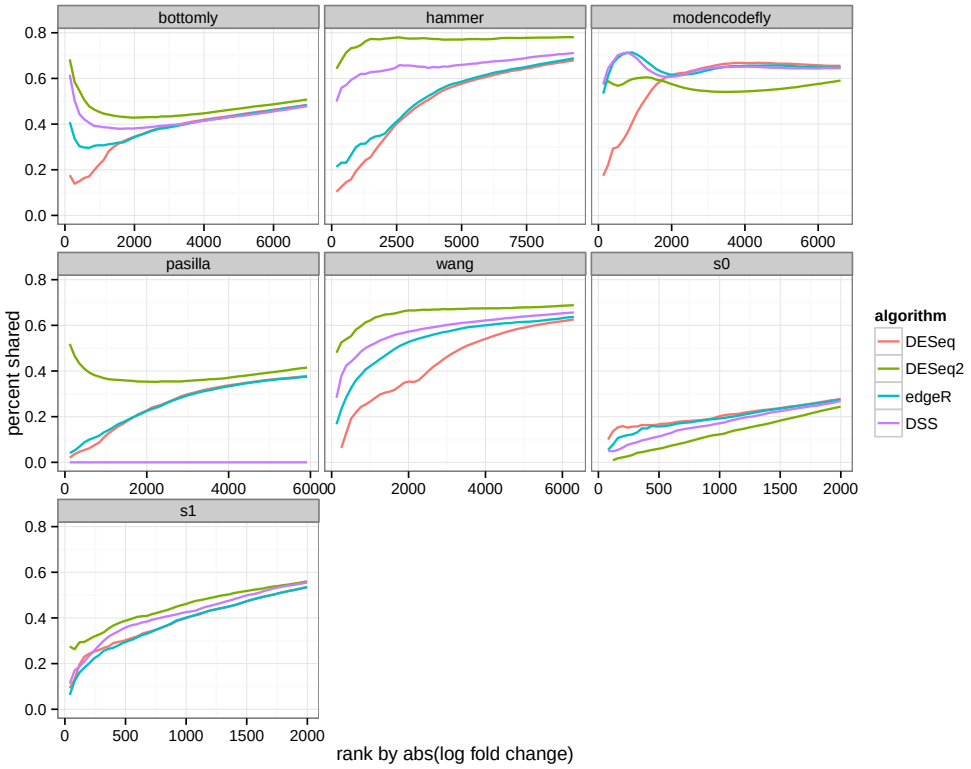

Figure 4.12: Percent of shared genes top-ranked by log fold change in the small and large subsets of data. For a given number n of genes with highest absolute log fold change in the small subset (x-axis), the percent of genes which are shared in the top-ranked genes from the large subset is plotted (y-axis). The *DESeq2* package has the most concordant rankings in five out of the six non-null datasets.

on the intersection of each gene set and a set of significant genes created by thresholding on significance level. The gene sets can then be ranked by the hypergeometric test p-value. One problem with this approach is that the threshold on significance might lead to a situation where small changes to the model or adding a new sample might result in drastic changes to the GSEA results. This is considered by Jaffe et al. [114], who propose bootstrap resampling of the samples in order to derive more stable GSEA results.

An alternative approach possible using the *DESeq2* framework is to define interesting gene sets as those with many genes with high or low MAP log fold changes, so called "continuous GSEA". The MAP log fold changes are less likely than the p-values to change drastically with small changes to the model or the set of samples; this is because p-values involve integration of the tail of distributions which are highly influenced by small changes in sample size and dispersion estimates. Two approaches to continuous GSEA are as follows: (1) for each gene set, perform a t-test of the log fold changes for the genes in the set against the genes out of the set; (2) add the log fold changes for the genes in the set and divide by the square root of the size of the set[5]. In the case of (1), the gene sets can be ranked by the t-statistic, and in the case of (2), the gene sets can be ranked by the absolute value of the scaled sums, which should be approximately normally distributed. The MAP

[5]This approach is demonstrated in the Bioconductor package *Category*.

log fold changes available with *DESeq2* allow for such an analysis, whereas top-ranked unmoderated log fold changes are a mix of "true" large log fold changes and large log fold changes which arise from highly variable, low count genes.

Another novel analysis for RNA-Seq data possible with *DESeq2* is to use the estimates of standard error of log fold changes in order to extract those genes whose log fold change is greater than a specified level, i.e. $\beta_{ij} > \theta$. This might be useful if an investigator is only interested in statistically significant and large log fold changes, rather than all non-zero log fold changes. Such a method has already been proposed for microarray data [115]. Using *DESeq2*, p-values for such a test can be generated in a manner very similar to the Wald test p-values. As with the Wald test, a distribution is centered on the log fold change estimate with variance equal to the squared standard error, though now integration of the tail begins at θ rather than at 0. One can also combine two one-sided tests in order to test for those genes whose log fold changes are less than θ and greater than $-\theta$, i.e. those genes which do not change much across conditions.

Chapter 5

Hierarchical Bayes modeling of cell-type-specific glucocorticoid receptor binding patterns

5.1 Introduction

The glucocorticoid hormone triggers a variety of responses across different cell types, including gluconeogenesis (hence the prefix "gluco-") and suppression of inflammation. Glucocorticoids are commonly used to treat the symptoms of overactive immune systems, as in asthma and autoimmune disorders. The glucocorticoid response is mediated through the glucocorticoid receptor (GR), a transcription factor which binds glucocorticoid (or similar synthetic molecules like dexamethasone). Upon activation by the hormone, GR is able to translocate to the nucleus and bind to a 15 base pair DNA recognition sequence, or "motif". A canonical motif is listed in transcription factor motif databases, though variations of the particular sequence have been shown to have functional consequences on the regulatory activity of GR [116]. GR is also able to bind to DNA indirectly, in binding to a protein which is itself bound to DNA.

If DNA sequence alone determined GR binding patterns, one would expect to find GR bound to millions of sites along the mammalian genome after treating cells with hormone. However, one typically finds tens of thousands of GR binding sites after hormone treatment. One factor which limits the possible universe of sites where GR can bind is chromatin accessibility. A recent study estimates that up to 80-90% of GR binding sites are "predetermined" by pre-hormone-treatment chromatin accessibility [117]. This was demonstrated through GR ChIP-Seq and DNase-Seq of mouse cells before and after hormone stimulation. The post-hormone GR ChIP-Seq peaks can be shown to line up directly with the pre-hormone-treatment DNase hypersensitive sites (DHS). This can also be seen for human cells, as shown in Figure 5.1. Furthermore, Pique-Regi et al. [118] show that, for many transcription factors, the combination of DNase hypersensitivity and a suitable recognition sequence is often enough to predict binding events. Other informative factors influencing binding include histone modifications on nucleosomes near the binding sites [118].

Many bioinformatic analyses in recent years have attempted to determine those chromatin and sequence features which can be used to predict the genome-wide binding patterns of transcription factors. An alternative question, perhaps more relevant to biologists studying cell-type-specific regulatory response, is how the relationship of various chromatin and sequence features to transcription factor binding might change across experiments or across

different cell types. In this chapter, I try to answer the latter question using a statistical framework known as hierarchical models. As reviewed by Ji and Liu [119], hierarchical models are useful for borrowing information across many genomics ranges to estimate model parameters. I will show that these models can also be used to borrow information across experiments.

In this chapter, a hierarchical Bayes model is used to correlate GR binding activity in DHS with a combination of pre-hormone-treatment chromatin features, including chromatin accessibility, histone modifications and histone variants, and sequence features including predicted binding affinity from DNA sequence. The analysis focuses on pre-hormone-treatment DHS, as these constitute the majority of GR binding sites and the DHS provide a convenient, limited set of genomic ranges at which to quantify the various chromatin features. Pre-hormone-treatment chromatin state is measured using publicly available DNase-Seq and histone modification ChIP-Seq experiments for the same cell type as the GR ChIP experiment. The model is interpreted as an exploratory, hypothesis-generating step towards further investigation of how genomic sequence combines with cell-type-specific chromatin state to produce a diversity of cellular responses to hormone.

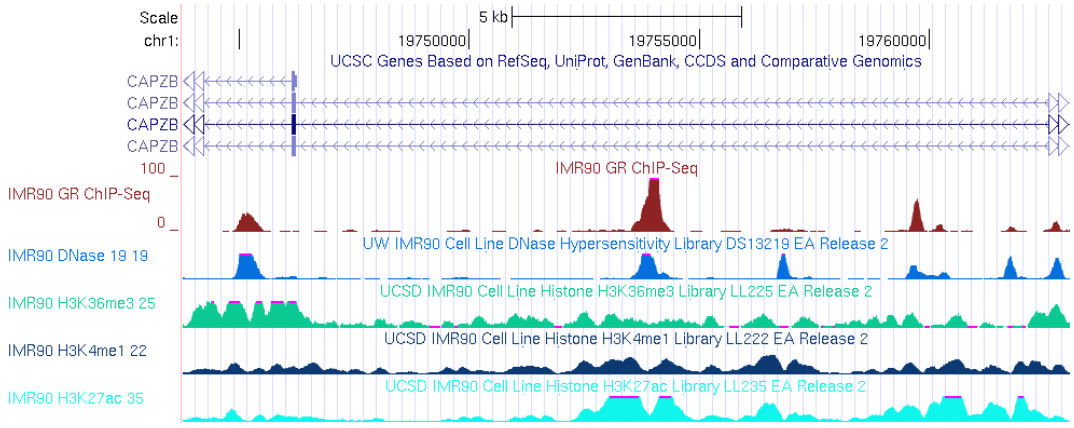

Figure 5.1: GR binds to pre-hormone-treatment DHS. A screenshot from the Roadmap Epigenomics Genome Browser, http://www.epigenomebrowser.org/, showing experimental tracks for the IMR90 cell type. The top track with dark red peaks shows the read density from a GR ChIP-Seq performed in the IMR90 cell type by the Transcriptional Regulation Group (TRG) in Berlin. The next track with blue peaks shows the read density from a DNase-Seq experiment in pre-hormone IMR90 cells performed at the University of Washington. The following three tracks show ChIP-Seq read densities for various histone modifications in pre-hormone IMR90 cells performed at the University of California in San Diego.

5.2 Methods

5.2.1 Sequencing data preparation

Three human cell lines were used for comparison of GR binding patterns and chromatin state:

- A549: adenocarcinomic lung epithelial cell
- IMR90: fetal lung fibroblast
- K562: myelogenous leukemia cell

The DHS as annotated by the ENCODE project [3] for the three cell types were used to compare binding patterns and quantify chromatin state. These DHS are available from the UCSC Genome Browser as "narrowPeak" files[1]. The DHS annotation files were subsequently filtered by excluding regions which had any intersection with regions of the RepeatMasker[2] track with score greater than 1000. The number of remaining DHS for each cell is: 105,121 for A549, 127,803 for IMR90 and 97,304 for K562.

GR ChIP-Seq and control (referred to as "Input") experiments were performed in the three cell types by the Transcriptional Regulation Group (TRG) of the Max Planck Institute for Molecular Genetics in Berlin. The sequenced reads were mapped to the hg19 genome using Bowtie version 1.0.0 [120] with default parameter settings. The count of reads falling in equal sized genomic ranges was taken as a quantitative measure of transcription factor binding, DNase hypersensitivity or histone modification levels. Read counts from a control experiment were also used in modeling. For determining overlap with annotated genomic ranges (promoters, exons, introns), peak calling was performed on the TRG ChIP-Seq data using the MACS software [121] version 1.4.1 with default settings, matching the annotated peaks available from the ENCODE project.

DNase-Seq mapped sequence data (BAM files) from the ENCODE project [3] were downloaded for all cells from the UCSC Genome Browser website. The ChIP-Seq experiment data targeting the following histone variants and histone modifications were also downloaded: H2A.Z, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3, H4K20me1. For A549 and K562 cells, these data were generated by the ENCODE project [3]. For IMR90 histone variant and histone modification data was generated by the Roadmap Epigenomics Mapping project [122], and downloaded from the NCBI GEO website under the series number GSE16256. For A549 cell, two replicates of GR ChIP (stimulated with 100nm dexamethasone) and one replicate of Input (labelled `Rxlch` for reverse crosslinked chromatin) were also downloaded.

GR ChIP, Input, and DNase-Seq reads were counted in 200 base pair genomic ranges centered on the annotated DHS peak from the narrowPeak files. As the histone modifications occur on nucleosomes typically adjacent to the peak of DNase-Seq read density, genomic ranges of 1600 base pair were used for read counting. The `countBamInGRanges` function was used for read counting, from the *exomeCopy* Bioconductor package.

[1]The DHS files of the form `wgEncodeOpenChromDnaseA549Pk.narrowPeak` are available at http://genome.ucsc.edu/

[2]A.F.A. Smit, R. Hubley and P. Green, RepeatMasker at http://repeatmasker.org, downloaded Feb. 2012.

5.2.2 Motif score calculation

Motif scores were calculated using the GR motif MA0113.1 from the JASPAR database [123] and scanning both strands in a 180 base pair range centered on the DHS peak. The maximum log odds score was assigned to each DHS location. Scanning was performed using the `PWMscoreStartingAt` function of the *Biostrings* Bioconductor package, with a 0-order background model for DNA with 42% GC content.

5.2.3 Hierarchical Bayes modeling

A hierarchical Bayes model [124] was constructed to correlate binding of GR (as measured by ChIP-Seq read counts) with chromatin features and motif score in DHS across experiments and across cell types. The log of read counts for various chromatin features (plus a pseudocount of 1) and the motif score over the annotated DHS of a cell type are arranged as columns of a matrix X, depicted on the right side of Figure 5.2. The chromatin feature matrix X is identical for experiments of the same cell type, except the Input feature, which is paired with the GR ChIP experiment: each of the TRG IMR90 ChIP-Seq experiments is paired with its own Input experiment, and the A549 ENCODE GR ChIP experiments are paired with the A549 ENCODE Input experiment. This matrix X is then centered and scaled to have columns with zero mean and unit standard deviation. For an experiment k and a genomic range i centered on a DHS, the count of GR ChIP-Seq reads is written as K_{ik}, following a Poisson distribution:

$$\begin{aligned} K_{ik} &\sim \text{Poisson}(\mu_{ik}) \\ \log(\mu_{ik}) &= \beta_{0k} + X_{i*k}\vec{\beta}_{*k} \end{aligned} \tag{5.1}$$

The β_{0k} coefficient is the intercept, which controls for sequencing depth. X_{i*k} is the i-th row of the matrix X for the cell type of sample k. The β_{jk} coefficient is the multiplicative effect of chromatin feature j on the GR ChIP-Seq read counts for experiment k. The β coefficients are given normal prior distributions. The distribution of K_{ij} can then be described as Poisson log normal, because the log of the mean parameter is a linear combination of normal random variables (each multiplied by a constant value in X_{i*k}). To specify that β_{jk} for experiments k of the same cell type ($\text{ct}(k)$) are related, they are given a shared prior mean $\nu_{j,\text{ct}(k)}$:

$$\begin{aligned} \beta_{0k} &\sim \mathcal{N}(0, \sigma_0^2) \\ \beta_{jk} &\sim \mathcal{N}(\nu_{j,\text{ct}(k)}, \sigma_\beta^2) \end{aligned} \tag{5.2}$$

where $\nu_{j,\text{ct}(k)}$ is a variable summarizing the effect of chromatin feature j across all experiments k which share the same cell type, $\text{ct}(k)$. $\nu_{j,\text{ct}(k)}$ has a prior distribution centered

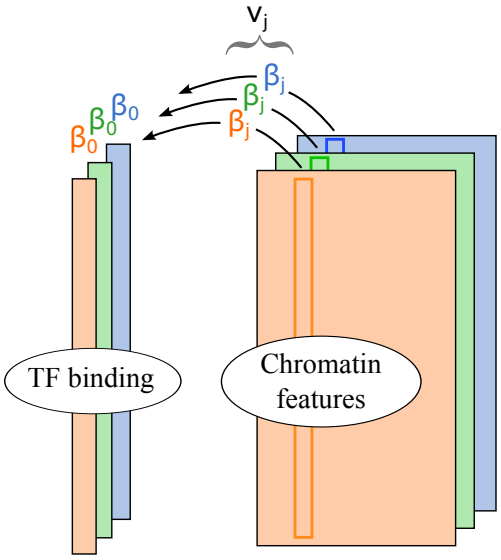

Figure 5.2: The hierarchical model for a single cell type, used to compare the correlation of chromatin features on transcription factor binding. In this diagram, the transcription factor binding strength as measured by ChIP-Seq reads (columns on left) is modeled based on various chromatin features (matrices on the right). The colors represent multiple experiments for a given cell type, though only a single matrix of chromatin features is currently available for each cell type. The height of the matrices represents the number of DHS. The coefficients β_j for the j-th chromatin feature share a common prior distributed with mean ν_j.

on a variable λ_j, which summarizes the effect of chromatin feature j across all cell types. Each cell type has its own variance term $\sigma^2_{\nu_{\text{ct}(k)}}$:

$$\nu_{j,\text{ct}(k)} \sim \mathcal{N}(\lambda_j, \sigma^2_{\nu_{\text{ct}(k)}}) \tag{5.3}$$

λ_j has a prior centered on zero:

$$\lambda_j \sim \mathcal{N}(0, \sigma^2_\lambda) \tag{5.4}$$

The practical consequence of a zero-mean prior is to prefer a small value, unless the likelihood function is strongly peaked at a non-zero value. The square root of the variances σ^2 used in the above equations are all given a gamma prior with mean 1:

$$\sigma_0, \sigma_\beta, \sigma_{\nu_{\text{ct}(k)}}, \sigma_\lambda \sim \Gamma(\alpha = 10, \beta = 10) \tag{5.5}$$

A diagram of the levels of the hierarchical model is presented in Figure 5.3. The posterior of the model parameters conditioning on the observed data was sampled using the Stan C++ MCMC package and *rstan* R package[3] [125]. The model was run for 4 chains for 4000 iterations (the first 2000 iterations discarded as burn-in) using the "no U-turn"

[3]Stan project description at http://mc-stan.org/.

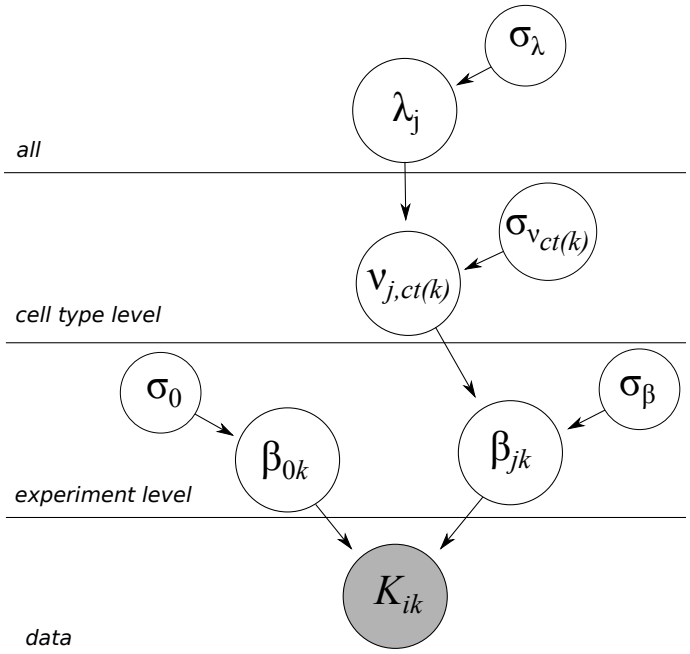

Figure 5.3: The full hierarchical model provided in Eqs. 5.1-5.5. The grey node at the bottom indicates the data: the read counts at genomic range i for experiment k, with a distribution depending on its parent nodes. The nodes β_{0k} and β_{jk} are the intercept and multiplicative effects of chromatin feature j in experiment k. The node $\nu_{j,\mathrm{ct}(k)}$ is the effect of chromatin feature j in the cell type of experiment k. The node λ_j is the effect of chromatin feature j across all cell types. The $\nu_{j,\mathrm{ct}(k)}$ are particularly of interest, as these variables summarize the cell-type-specific chromatin effect across multiple experiments.

setting. `R-hat` values near 1 were used as a convergence diagnostic, provided in Supplementary Table B.6 [126]. Example code for the `stan` model specification is provided in the Appendix C.

5.3 Results

5.3.1 Genomic location of GR binding

Before interpreting the results of the hierarchical model, a comparison is made of the genomic locations of DHS and those DHS which are bound by GR. This comparison shows that GR binding in open chromatin is depleted of open promoters (Figure 5.4). Across all cell types examined by the TRG, the amount of promoter-proximal DHS which are bound by GR is about half of the total promoter-proximal DHS. This finding matches that of John et al. [117] and Grontved et al. [127] found in mouse cells, with a similar amount of depletion of promoter-proximal GR binding sites when comparing the proportion of pre-hormome-treated promoter-proximal DHS.

Genomic location analysis was also performed on the annotated binding peaks of various proteins in A549 cells from the ENCODE project using the annotated DHS for A549 cells. The corresponding plots in Figure 5.5 show that certain proteins associated with transcription, such as Pol2, TAF1, PBX3, and ETS1, when bound to DHS are enriched near promoters. Other proteins show moderate promoter-proximal depletion, including FOSL2, a component of the AP-1 complex, a known cofactor of GR [128]. However, the depletion observed by John et al. [117], Grontved et al. [127] and in the data of the TRG is not consistently reproduced in the ENCODE A549 GR ChIP-Seq experiments, with one replicate showing a slight increase of promoter-proximal binding and one showing a slight decrease.

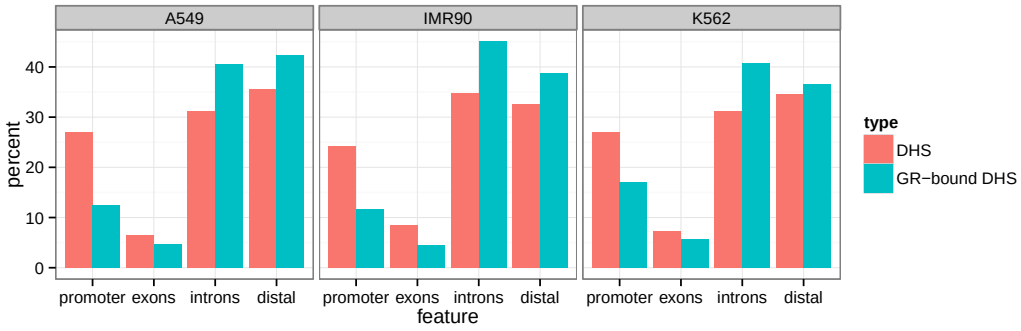

Figure 5.4: Genomic location of DHS and GR-bound DHS, using ChIP-Seq data performed by the TRG. The number of promoter-proximal (±2.5 kb from TSS) DHS which are bound by GR upon hormone treatment ranges from around 50-75% of the number of promoter-proximal DHS. GR tends to bind instead to intronic and distal DHS.

5.3.2 Interpretation of hierarchical model parameters

The hierarchical model allows comparisons to be made across experiments and across cell types. The posterior distributions for parameters at all levels of the model are plotted for each chromatin feature and motif score in Figure 5.6. The experiment-level parameters, β, have very narrow posterior distributions. The cell-type-level and across-cell-type parameters ν and λ have wider posteriors which sometimes overlap zero, indicating the uncertainty of these effects when experiments or cell types have parameters with different sign. The cell type parameter ν has different variance for each cell type, with large sample size (A549 has three experiments, IMR90 has two and K562 has only one) providing smaller variance estimates.

The small intervals for the β posterior distributions can be explained by the very large number of genomic ranges represented in the data. For a simple Bayes calculation, as the sample size increases, the mean of the posterior distribution converges to the MLE and the standard deviation of the posterior converges to the standard error of the MLE [129]. This can be restated, that the likelihood becomes increasingly peaked at the MLE, and therefore the prior has negligible effect on the posterior. The posterior distributions for

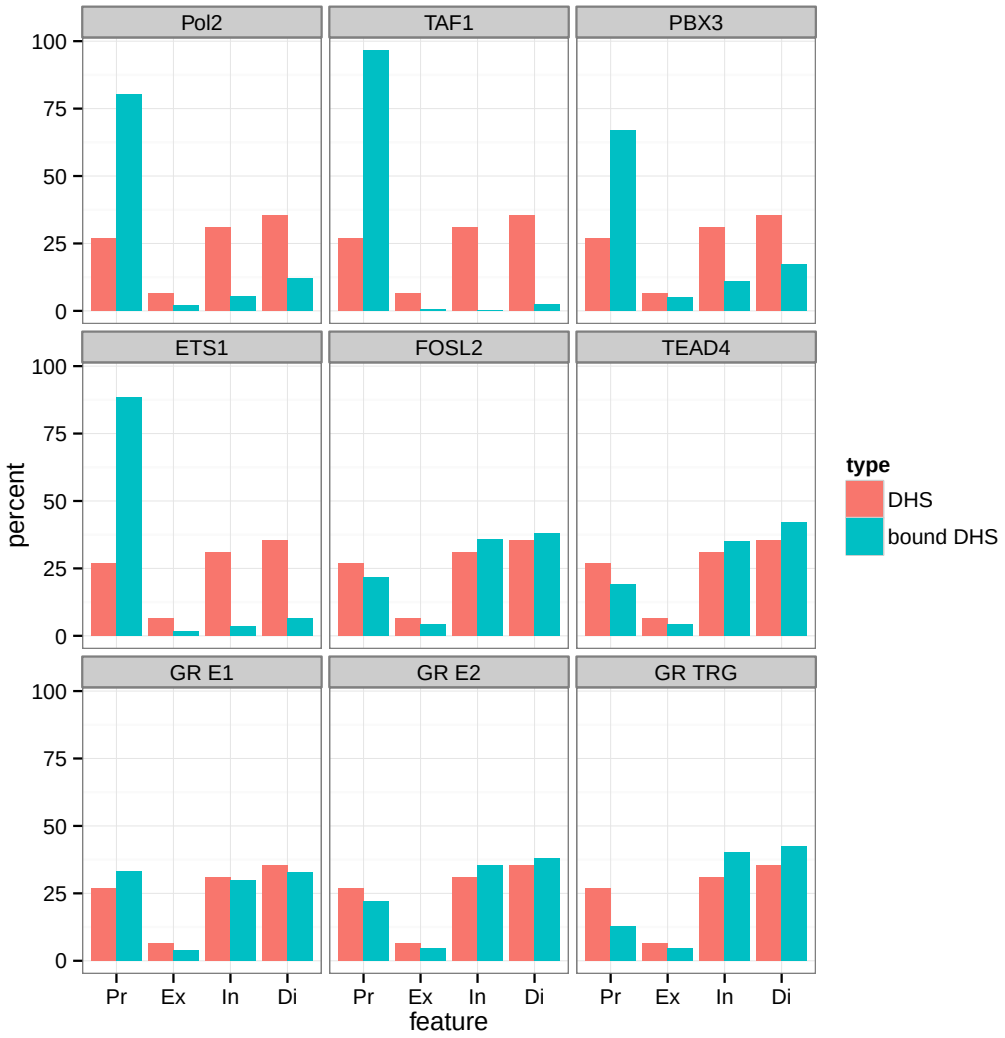

Figure 5.5: Genomic location of DHS and protein-bound DHS in A549 cells (Pr = promoters, Ex = exons, In = introns, Di = distal). All plots show annotated ChIP-Seq peaks from the ENCODE project except the last, which is the A549 ChIP-Seq peaks from the TRG also shown in Figure 5.4. "GR E1" and "GR E2" represent two replicates of a GR ChIP-Seq experiment in hormone-treated cells.

experiment level parameters β are very similar to the maximum likelihood estimates using a Poisson generalized linear model, as is shown in Supplemental Table B.7.

The parameters are mostly consistent across cell types and in general reflect known properties of distal regulatory elements and enhancers. For example, two consistently positive parameters are for H3K4me1, typically used to identify enhancers, and for H3K27ac, which is used to identify active enhancers [130]. Other consistently positive parameters are found for DNase-Seq and Input (the control experiment). Though the genomic ranges used for modeling are all DHS, increased DNase hypersensitivity as measured by the number of DNase-Seq reads correlates with higher GR binding, which accords with the literature regarding GR [117] and transcription factors in general [118].

DHS with high scoring GR motif also correlate with higher levels of GR binding, as would be expected. As the chromatin features and sequence features are all scaled to have unit variance, the size of the coefficients can be used to compare overall association of features with GR binding. In this respect, one standard deviation of the motif score has comparable positive effect on binding as one standard deviation of DNase, although H3K27ac appears even more associated with GR binding than motif score.

Three consistently negatively associated chromatin features are H3K36me3, H3K79me2 and H2A.Z. H3K36me3, typically found along the gene body, is associated with transcription elongation [131]. H3K36me3 has been linked to histone deacetylation (removal of acetyl groups from histone tails) to suppress cryptic transcription [132]. Cryptic transcription occurs when an RNA polymerase starts transcribing from an intragenic region rather than binding to the promoter. Therefore the negative association with H3K36me3 might reflect an underlying mechanism by which GR is directed to more distal regulatory elements. An explanation for the consistent negative coefficient for H3K79me2, a histone modification linked to cell cycle, is not obvious. The consistent negative association of H2A.Z with GR binding might at first appear incompatible with the findings reported by John et al. [133], that H2A.Z in highly enriched at GR binding sites. However, this study compared H2A.Z levels at GR binding sites with two nearby regions which were not DHS, while the model presented here focuses only on the universe of DHS. H2A.Z might therefore be correlated with GR binding patterns genome-wide, though negatively correlated when limiting the comparison to only DHS. As the large majority of GR binding sites are in DHS, the latter might be a more meaningful comparison.

5.3.3 Cell-type-specific parameters are typical promoters marks

A number of chromatin features have parameters with different sign across cell types, including H3K27me3 and H3K4me3. H3K4me3 is a histone modification typically found at high-CpG promoters (HCP) [134], and among low-CpG promoters (LCP) has been shown to have high predictive power for active transcription [135]. The negative correlation of H3K4me3 with GR binding in A549 and IMR90 cells might therefore be related to the depletion of promoters in the set of DHS bound by GR. In addition, some chromatin features have parameters with a different sign within a cell type, notable the H3K9ac parameter for A549 cells. This modification has a strong negative parameter for the ChIP-Seq experiment performed by the TRG, but a positive parameter for both replicates from the ENCODE project. H3K9ac is a histone modification typical for promoters of actively transcribed genes and is present, at "bivalent promoters" marked with both H3K4me3 and H3K27me3 [136]. These cell-type- and experiment-specific associations are being investigated further by the TRG.

5.3.4 Explanatory power of the model

The percent of variance of log-scale GR binding explained by the hierarchical model across the different experiments ranges from 31% to 59%. The percent variance for each exper-

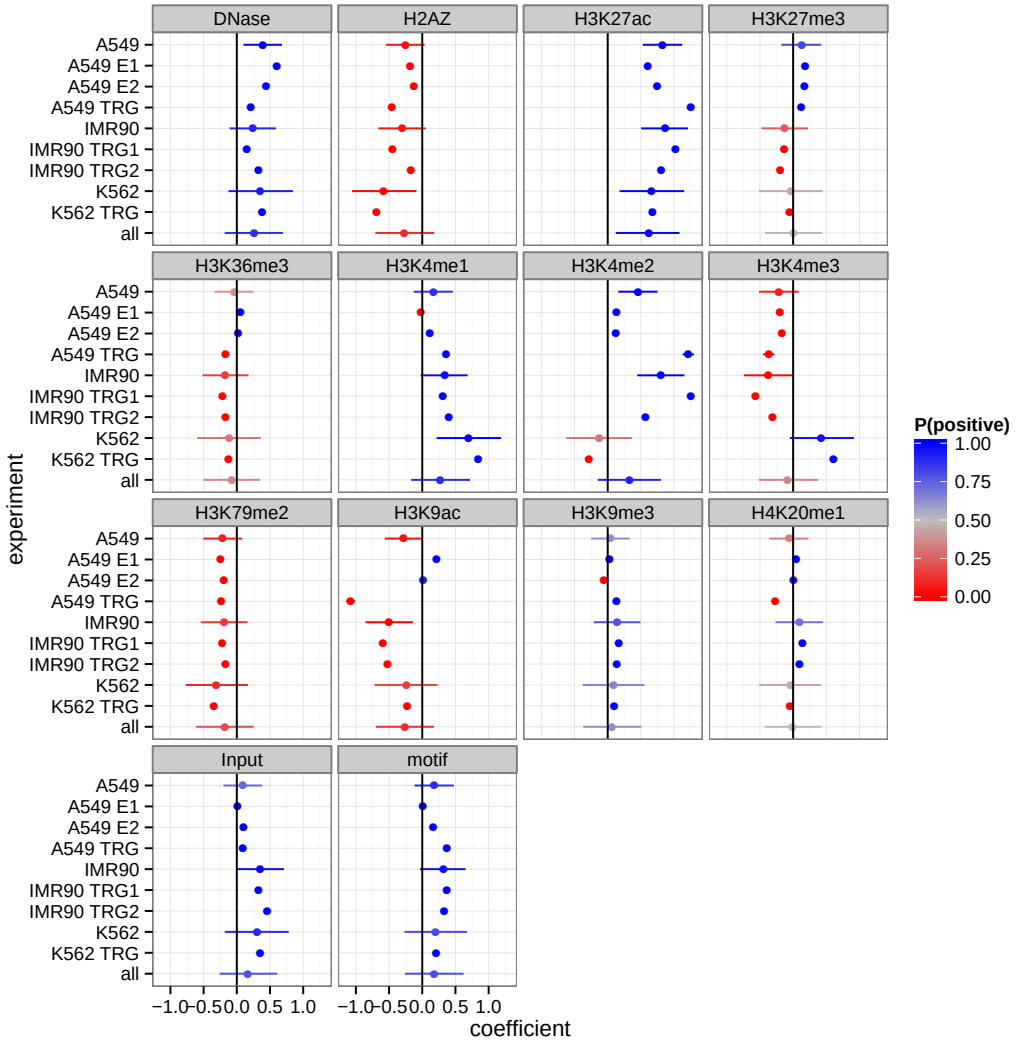

Figure 5.6: Posterior mean and 95% quantile-based interval for parameters of the model. The individual experiment parameters, β, are labeled with the cell type, lab and replicate number, e.g. "A549 E1" for A549 cell type, ENCODE project, replicate 1. The cell type parameters, ν, are labelled only with the cell names, e.g. "A549". The parameter across all cell types, λ, is labelled "all". The parameters are mostly consistent in terms of sign across experiments and cell types, with exceptions including H3K27me3, H3K4me2, H3K4me3 and H3K9ac. The individual experiment parameters have comparatively small 95% intervals due to the large number of genomic ranges used for modeling.

iment is provided in Table 5.1. These are the squared Pearson correlations between log GR ChIP-Seq read counts and the log of the fitted means, $\log(\mu_{ik})$. The fitted means are obtained from the posterior means of the β coefficients of the hierarchical model. As the model is built on a quantitative measure of binding, the results are not directly comparable with other methods which mainly focus on prediction of a binary variable indicating the binding of a protein [118]. As the number of locations (on the order of 10^5) is much larger than the number of features (14), the explanatory power of the model can be described by comparing the training data K_{ik} with the fitted values. If the number of features were close to the number of locations, then using the training data for comparison would be problematic due to overfitting of model parameters.

cell type	lab	% variance
A549	TRG	31
A549	ENCODE	35
A549	ENCODE	35
IMR90	TRG	57
IMR90	TRG	50
K562	TRG	59

Table 5.1: Percent variance of log GR ChIP-Seq read counts explained by the hierarchical model.

5.3.5 GR motif score distribution at DHS and promoters

A simple explanation for the depletion of promoter-proximal GR binding is that the promoters lack the proper motif to support GR binding. To investigate the hypothesis that promoter-proximal depletion is driven by motif score, the distribution of motif scores were calculated for DHS grouped by GR peak presence and promoter proximity. The score distributions are shown in Figure 5.7. Average motif scores for GR-bound DHS are slightly elevated above the non-bound DHS and scores for distal DHS are slightly elevated over proximal. Overall, the groups have largely overlapping motif score distributions, so a more formal approach is needed to test the hypothesis.

The contribution of motif score and promoter proximity to GR binding was quantified and statistically tested using an analysis of deviance. The effect of promoter proximity on GR binding can be tested, while controlling for a lack of sequence motif which might also explain the observed depletion of promoter-proximal binding. A binary variable was defined for each DHS, indicating if the DHS overlapped a GR peak. The GR peak variable was modeled using a logistic regression, with independent variables including motif score vs a model with motif score and an indicator of promoter proximity. Table 5.2 shows a comparison of deviance of the GR peak variable explained by promoter proximity, controlling for motif score, as a percent of the deviance explained by motif score alone. The deviance explainable by a parameter is defined as the reduction in residual deviance from adding that term to the model. The p-values for the analysis of residual deviance explained by promoter proximity were less than 2×10^{-16}, indicating that promoter proximity explains a significant portion of the presence of GR peaks, while controlling for GR motif score.

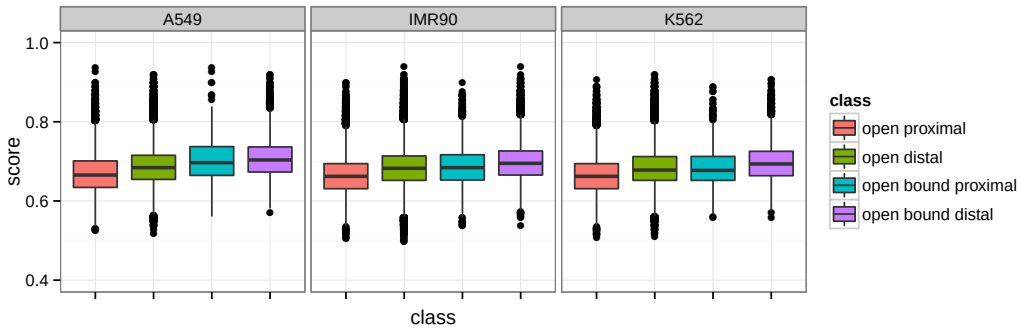

Figure 5.7: Distribution of motif scores for various groups: DHS sites which are proximal or distal to promoters, and DHS sites which are bound by GR and proximal or distal to promoters. Bound DHS have slightly higher score distributions compared to the universe of all DHS, and distal DHS have slightly higher score distributions than promoter-proximal DHS. Distributions of closed and open binding sites are compared in Supplementary Figure A.4

Therefore, the depletion of GR binding at promoter-proximal DHS cannot be explained entirely by the lack of high scoring sequences for the GR motif.

cell type	% deviance
A549	45
IMR90	52
K562	36

Table 5.2: Percent deviance of GR peak presence at DHS explained by promoter proximity while controlling for motif score, as a percent the deviance explained by motif score alone.

5.4 Discussion

In this chapter, I present a hierarchical model used to correlate the binding patterns of the glucocorticoid receptor to chromatin and sequence features in sites of chromatin accessibility. This Bayesian model allows for the posterior distributions of parameters to be compared across experiments and cell types. The results of the model are useful for hypothesis generation, leading to experiments which can test the causality of any interesting relationships which arise from the model. This follow-up experimentation is critical to determine whether histone modifications or other proteins associated with histone modifications might somehow exert an influence on transcription factor binding, or whether they are merely correlative.

As the modeling was performed in collaboration with the TRG, a number of hypotheses from this project are now being tested using GAL4 fusion proteins. In these experiments, a binding site for the protein GAL4 is positioned upstream from a GR binding site which is itself upstream from a minimal promoter and a luciferase reporter gene. The luciferase reporter allows for quantitive measurement of the regulatory activity of the GR. A protein is then added which consists of the GAL4 DNA-binding domain fused to an enzyme which

can deposite the modification of interest, ideally with some specificity for the residue of the histone tail (e.g. deposit trimethylation on the H3K4 residue). This experiment tests whether a perturbation of the system – e.g. adding H3K4me3 – can reduce GR regulatory activity as measured through the transcription of the reporter gene. Preliminary results with GAL4 fusion proteins in U2OS cells (a human osteosarcoma cell line) suggest that trimethylation of H3K4 mediated by the WDR5 enzyme, acetylation of H3K9 mediated by the GCN5 enzyme, and recruitment of H2A.Z all result in reduced GR-dependent transcription at the reporter, arguing for a causative connection between these histone modifications to GR binding and regulatory activity.

The hierarchical model described in this chapter, while providing a framework for picking apart experimental effects, does have room for improvement. Sampling the posterior is slow when using all DHS ($\sim 10^5$ per cell type) and all of the chromatin features assayed by the ENCODE project across cell types (4,000 MCMC iterations last around 4 hours for 20,000 locations). One solution would be to reduce the number of chromatin features through principal component analysis, as the chromatin features are highly correlated and might lead to many rejected proposals in the parameter space and extended running time. However, linear combinations of features are not as meaningful to biological collaborators as the chromatin features themselves. Another limitation of the current model is that it only includes a single sequence feature, representing the best match in the DHS to the canonical GR motif; an expanded model which allows for alternative motifs might be better at explaining the variance seen in GR binding. The model would also benefit from other datasets for the transcription factor ChIP-Seq, which might help resolve the inconsistent parameters for some chromatin features seen across labs. An extension which offers a layer for multiple replicates of the chromatin feature matrix X is also theoretically possible, although this would lead to increased difficulty in sampling due to the increased size of the parameter space.

Bibliography

[1] Michael I. Love, Alena Myšičková, Ruping Sun, Vera Kalscheuer, Martin Vingron, and Stefan A. Haas. Modeling read counts for CNV detection in exome sequencing data. *Stat Appl Genet Mol Biol*, 10(1), November 2011.

[2] J. B. Gurdon, T. R. Elsdale, and M. Fischberg. Sexually mature individuals of xenopus laevis from the transplantation of single somatic nuclei. *Nature*, 182(4627): 64–65, July 1958.

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

[4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, October 1999.

[5] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, October 2000.

[6] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99(10):6567–6572, May 2002.

[7] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *J Comput Biol*, 6(3-4):281–297, 1999.

[8] Anja von Heydebreck, Wolfgang Huber, Annemarie Poustka, and Martin Vingron. Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics*, 17(suppl 1):S107–S114, June 2001.

[9] Jamal Tazi, Nadia Bakkour, and Stefan Stamm. Alternative splicing and disease. *Biochim Biophys Acta*, 1792(1):14–26, January 2009.

[10] Donald E. Olins and Ada L. Olins. Chromatin history: our view from the bridge. *Nat Rev Mol Cell Biol*, 4(10):809–814, October 2003.

[11] Mikael Huss. Introduction into the analysis of high-throughput-sequencing based epigenome data. *Brief Bioinform*, 11(5):512–523, September 2010.

[12] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, Scott S. Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, September 2012.

[13] Peter J. Campbell, Philip J. Stephens, Erin D. Pleasance, Sarah O'Meara, Heng Li, Thomas Santarius, Lucy A. Stebbings, Catherine Leroy, Sarah Edkins, Claire Hardy, Jon W. Teague, Andrew Menzies, Ian Goodhead, Daniel J. Turner, Christopher M. Clee, Michael A. Quail, Antony Cox, Clive Brown, Richard Durbin, Matthew E. Hurles, Paul A. W. Edwards, Graham R. Bignell, Michael R. Stratton, and P. Andrew Futreal. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*, 40(6):722–729, June 2008.

[14] Seungtai Yoon, Zhenyu Xuan, Vladimir Makarov, Kenny Ye, and Jonathan Sebat. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res*, 19(9):1586–1592, September 2009.

[15] Can Alkan, Jeffrey M. Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O. Kitzman, Carl Baker, Maika Malig, Onur Mutlu, S. Cenk Sahinalp, Richard A. Gibbs, and Evan E. Eichler. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet*, 41(10):1061–1067, October 2009.

[16] Marc Sultan, Marcel H. Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, Tatjana Borodina, Aleksey Soldatov, Dmitri Parkhomchuk, Dominic Schmidt, Sean O'Keeffe, Stefan Haas, Martin Vingron, Hans Lehrach, and Marie-Laure Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960, August 2008.

[17] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods*, 5(7):621–628, July 2008.

[18] John C. Marioni, Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18(9):1509–1517, September 2008.

[19] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[20] Thomas D. Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, April 2010.

[21] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, November 2009.

[22] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106+, October 2010.

[23] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, May 2010.

[24] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome Res*, 22(10):2008–2017, October 2012.

[25] David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-Wide mapping of in vivo Protein-DNA interactions. *Science*, 316(5830):1497–1502, June 2007.

[26] Gregory E. Crawford, Ingeborg E. Holt, James Whittle, Bryn D. Webb, Denise Tai, Sean Davis, Elliott H. Margulies, YiDong Chen, John A. Bernat, David Ginsburg, Daixing Zhou, Shujun Luo, Thomas J. Vasicek, Mark J. Daly, Tyra G. Wolfsberg, and Francis S. Collins. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res*, 16(1):123–131, January 2006.

[27] Alan P. Boyle, Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. High-Resolution mapping andcharacterization of open chromatin across the genome. *Cell*, 132(2):311–322, January 2008.

[28] Hualin Xi, Hennady P. Shulha, Jane M. Lin, Teresa R. Vales, Yutao Fu, David M. Bodine, Ronald D. G. McKay, Josh G. Chenoweth, Paul J. Tesar, Terrence S. Furey, Bing Ren, Zhiping Weng, and Gregory E. Crawford. Identification and characterization of cell TypeSpecific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet*, 3(8):e136+, August 2007.

[29] Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shoresh, Lucas D. Ward, Charles B. Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis, and Bradley E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, May 2011.

[30] Lingyun Song, Zhancheng Zhang, Linda L. Grasfeder, Alan P. Boyle, Paul G. Giresi, Bum-Kyu Lee, Nathan C. Sheffield, Stefan Gräf, Mikael Huss, Damian Keefe, Zheng Liu, Darin London, Ryan M. McDaniell, Yoichiro Shibata, Kimberly A. Showers, Jeremy M. Simon, Teresa Vales, Tianyuan Wang, Deborah Winter, Zhuzhu Zhang, Neil D. Clarke, Ewan Birney, Vishwanath R. Iyer, Gregory E. Crawford, Jason D.

Lieb, and Terrence S. Furey. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res*, 21(10):1757–1767, October 2011.

[31] Roger S. Lasken. Single-cell genomic sequencing using multiple displacement amplification. *Curr Opin Microbiol*, 10(5):510–516, October 2007.

[32] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W. Richard McCombie, James Hicks, and Michael Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341): 90–94, April 2011.

[33] Iwanka Kozarewa, Zemin Ning, Michael A. Quail, Mandy J. Sanders, Matthew Berriman, and Daniel J. Turner. Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods*, 6(4):291–295, April 2009.

[34] Yuval Benjamini and Terence P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res*, 40(10):e72, May 2012.

[35] R. A. Fisher. The negative binomial distribution. *Ann Eug*, 11(1):182–187, January 1941.

[36] F. J. Anscombe. The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics*, 5(2):165+, June 1949.

[37] C. I. Bliss and R. A. Fisher. Fitting the negative binomial distribution to biological data. *Biometrics*, 9(2), 1953.

[38] M. G. Bulmer. On fitting the poisson lognormal distribution to Species-Abundance data. *Biometrics*, 30(1):101–110, March 1974.

[39] Naim Rashid, Paul Giresi, Joseph Ibrahim, Wei Sun, and Jason Lieb. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol*, 12(7):R67+, 2011.

[40] Steven P. Lund, Dan Nettleton, Davis J. McCarthy, and Gordon K. Smyth. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat Appl Genet Mol Biol*, 11(5), 2012.

[41] R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the GaussNewton method. *Biometrika*, 61(3):439–447, December 1974.

[42] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat*, 5(3):1752–1779, 2011.

[43] Shengyu Ni and Martin Vingron. R2KS: a novel measure for comparing gene expression based on ranked gene lists. *J Comput Biol*, 19(6):766–775, June 2012.

[44] Jun Li and Robert Tibshirani. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-seq data. *Stat Methods Med Res*, November 2011.

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

[46] Wolfgang Huber, Anja von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1(suppl 1): S96–S104, July 2002.

[47] B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18(suppl 1): S105–S110, July 2002.

[48] Charity W. Law, Yunshun Chen, Wei Shi, and Gordon K. Smyth. Voom! precision weights unlock linear model analysis tools for RNA-seq read counts. Technical report, Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, May 2013.

[49] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *J R Stat Soc*, 135(3):370–384, 1972.

[50] R. Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, February 1977.

[51] Jonathan Sebat, B. Lakshmi, Dheeraj Malhotra, Jennifer Troge, Christa Lese-Martin, Tom Walsh, Boris Yamrom, Seungtai Yoon, Alex Krasnitz, Jude Kendall, Anthony Leotta, Deepa Pai, Ray Zhang, Yoon-Ha Lee, James Hicks, Sarah J. Spence, Annette T. Lee, Kaija Puura, Terho Lehtimäki, David Ledbetter, Peter K. Gregersen, Joel Bregman, James S. Sutcliffe, Vaidehi Jobanputra, Wendy Chung, Dorothy Warburton, Mary-Claire King, David Skuse, Daniel H. Geschwind, T. Conrad Gilliam, Kenny Ye, and Michael Wigler. Strong association of de novo copy number mutations with autism. *Science*, 316(5823):445–449, April 2007.

[52] Joseph T. Glessner, Kai Wang, Guiqing Cai, Olena Korvatska, Cecilia E. Kim, Shawn Wood, Haitao Zhang, Annette Estes, Camille W. Brune, Jonathan P. Bradfield, Marcin Imielinski, Edward C. Frackelton, Jennifer Reichert, Emily L. Crawford, Jeffrey Munson, Patrick M. A. Sleiman, Rosetta Chiavacci, Kiran Annaiah, Kelly Thomas, Cuiping Hou, Wendy Glaberson, James Flory, Frederick Otieno, Maria Garris, Latha Soorya, Lambertus Klei, Joseph Piven, Kacie J. Meyer, Evdokia Anagnostou, Takeshi Sakurai, Rachel M. Game, Danielle S. Rudd, Danielle Zurawiecki, Christopher J. McDougle, Lea K. Davis, Judith Miller, David J. Posey, Shana Michaels, Alexander Kolevzon, Jeremy M. Silverman, Raphael Bernier, Susan E. Levy, Robert T. Schultz, Geraldine Dawson, Thomas Owley, William M. McMahon, Thomas H. Wassink, John A. Sweeney, John I. Nurnberger, Hilary Coon, James S. Sutcliffe, Nancy J. Minshew, Struan F. A. Grant, Maja Bucan, Edwin H. Cook,

Joseph D. Buxbaum, Bernie Devlin, Gerard D. Schellenberg, and Hakon Hakonarson. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, 459(7246):569–573, April 2009.

[53] David St Clair. Copy number variation and schizophrenia. *Schizophr Bull*, 35(1): 9–12, January 2009.

[54] Enrique Gonzalez, Hemant Kulkarni, Hector Bolivar, Andrea Mangano, Racquel Sanchez, Gabriel Catano, Robert J. Nibbs, Barry I. Freedman, Marlon P. Quinones, Michael J. Bamshad, Krishna K. Murthy, Brad H. Rovin, William Bradley, Robert A. Clark, Stephanie A. Anderson, Robert J. O'Connell, Brian K. Agan, Seema S. Ahuja, Rosa Bologna, Luisa Sen, Matthew J. Dolan, and Sunil K. Ahuja. The influence of CCL3L1 Gene-Containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, 307(5714):1434–1440, March 2005.

[55] I. Madrigal, L. Rodríguez-Revenga, L. Armengol, E. González, B. Rodriguez, C. Badenas, A. Sánchez, F. Martínez, M. Guitart, I. Fernández, J. A. Arranz, Mi Tejada, L. A. Pérez-Jurado, X. Estivill, and M. Milà. X-chromosome tiling path array detection of copy number variants in patients with chromosome x-linked mental retardation. *BMC Genomics*, 8:443+, November 2007.

[56] Dirk-Jan Kleinjan and Veronica van Heyningen. Position effect in human genetic disease. *Hum Mol Genet*, 7(10):1611–1618, September 1998.

[57] Donald F. Conrad, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, T. Daniel Andrews, Chris Barnes, Peter Campbell, Tomas Fitzgerald, Min Hu, Chun H. Ihm, Kati Kristiansson, Daniel G. MacArthur, Jeffrey R. MacDonald, Ifejinelo Onyiah, Andy W. Pang, Sam Robson, Kathy Stirrups, Armand Valsesia, Klaudia Walter, John Wei, Chris Tyler-Smith, Nigel P. Carter, Charles Lee, Stephen W. Scherer, and Matthew E. Hurles. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, October 2009.

[58] Andy Pang, Jeffrey MacDonald, Dalila Pinto, John Wei, Muhammad Rafiq, Donald Conrad, Hansoo Park, Matthew Hurles, Charles Lee, J. Craig Venter, Ewen Kirkness, Samuel Levy, Lars Feuk, and Stephen Scherer. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol*, 11(5):R52+, 2010.

[59] Paul Medvedev, Monica Stanciu, and Michael Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods*, 6(11 Suppl):S13–S20, November 2009.

[60] E. S. Venkatraman and Adam B. Olshen. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663, March 2007.

[61] J. Fridlyand. Hidden markov models approach to the analysis of array CGH data. *J Mult Anal*, 90(1):132–153, July 2004.

[62] J. C. Marioni, N. P. Thorne, and S. Tavaré. BioHMM: a heterogeneous hidden markov model for segmenting array CGH data. *Bioinformatics*, 22(9):1144–1146, May 2006.

[63] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc IEEE*, 77(2):257–286, February 1989.

[64] Yuval Benjamini and Terence P. Speed. Estimation and correction for GC-content bias in high throughput sequencing. Technical report, University of California at Berkeley, June 2011.

[65] Valentina Boeva, Andrei Zinovyev, Kevin Bleakley, Jean-Philippe Vert, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, 27(2):268–269, January 2011.

[66] Christopher A. Miller, Oliver Hampton, Cristian Coarfa, and Aleksandar Milosavljevic. ReadDepth: A parallel r package for detecting copy number alterations from short sequencing reads. *PLoS ONE*, 6(1):e16327+, January 2011.

[67] Derek Y. Chiang, Gad Getz, David B. Jaffe, Michael J. T. O'Kelly, Xiaojun Zhao, Scott L. Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S. Lander. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods*, 6(1):99–103, November 2008.

[68] Chao Xie and Martti Tammi. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, 10(1):80+, 2009.

[69] Sergii Ivakhno, Tom Royce, Anthony J. Cox, Dirk J. Evers, R. Keira Cheetham, and Simon Tavaré. CNAsega novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics*, 26(24): 3051–3058, December 2010.

[70] Jeremy J. Shen and Nancy R. Zhang. Change-Point model on Non-Homogeneous poisson processes with application in copy number profiling by Next-Generation DNA sequencing. Technical report, Division of Biostatistics, Stanford University, March 2011.

[71] Jarupon F. Sathirapongsasuti, Hane Lee, Basil A. J. Horst, Georg Brunner, Alistair J. Cochran, Scott Binder, John Quackenbush, and Stanley F. Nelson. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, 27(19):2648–2654, October 2011.

[72] Olivier Harismendy, Pauline C. Ng, Robert L. Strausberg, Xiaoyun Wang, Timothy B. Stockwell, Karen Y. Beeson, Nicholas J. Schork, Sarah S. Murray, Eric J. Topol, Samuel Levy, and Kelly A. Frazer. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol*, 10(3):R32+, 2009.

[73] Dale J. Hedges, Toumy Guettouche, Shan Yang, Guney Bademci, Ashley Diaz, Ashley Andersen, William F. Hulme, Sara Linker, Arpit Mehta, Yvonne J. K. Edwards, Gary W. Beecham, Eden R. Martin, Margaret A. Pericak-Vance, Stephan Zuchner, Jeffery M. Vance, and John R. Gilbert. Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. *PLoS ONE*, 6(4):e18595+, April 2011.

[74] Daniel S. Herman, G. Kees Hovingh, Oleg Iartchouk, Heidi L. Rehm, Raju Kucherlapati, J. G. Seidman, and Christine E. Seidman. Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat Methods*, 6(7): 507–510, July 2009.

[75] Alex S. Nord, Ming Lee, Mary-Claire C. King, and Tom Walsh. Accurate and exact CNV identification from targeted high-throughput sequence data. *BMC Genomics*, 12(1):184+, 2011.

[76] Yingrui Li, Nicolas Vinckenbosch, Geng Tian, Emilia Huerta-Sanchez, Tao Jiang, Hui Jiang, Anders Albrechtsen, Gitte Andersen, Hongzhi Cao, Thorfinn Korneliussen, Niels Grarup, Yiran Guo, Ines Hellman, Xin Jin, Qibin Li, Jiangtao Liu, Xiao Liu, Thomas Sparso, Meifang Tang, Honglong Wu, Renhua Wu, Chang Yu, Hancheng Zheng, Arne Astrup, Lars Bolund, Johan Holmkvist, Torben Jorgensen, Karsten Kristiansen, Ole Schmitz, Thue W. Schwartz, Xiuqing Zhang, Ruiqiang Li, Huanming Yang, Jian Wang, Torben Hansen, Oluf Pedersen, Rasmus Nielsen, and Jun Wang. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet*, 42(11):969–972, November 2010.

[77] Kim D. Pruitt, Jennifer Harrow, Rachel A. Harte, Craig Wallin, Mark Diekhans, Donna R. Maglott, Steve Searle, Catherine M. Farrell, Jane E. Loveland, Barbara J. Ruef, Elizabeth Hart, Marie-Marthe M. Suner, Melissa J. Landrum, Bronwen Aken, Sarah Ayling, Robert Baertsch, Julio Fernandez-Banet, Joshua L. Cherry, Val Curwen, Michael Dicuccio, Manolis Kellis, Jennifer Lee, Michael F. Lin, Michael Schuster, Andrew Shkeda, Clara Amid, Garth Brown, Oksana Dukhanina, Adam Frankish, Jennifer Hart, Bonnie L. Maidak, Jonathan Mudge, Michael R. Murphy, Terence Murphy, Jeena Rajan, Bhanu Rajput, Lillian D. Riddick, Catherine Snow, Charles Steward, David Webb, Janet A. Weber, Laurens Wilming, Wenyu Wu, Ewan Birney, David Haussler, Tim Hubbard, James Ostell, Richard Durbin, and David Lipman. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res*, 19(7):1316–1323, July 2009.

[78] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria., 2011.

[79] David Weese, Anne-Katrin Emde, Tobias Rausch, Andreas Döring, and Knut Reinert. RazerSfast read mapping with sensitivity control. *Genome Res*, 19(9):1646–1654, September 2009.

[80] J. Zhang, L. Feuk, G. E. Duggan, R. Khaja, and S. W. Scherer. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet Genome Res*, 115(3-4):205–214, 2006.

[81] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010.

[82] Brian J. O'Roak, Pelagia Deriziotis, Choli Lee, Laura Vives, Jerrod J. Schwartz, Santhosh Girirajan, Emre Karakoc, Alexandra P. MacKenzie, Sarah B. Ng, Carl Baker, Mark J. Rieder, Deborah A. Nickerson, Raphael Bernier, Simon E. Fisher, Jay Shendure, and Evan E. Eichler. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet*, 43(6):585–589, June 2011.

[83] Mark D. Robinson and Gordon K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, November 2007.

[84] Thomas Hardcastle and Krystyna Kelly. baySeq: Empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11 (1):422+, 2010.

[85] Mark A. Van De Wiel, Gwenaël G. R. Leday, Luba Pardo, Håvard Rue, Aad W. Van Der Vaart, and Wessel N. Van Wieringen. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 14(1):113–128, January 2013.

[86] Kasper D. Hansen, Rafael A. Irizarry, and W. U. Zhijin. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, 13 (2):204–216, April 2012.

[87] Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. GC-content normalization for RNA-seq data. *BMC Bioinformatics*, 12(1):480+, December 2011.

[88] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol*, 8(1):37–52, 2001.

[89] Florian Hahne, Wolfgang Huber, Robert Gentleman, and Seth Falcon. *Bioconductor Case Studies*. Springer New York, New York, NY, 2008. ISBN 978-0-387-77239-4.

[90] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, corrected edition, August 2003. ISBN 0387952845.

[91] Jianxing Feng, Clifford A. Meyer, Qian Wang, Jun S. Liu, X. Shirley Liu, and Yong Zhang. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics*, 28(21):2782–2788, November 2012.

[92] Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans Am Math Soc*, 54(3):426–482, November 1943.

[93] Hao Wu, Chi Wang, and Zhijin Wu. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, 14(2): 232–243, April 2013.

[94] D. R. Cox and N. Reid. Parameter orthogonality and approximate conditional inference. *J R Stat Soc*, 49(1):1–39, 1987.

[95] Davis J. McCarthy, Yunshun Chen, and Gordon K. Smyth. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res*, 40(10):4288–4297, May 2012.

[96] Peter McCullagh and J. A. Nelder. *Generalized Linear Models*. Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, London, United Kingdom, second edition edition, August 1989. ISBN 0412317605.

[97] Milton Abramowitz and Irene Stegun. *Handbook of Mathematical Functions*. Dover books on mathematics. Dover Publications, 1 edition, June 1972. ISBN 0486612724.

[98] Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pac J Math*, 16(1):1–3, January 1966.

[99] Mee Y. Park. *Generalized Linear Models with Regularization*. PhD thesis, Stanford University, Department of Statistics Sequoia Hall 390 Serra Mall Stanford University Stanford, CA, September 2006.

[100] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 33(1):1–22, 2010.

[101] Erika Cule, Paolo Vineis, and Maria De Iorio. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12(1):372+, 2011.

[102] R. Dennis Cook and Sanford Weisberg. *Residuals and Influence in Regression (Monographs on Statistics and Applied Probability, 18)*. Springer, 1st edition, October 1982. ISBN 041224280X.

[103] Felix Haglund, Ran Ma, Mikael Huss, Luqman Sulaiman, Ming Lu, Inga-Lena Nilsson, Anders Höög, Christofer C. Juhlin, Johan Hartman, and Catharina Larsson. Evidence of a functional estrogen receptor in parathyroid adenomas. *J Clin Endocrinol Metab*, September 2012.

[104] Alyssa Frazee, Ben Langmead, and Jeffrey Leek. ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12 (1):449+, 2011.

[105] Daniel Bottomly, Nicole A. Walter, Jessica Ezzell E. Hunter, Priscila Darakjian, Sunita Kawane, Kari J. Buck, Robert P. Searles, Michael Mooney, Shannon K. McWeeney, and Robert Hitzemann. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-seq and microarrays. *PLoS ONE*, 6(3):e17820+, March 2011.

[106] Paul Hammer, Michaela S. Banck, Ronny Amberg, Cheng Wang, Gabriele Petznick, Shujun Luo, Irina Khrebtukova, Gary P. Schroth, Peter Beyerlein, and Andreas S. Beutler. mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. *Genome Res*, 20(6):847–860, June 2010.

[107] Brenton R. Graveley, Angela N. Brooks, Joseph W. Carlson, Michael O. Duff, Jane M. Landolin, Li Yang, Carlo G. Artieri, Marijke J. van Baren, Nathan Boley, Benjamin W. Booth, James B. Brown, Lucy Cherbas, Carrie A. Davis, Alex Dobin, Renhua Li, Wei Lin, John H. Malone, Nicolas R. Mattiuzzo, David Miller, David Sturgill, Brian B. Tuch, Chris Zaleski, Dayu Zhang, Marco Blanchette, Sandrine Dudoit, Brian Eads, Richard E. Green, Ann Hammonds, Lichun Jiang, Phil Kapranov, Laura Langton, Norbert Perrimon, Jeremy E. Sandler, Kenneth H. Wan, Aarron Willingham, Yu Zhang, Yi Zou, Justen Andrews, Peter J. Bickel, Steven E. Brenner, Michael R. Brent, Peter Cherbas, Thomas R. Gingeras, Roger A. Hoskins, Thomas C. Kaufman, Brian Oliver, and Susan E. Celniker. The developmental transcriptome of drosophila melanogaster. *Nature*, 471(7339):473–479, March 2011.

[108] Angela N. Brooks, Li Yang, Michael O. Duff, Kasper D. Hansen, Jung W. Park, Sandrine Dudoit, Steven E. Brenner, and Brenton R. Graveley. Conservation of an RNA regulatory map between drosophila and mammals. *Genome Res*, 21(2): 193–202, February 2011.

[109] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221): 470–476, November 2008.

[110] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc*, 57(1):289–300, 1995.

[111] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.

[112] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(Database issue):D109–D114, January 2012.

[113] David Croft. *Building Models Using Reactome Pathways as Templates*, volume 1021 of *Methods in Molecular Biology*, chapter 14, pages 273–283. Humana Press, Totowa, NJ, 2013. ISBN 978-1-62703-449-4.

[114] Andrew E. Jaffe, John D. Storey, Hongkai Ji, and Jeffrey T. Leek. Gene set bagging for estimating replicability of gene set analyses, January 2013.

[115] Davis J. McCarthy and Gordon K. Smyth. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, 25(6):765–771, March 2009.

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

[117] Sam John, Peter J. Sabo, Robert E. Thurman, Myong-Hee Sung, Simon C. Biddie, Thomas A. Johnson, Gordon L. Hager, and John A. Stamatoyannopoulos. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet*, 43(3):264–268, March 2011.

[118] Roger Pique-Regi, Jacob F. Degner, Athma A. Pai, Daniel J. Gaffney, Yoav Gilad, and Jonathan K. Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res*, 21(3):447–455, March 2011.

[119] Hongkai Ji and X. Shirley Liu. Analyzing 'omics data using hierarchical models. *Nat Biotechnol*, 28(4):337–340, April 2010.

[120] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25–10, March 2009.

[121] Yong Zhang, Tao Liu, Clifford Meyer, Jerome Eeckhoute, David Johnson, Bradley Bernstein, Chad Nusbaum, Richard Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based analysis of ChIP-seq (MACS). *Genome Biol*, 9(9):R137+, 2008.

[122] Bradley E. Bernstein, John A. Stamatoyannopoulos, Joseph F. Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A. Marra, Arthur L. Beaudet, Joseph R. Ecker, Peggy J. Farnham, Martin Hirst, Eric S. Lander, Tarjei S. Mikkelsen, and James A. Thomson. The NIH roadmap epigenomics mapping consortium. *Nat Biotechnol*, 28(10):1045–1048, October 2010.

[123] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W. Wasserman, and Boris Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–D94, January 2004.

[124] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis*. Chapman & Hall/CRC, 2004. ISBN 158488388.

[125] Matthew D. Hoffman and Andrew Gelman. The No-U-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. Technical report, Princeton University, January 2012.

[126] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Stat Science*, 7(4):457–472, 1992.

[127] Lars Grontved, Sam John, Songjoon Baek, Ying Liu, John R. Buckley, Charles Vinson, Greti Aguilera, and Gordon L. Hager. C/EBP maintains chromatin accessibility in liver and facilitates glucocorticoid receptor recruitment to steroid response elements. *EMBO J*, 32(11):1568–1583, May 2013.

[128] Simon C. Biddie, Sam John, Pete J. Sabo, Robert E. Thurman, Thomas A. Johnson, R. Louis Schiltz, Tina B. Miranda, Myong-Hee H. Sung, Saskia Trump, Stafford L. Lightman, Charles Vinson, John A. Stamatoyannopoulos, and Gordon L. Hager.

Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol Cell*, 43(1):145–155, July 2011.

[129] John A. Rice. *Mathematical statistics and data analysis*. Thomson/Brooks/Cole, 2007. ISBN 9780534399429.

[130] Menno P. Creyghton, Albert W. Cheng, Grant G. Welstead, Tristan Kooistra, Bryce W. Carey, Eveline J. Steine, Jacob Hanna, Michael A. Lodato, Garrett M. Frampton, Phillip A. Sharp, Laurie A. Boyer, Richard A. Young, and Rudolf Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*, 107(50):21931–21936, December 2010.

[131] Matthew G. Guenther, Stuart S. Levine, Laurie A. Boyer, Rudolf Jaenisch, and Richard A. Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88, July 2007.

[132] Michael J. Carrozza, Bing Li, Laurence Florens, Tamaki Suganuma, Selene K. Swanson, Kenneth K. Lee, Wei-Jong Shia, Scott Anderson, John Yates, Michael P. Washburn, and Jerry L. Workman. Histone h3 methylation by set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell*, 123 (4):581–592, November 2005.

[133] Sam John, Peter J. Sabo, Thomas A. Johnson, Myong-Hee Sung, Simon C. Biddie, Stafford L. Lightman, Ty C. Voss, Sean R. Davis, Paul S. Meltzer, John A. Stamatoyannopoulos, and Gordon L. Hager. Interaction of the glucocorticoid receptor with the chromatin landscape. *Mol Cell*, 29(5):611–624, March 2008.

[134] Tarjei S. Mikkelsen, Manching Ku, David B. Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-Kyung Kim, Richard P. Koche, William Lee, Eric Mendenhall, Aisling O/'Donovan, Aviva Presser, Carsten Russ, Xiaohui Xie, Alexander Meissner, Marius Wernig, Rudolf Jaenisch, Chad Nusbaum, Eric S. Lander, and Bradley E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, August 2007.

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

[136] Krishanpal Karmodiya, Arnaud R. Krebs, Mustapha Oulad-Abdelghani, Hiroshi Kimura, and Laszlo Tora. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics*, 13(1):424+, August 2012.

[137] Robert C. Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki,

Colin Smith, Gordon Smyth, Luke Tierney, Jean Y. Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80+, 2004.

List of Figures

List of Tables

Appendix A

Supplementary Figures

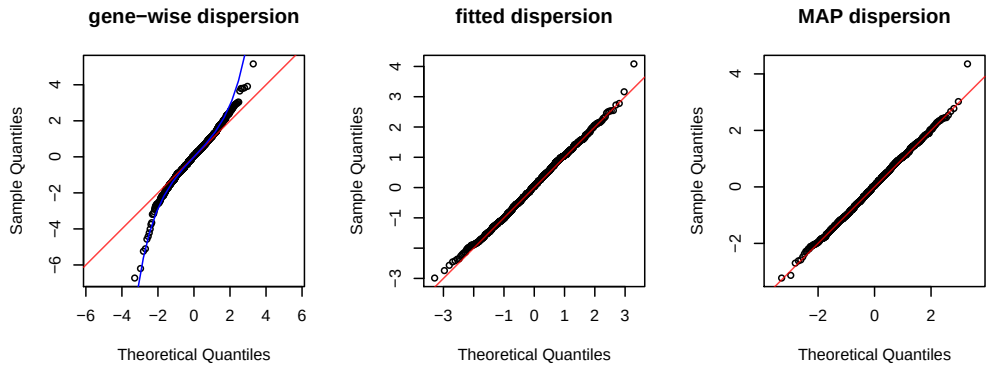

Figure A.1: Quantile-quantile plot for Wald statistics from simulated null data compared to $\mathcal{N}(0,1)$ theoretical quantiles. Wald statistics are shown which use various dispersion estimates. Using the gene-wise dispersion estimates results in Wald-statistics which closely follow a t-distribution with $(m-p)$ degrees of freedom (blue curve). Using either the fitted or MAP dispersion estimates results in Wald statistics which more closely follow the expected quantiles of the normal distribution (red line). However, when the dispersion prior variance is large, the distribution of Wald statistics using MAP dispersion estimates for null data has wider tails than the normal distribution.

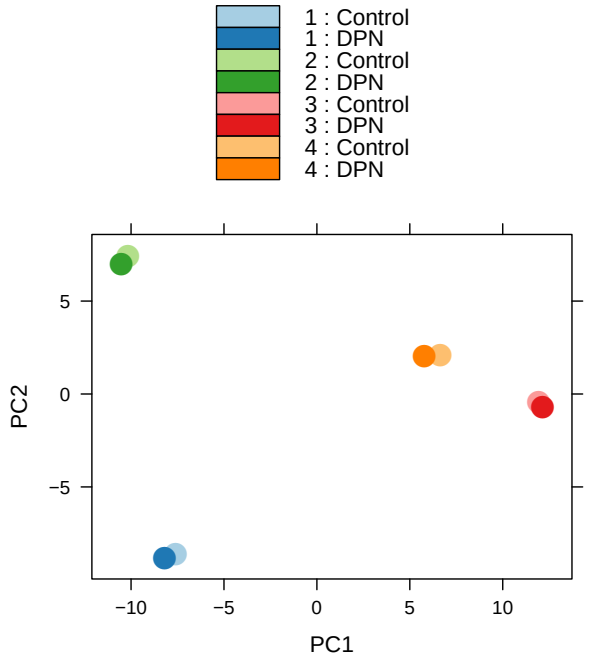

Figure A.2: Principal component plot of 8 samples from the experiment of Haglund et al. [103], at 48 hours from control and DPN treatment. The raw counts are first rlog transformed using a dispersion estimate blind to the information about the patient or treatment type. The patient type, denoted by color, explains a larger part of the total variance than the treatment type.

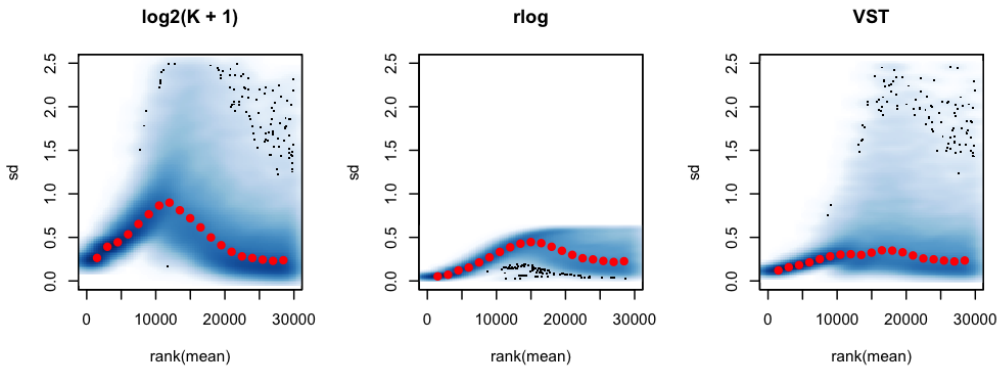

Figure A.3: Standard deviation dependence on the mean using the log transformation, the rlog transformation and using a variance stabilizing transformation (VST) as described in Anders and Huber [22]. For each gene, the standard deviation across samples is plotted against the gene rank by normalized mean count. The rlog transformation helps to stabilize the variance, though in this case the VST appears to stabilize the variance better, with the mean standard deviation (red points) mostly flat across the rank of mean count.

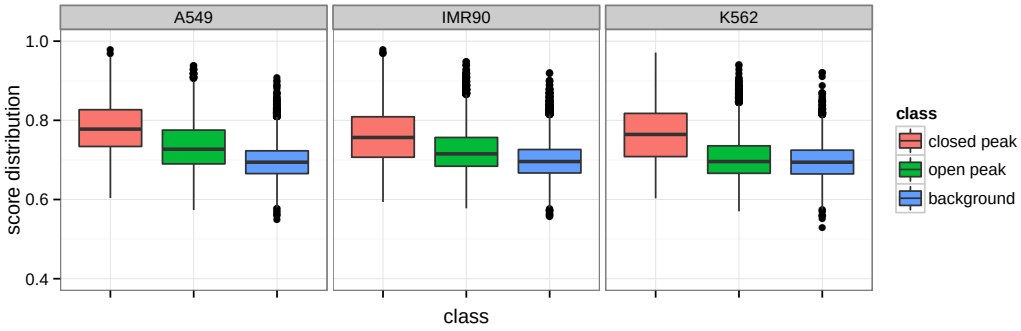

Figure A.4: Distribution of top scoring motifs for various regions: GR peaks not overlapping any DHS (closed peaks), GR peaks overlapping DHS, (open peaks), and regions which are 2 kb randomly upstream or downstream from a peak (background). For all cell types, the motif scores are elevated for the closed peaks over background. For A549 and IMR90 cells, the open peaks also have elevated score distributions over background.

Appendix B

Supplementary Tables

μ_0	NB GLM	log t-test	sqrt t-test
16	0.09	0.09	0.09
64	0.11	0.11	0.11
256	0.09	0.11	0.11
1024	0.10	0.12	0.12

Table B.1: Type I error control for discrete distributions and t-tests on transformed counts. Shown is the proportion of tests with p-value < 0.1, for Poisson data in two groups with three samples each, with no true difference, over 1000 replications.

m	p	α	theor. var.	sample var.
4	2	0.05	1.645	1.909
4	2	0.20	1.645	1.781
8	2	0.05	0.395	0.402
8	2	0.20	0.395	0.394
8	3	0.05	0.490	0.550
8	3	0.20	0.490	0.466
16	2	0.05	0.154	0.159
16	2	0.20	0.154	0.143
16	3	0.05	0.166	0.168
16	3	0.20	0.166	0.161

Table B.2: Theoretical and sample variance of log dispersion estimates for various combinations of sample size m, number of parameters p and true dispersion α. The estimates are the *DESeq2* gene-wise estimates from 4000 simulated genes with negative binomial counts with a mean of 1024. The sample variance of the log dispersion estimates is generally close to the approximate theoretical variance derived in Chapter 4.

	condition 1	condition 2
bottomly	C57BL/6J	DBA/2J
hammer	control	L5 SNL
modencodefly	larvae	adult
pasilla	untreated	treated
wang	not cerebellum	cerebellum

Table B.3: Condition of interest used for testing *DESeq2* against other software packages.

	genes	samples
bottomly	13932	21
hammer	18635	8
modencodefly	13224	12
pasilla	11836	7
wang	12596	14
simulated	4000	12

Table B.4: Dimensions of RNA-Seq datasets used for testing *DESeq2* against other software packages. The number of genes counts only those genes with non-zero sum of counts across all samples.

	DESeq	DESeq2	edgeR	DSS	baySeq
bottomly	67	35	31	12	5794
hammer	68	21	19	13	4464
modencodefly	55	14	12	8	6947
pasilla	47	13	12	9	3383
wang	63	17	13	8	4735
s0	13	4	4	3	740
s1	13	4	4	3	739

Table B.5: Timing of differential expression packages on full datasets in seconds.

Min	1st Q	Median	Mean	3rd Q	Max
0.9996	0.9998	1.0000	1.0000	1.0000	1.0040

Table B.6: Convergence diagnostic: `R-hat` values for all parameters of the hierarchical model [126]. `R-hat` is defined as the square root of the variance of the mixture of chains divided by the average within-chain variance. When the MCMC samples of the model parameters have converged on the posterior distribution, then the variance of the mixture of the chains should be equal to the variance of the individual chains, so `R-hat` should be equal to 1. Values greater than 1 indicate that the sampler might not have yet converged on the posterior distribution.

model:	Hier. Bayes	GLM	Hier. Bayes	GLM
statistic:	posterior mean	MLE	posterior SD	MLE SE
Intercept	-0.836	-0.838	0.012	0.012
DNase	0.212	0.212	0.008	0.009
Input	0.093	0.092	0.006	0.007
H2AZ	-0.457	-0.455	0.015	0.015
H3K27ac	1.253	1.261	0.026	0.026
H3K27me3	0.123	0.124	0.013	0.013
H3K36me3	-0.169	-0.168	0.015	0.015
H3K4me1	0.359	0.352	0.018	0.018
H3K4me2	1.213	1.234	0.044	0.044
H3K4me3	-0.367	-0.382	0.044	0.044
H3K79me2	-0.233	-0.231	0.015	0.015
H3K9ac	-1.078	-1.092	0.033	0.033
H3K9me3	0.132	0.132	0.011	0.011
H4K20me1	-0.272	-0.273	0.010	0.010
motif	0.372	0.372	0.006	0.007

Table B.7: Hierarchical model posterior mean and standard deviation (SD) compared to the maximum likelihood estimate (MLE) and its standard error (MLE SE). The experiment level coefficients β from the hierarchical model for the A549 cell, TRG are compared to maximum likelihood estimates and their standard error using a Poisson generalized linear model. Due to the large sample size, the coefficients and standard deviations / standard errors are very similar for the Bayesian analysis and the "frequentist" analysis (i.e. the MLE estimates and standard errors).

Appendix C

Software

Bioconductor packages

I have implemented two of the methods described in this thesis as R/Bioconductor packages [78, 137]. Both packages include documentation of every function, as well as "Sweave vignettes", detailed workflows with sample code which is tested daily on actual datasets.

- The method described in Chapter 3 for exome-enriched DNA-Seq and copy number variants is implemented in the R/Bioconductor package *exomeCopy*, available since October 2011 at:
  http://www.bioconductor.org/packages/release/bioc/html/exomeCopy.html.
- The method described in Chapter 4 for RNA-Seq and differential gene expression is implemented in the R/Bioconductor package *DESeq2*, available since March 2013 at:
  http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html.

Hierarchical model code

Example `stan` code for the hierarchical model presented in Chapter 5 follows, for only a single experiment and a single cell type.

```
data {
  int N1;
  int M;
  int K;
  int y11[N1];
  matrix[N1,M] x11;
}
parameters {
  vector[M] beta11;
  vector[M] nu1;
  vector[M] lambda;
  real beta011;
  real<lower=0> sigma0;
  real<lower=0> sigma_beta;
  real<lower=0> sigma_nu1;
  real<lower=0> sigma_lambda;
```

```
}
model {
  y11 ~ poisson_log(beta011 + x11 * beta11);
  beta011 ~ normal(0, sigma0);
  beta11 ~ normal(nu1, sigma_beta);
  nu1 ~ normal(lambda, sigma_nu1);
  lambda ~ normal(0, sigma_lambda);
  sigma0 ~ gamma(10,10);
  sigma_beta ~ gamma(10,10);
  sigma_nu1 ~ gamma(10,10);
  sigma_lambda ~ gamma(10,10);
}
```

Appendix D

Notation

D.1 Acronyms

arrayCGH	array-based comparative genomic hybridization
bp	base pair
cDNA	complementary DNA
ChIP	chromatin immunoprecipitation
CNV	copy number variant
mRNA	messenger RNA
DESeq	differential expression for sequence counts
DHS	DNase hypersensitive site
FDR	false discovery rate
GLM	generalized linear model
GR	glucocorticoid receptor
HMM	hidden Markov model
kb	kilobase (10^3 base pairs)
LFC	log fold change
MAP	maximum *a posteriori*
Mb	megabase (10^6 base pairs)
MLE	maximum likelihood estimate
PCA	principal component analysis
PCR	polymerase chain reaction
rlog	regularized log transformation
SD	standard deviation
SE	standard error
TSS	transcription start site
VSN	variance stabilizing normalization
VST	variance stabilizing transformation
XLID	X-linked intellectual disability

D.2 Symbols

K	read counts
X	design matrix
x_{i*}	the i-th row of X
$\vec{\beta}$	column vector of coefficients
μ	mean parameter
α	dispersion parameter
σ^2	variance parameter
$\mathcal{N}$	normal distribution
NB	negative binomial distribution

Appendix E

Curriculum Vitae

For reasons of data protection, the Curriculum Vitae is not published in the online version.

For reasons of data protection, the Curriculum Vitae is not published in the online version.

For reasons of data protection, the Curriculum Vitae is not published in the online version.

Appendix F

Zusammenfassung

Mit Hochdurchsatz-Sequenzierverfahren (HTS) bezeichnet man das gleichzeitige Sequenzieren von Millionen von DNA-Fragmenten, welche entweder zur Genomrekonstrution genutzt oder auf ein bestehendes Referenzgenom aligniert werden können. Das Protokoll kann erweitert werden, um verschiedene biologische Zustände der Zelle, wie z.B. die Anzahl an DNA-Kopien, mRNA-Abundanzen oder verschiedene Chromatin-Eigenschaften, zu messen. Diese Hochdurchsatzverfahren ermöglichen biologische Zustände genomweit mit einem einzigen Experiment zu quantifizieren. Obwohl diese Experimente oft nur eine begrenzte Stichprobengrösse haben, liefern sie dennoch Informationen zu tausenden Genomregionen und ermöglichen das Erstellen robuster statistische Modelle, um technische Fehler zu reduzieren.

In dieser Arbeit entwickle ich drei statistische Modelle basierend auf HTS-Daten um konkrete biologische Fragen zu beantworten.

Im ersten Teil wird ein *hidden* Markov-Modell entworfen, um Kopienzahlvariationen (CNVs) in einzelnen Patienten zu detektieren. Das Modell berücksichtigt hierbei technische Artefakte wie z.B. die variable HTS-Effizienz abhängig vom lokalen GC-Gehalt. Angewendet auf eine Studie mit 248 männlichen Patienten, sagt das Modell 16 grosse CNVs voraus, wovon 10 CNVs getestet und experimentell validiert wurden. Im Vergleich mit anderen Segmentierungsalgorithmen zeigt die vorgestellte Software auf simulierten CNVs eine höhere Sensitivität bei gleicher Anzahl prognostizierter CNVs.

Im zweiten Teil wird die Parameterabschätzung in einem statistisches Modell zur Identifizierung von differentieller Genexpression in RNA-Seq-Daten verbessert. Dies umfasst die Benutzung von empirischen Bayes'schen *a-priori*-Wahrscheinlichkeiten, welche über alle Gene geschätzt werden. Hierdurch werden unsichere Schätzungen der Varianz-Parameter und der Expressionsänderung einzelner Gene korrigiert. Das verbesserte Modell ist sensitiver und zusätzlich robuster in der Schätzung der Expressionsänderung im Vergleich zu alternativen Softwarepaketen.

Im letzten Teil wird ein hierarchisches Bayes'sches Modell verwendet um in zugänglichen Chromatinregionen den Zusammenhang zwischen der Bindung eines Transkriptionsfaktors und Chromatin- und Sequenz-Eigenschaften zu beschreiben. Dieses Modell umfasst drei Ebenen: den Vergleich einzelner Experimente, Experimente des gleichen Zelltyps oder Experimente über alle Zelltypen. Das Modell dient der Hypothesengenerierung für das DNA-Bindungsverhalten eines Transkriptionskriptionsfaktors. Dies wird am Beispiel des Glucocorticoid-Rezeptors veranschaulicht.

Zusammenfassend beschreibt diese Arbeit eine Sammlung statistischer Methoden für die Modellierung von HTS-Daten, die in verschiedenen biologischen Bereichen verwendet werden kann. Diese Methoden bilden einen allgemeinen Rahmen zur robusten Schätzungen von Variablen und zum Testen von Hypothesen.

Appendix G

Summary

High-throughput sequencing (HTS) refers to the simultaneous sequencing of millions of fragments of DNA, which can be either assembled to reconstitute a genome, or aligned to an existing reference genome. The protocol can be extended to assay a wide variety of biological states of the cell, including DNA copy number, mRNA abundance and various properties of chromatin. HTS experiments allow for these biological states to be quantified as read counts at genome-wide scale with a single experiment. Though the experiments are expensive and often datasets are produced with limited sample size, information can be shared across thousands of genomic ranges in order to obtain robust models which control for technical biases.

In this thesis, I present three statistical models for analyzing HTS read count data, aimed at answering concise biological questions.

First, a hidden Markov model is developed for detecting copy number variants (CNVs) in individual samples while controlling for technical artifacts, such as variation in read counts due to local GC-content. Applied to a study of 248 male patients with X-linked intellectual disability, the model predicts 16 large CNVs, of which 10 candidate disease-causing CNVs were tested and all experimentally validated. The proposed software is then compared with state-of-the-art segmentation algorithms on normalized data, showing higher sensitivity while controlling the total rate of predicted CNVs.

Second, improvements for parameter estimation are made for a statistical model of differential gene expression from RNA-Seq data. The improvements involve the use of empirical Bayes priors – priors estimated using the observations from all genes – in order to moderate otherwise noisy estimates of dispersion and fold changes for individual genes. The improved model shows increased sensitivity and more robust estimation of fold change in comparison with other differential expression software packages for RNA-Seq.

Finally, a hierarchical Bayes model is used to associate transcription factor binding with chromatin and sequence features in regions of accessible chromatin. The hierarchical model incorporates three levels of parameters: one for individual experiments, one for experiments of the same cell type and one across all cell types. The model parameters are used to generate hypotheses regarding the DNA-binding behavior of a transcription factor, the glucocorticoid receptor.

In summary, this thesis describes a set of statistical methods for HTS read count data which can be used across various biological domains. The methods form a framework for robust estimation of variables and hypothesis testing.

Appendix H

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, Juli 2013