AN ABSTRACT OF THE DISSERTATION OF

Gu Mi for the degree of Doctor of Philosophy in Statistics presented on June 10, 2014.

Title: Statistical Analysis of RNA Sequencing Count Data

Abstract approved: _____

Yanming Di Daniel W. Schafer

RNA-Sequencing (RNA-Seq) has rapidly become the *de facto* technique in transcriptome studies. However, established statistical methods for analyzing experimental and observational microarray studies need to be revised or completely re-invented to accommodate RNA-Seq data's unique characteristics. In this dissertation, we focus on statistical analyses performed at two particular stages in the RNA-Seq pipeline, namely, regression analysis of gene expression levels including tests for differential expression (DE) and the downstream Gene Ontology (GO) enrichment analysis.

The negative binomial (NB) distribution has been widely adopted to model RNA-Seq read counts for its flexibility in accounting for any extra-Poisson variability. Because of the relatively small number of samples in a typical RNA-Seq experiment, power-saving strategies include assuming some commonalities of the NB dispersion parameters across genes, via simple models relating them to mean expression rates. Many such NB dispersion models have been proposed, but there is limited research on evaluating model adequacy. We propose a simulation-based goodness-of-fit (GOF) test with diagnostic graphics to assess the NB assumption for a single gene via parametric bootstrap and empirical probability

plots, and assess the adequacy of NB dispersion models by combining individual GOF test p-values from all genes. Our simulation studies and real data analyses suggest the NB assumption is valid for modeling a gene's read counts, and provide evidence on how NB dispersion models differ in capturing the variation in the dispersion.

It is not well understood to what degree a dispersion-modeling approach can still be useful when a fitted dispersion model captures a significant part, but not all, of the variation in the dispersion. As a further step towards understanding the power-robustness trade-offs of NB dispersion models, we propose a simple statistic to quantify the inadequacy of a fitted NB dispersion model. Subsequent power-robustness analyses are guided by this estimated residual dispersion variation and other controlling factors estimated from real RNA-Seq datasets. The proposed measure for quantifying residual dispersion variation gives hints on whether we can gain statistical power by a dispersion-modeling approach. Our real-data-based simulations also provide benchmarking investigations into the power and robustness properties of the many NB dispersion methods in current RNA-Seq community.

For statistical tests of enriched GO categories, which aim to relate the outcome of DE analysis to biological functions, the transcript length becomes a confounding factor as it correlates with both the GO membership and the significance of the DE test. We propose to adjust for such bias using the logistic regression and incorporate the length as a covariate. The use of continuous measures of differential expression via transformations of DE test p-values also avoids the subjective specification of a p-value threshold adopted by contingency-table-based approaches. Simulation and real data examples indicate that enriched categories no longer favor longer transcripts after the adjustment, which justifies the effectiveness of our proposed method.

Statistical Analysis of RNA Sequencing Count Data

by

Gu Mi

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented June 10, 2014
Commencement June 2015

Doctor of Philosophy dissertation of Gu Mi presented on June 10, 2014.

APPROVED:

Co-Major Professor, representing Statistics

Co-Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Gu Mi, Author

ACKNOWLEDGEMENTS

Academic

Special thanks to my co-major professors, Dr. Yanming Di and Dr. Daniel W. Schafer, for their immense and unconditional guidance, support and encouragement in the past five years. Whenever I had difficulties in my research, they are always available to help with inspirational ideas. I am honored to be Yanming's research assistant and a member in the interdisciplinary group, which grants me the opportunity to collaborate with experts in finding solutions to many exciting problems. With a genuine passion for knowledge and research, Yanming has set a high standard for me to pursue in my future life. I am indebted to Dan for his sharp insights in defining research problems and careful manuscript revisions. Thank you Dan for teaching me how should a rigorous statistician contribute to scientific advances.

Thank you to Dr. Sarah Emerson and Dr. Yuan Jiang, not only for serving on my committee and teaching wonderful classes, but also for the many inspiring discussions on my research. I am also grateful to Dr. Jeffrey H. Chang for his warmhearted supports on manuscript revisions and many thought-provoking conversations throughout my entire Ph.D. study. Thank you Jeff for motivating my research on many interesting RNA-Seq problems.

I feel fortunate to be a student of Professors Virginia Lesser, Robert Smythe, Fred Ramsey, Alix Gitelman, Lisa Madsen, Paul Murtaugh, Cliff Pereira and Charlotte Wickham. Thank you all for showing me every fascinating aspect of statistics. I appreciate Dr. David Birkes and Dr. Lan Xue for their advanced-level classes—I can still remember the self-fulfillment after solving a difficult question, sometimes during Dr. Birkes' midnight office hours in Kidder Hall.

Thank you to my collaborators at the Center for Genome Research and Biocomputing (CGRB) and members of the Bioinformatics Users Group, who answered my (sometimes naive) biological questions with patience. Special thanks to Dr. Shawn O'Neil who brought me to the fantastic world of computational biology.

Thank you to Dr. Jin Zheng and Dr. Mark Farmen for supervising my summer internship at Eli Lilly and Company, and to many colleagues (especially Dr. Wei Shen and Dr. Yun-Fei Chen) in the Global Statistical Sciences (GSS) who helped me get familiarized with many statistical applications in the pharmaceutical industry. It has been a pleasure to work with you in the summer of 2013 in Indianapolis.

Personal

Thank you mom and dad. You are the ones who love me the most and sacrifice the most while I journey thousands of miles away from home. Thank you Luna for accompanying me these years in the US, China and Canada, and even more wonderful places in the future. Thank you to all my dear friends, especially Xuan, Shuping, Ziyi, Meian and Lili for the joy we had in the beautiful Cascadia, and my long-time friend Jinhang ("Blake") for studying together at SUNY-Cortland back in 2008, my short visits to Chicago and College Park later during my graduate studies. It is my great honor to have you all in my life, wherever I go.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

Figure | Page

LIST OF FIGURES (Continued)

Figure Page

LIST OF TABLES

LIST OF TABLES (Continued)

Table | Page

LIST OF APPENDIX FIGURES

Figure Page

Statistical Analysis of RNA Sequencing Count Data

1 Introduction

In the last decade, with the advent of high-throughput next-generation sequencing (NGS) technologies which generate unprecedented volumes of high-quality gene expression data, life scientists have the opportunity to obtain a progressively fuller knowledge of the quantitative and qualitative aspects of the transcriptome (Ozsolak and Milos [89]). The use of the NGS technologies for transcriptome studies at the nucleotide level is known as RNA sequencing (RNA-Seq), and it has revolutionized the manner in which eukaryotic transcriptomes are analyzed (Wang et al. [121]). Compared to hybridization-based microarray technologies introduced in the mid-1990s on which earlier genomics studies primarily relied, RNA-Seq has many clear advantages on the scope and complexity of biological questions it can answer, and the declining costs have led to explosive growth in its use for transcriptome studies. At the same time, RNA-Seq experiments have precipitated the flood of massive and complex datasets and the interpretations are not straightforward.

Understanding this wealth of RNA-Seq data requires sophisticated statistical methods and dedicated computational software that take into accounts some prominent features of this sequencing-based method (e.g., discrete, digital sequencing read counts with typically small sample sizes) and the sources of variations in the pipeline (e.g., read alignments and summarizations), among other factors that fundamentally differ from those in the microarray era. Statisticians and bioinformaticians have been developing novel and improved approaches for solving problems that arise at various stages of the RNA-Seq pipeline (see Section 1.1.3). In this dissertation, we restrict our attention to some *statistical* problems in two areas, namely regression analysis of gene expression levels including tests for differ-

ential expression (DE) and the downstream enrichment analysis using the Gene Ontology (GO).

The rest of this introductory chapter is divided into three parts. First, we start with an introduction to the biological problems that motivate the methodological development in statistics for RNA-Seq analysis; specifically, we give a brief overview of the high-throughput sequencing (HTS) technology (Section 1.1.2) and the RNA-Seq pipeline (Section 1.1.3); DE tests and GO enrichment analysis are particularly detailed in Sections 1.1.4 and 1.1.5, respectively. Second, we introduce some statistical techniques for modeling RNA-Seq read counts, including statistical models for count data (Section 1.2.1), generalized linear models (GLM) and negative binomial (NB) regression (Section 1.2.2), and NB dispersion models (Section 1.2.3). Since these biological and statistical concepts are shared among the three main chapters that follow, we therefore provide the relevant background and literature reviews in a unified manner in this chapter. We conclude in the third part with an outline of the objectives and structure of the dissertation in Section 1.3.

1.1 Biological Introduction and Motivation

1.1.1 The Central Dogma of Molecular Biology

The flow of genetic information in a biological system is revealed by the central dogma of molecular biology, which states that deoxyribonucleic acid (DNA) is *transcribed* into messenger ribonucleic acid (mRNA), and then *translated* into proteins that perform functions in cells (Crick [25]). Following this central dogma, we refer to "gene expression" as the process by which the information encoded in a gene is used to direct the assembly of a protein molecule. Though the machinery for expressing genes is highly sophisticated, it is clear that a gene's expression level (or activity), i.e. the amount of gene product (protein)

present and functioning in a cell, is proportional to the mRNA transcript abundance of that gene, which provides an accurate snapshot of the gene's expression profile. To understand the dynamic system, one essential component involves analyses of the transcriptome (the set of all RNA molecules in an organism), such as measuring transcriptome composition and discovering novel genes. In the next section, we will introduce in more details the state-of-the-art RNA-Seq technology and some of the"second-generation" sequencing platforms widely used for current transcriptome studies.

1.1.2 High-Throughput Sequencing Technology

Over the last ten years, the technologies of choice for large-scale genomic studies have been switching from DNA microarrays to sequencing-based approaches that directly determine complementary DNA (cDNA) sequence. These approaches have evolved from the early (first generation) Sanger sequencing of cDNA or EST libraries (Boguski et al. [14], Gerhard et al. [41]), to tag-based methods such as serial analysis of gene expression (SAGE; Velculescu et al. [117]), and to the recent development of novel HTS methods for RNA (RNA-Seq) that have fundamentally changed how we investigate transcriptomics. As RNA-Seq becomes the current technology of choice, we summarize some key advantages of RNA-Seq over the hybridization-based microarray technology (Wang et al. [121]): (1) RNA-Seq requires no existing knowledge about genome sequence as it explicitly sequences transcripts, which is ideal for studies of non-model organisms where genomic annotations are unavailable; (2) for the detection and quantification of alternative splicing events, RNA-Seq can reveal the location of transcription boundaries to a single-base resolution as well as sequence variations (e.g., single-nucleotide polymorphisms, SNPs), whereas specialized microarrays with probes spanning exon junctions have to be made for such purposes; (3)

RNA-Seq determines gene expressions more accurately as evidenced by the strong correlation with quantitative PCR analyses and RNA spike-in controls; (4) RNA-Seq has low background noise and a larger dynamic range of expression levels, whereas microarrays suffer from high technical noise due to cross-hybridization and lack sensitivity for highly or lowly-expressed genes.

Three major sequencing platforms have been released since 2011: Ion Torrent's Personal Genome Machine (PGM), Pacific Biosciences' RS and Illumina's MiSeq. Quail et al. [92] did a comprehensive comparison of these platforms with the popular Illumina Genome Analyzer (GA) and HiSeq 2000, and discussed their differences in terms of the generated data quality and the supported applications. The platform information for RNA-Seq experiments can be accessed via online databases and repositories, for example, the National Center for Biotechnology Information (NCBI). See Anders et al. [5] for the procedures of getting such meta-data via the NCBI Sequence Read Archive (SRA).

Along with the benefits RNA-Seq offers, many challenges and problems still remain unsolved at various stages in the RNA-Seq pipeline (see next section). For example, not all RNA molecules can be sequenced directly, so that large RNA molecules have to be fragmented into smaller pieces (200–500 base pairs) to be compatible with current deep-sequencing technologies. Such fragmentations introduce bias during library preparations, in addition to some other complications such as discrimination of abundant RNA species from polymerase chain reaction (PCR) artefacts, the inclusion of strand information, etc. (Wang et al. [121]). In addition, the so-called "transcript length bias" is present in RNA-Seq but not in microarrays. As pointed out in Oshlack and Wakefield [87], the confounding factor of transcript length will influence downstream GO analysis and need to be accounted for. Statistical adjustment methods have been proposed in Young et al. [126], Gao et al. [38] and Mi et al. [83]. Chapter 4 is devoted to the discussion of this problem.

The variability associated with sequencing-based technologies is often referred to as "technical variation" arising from measurement error (Marioni et al. [77], McCarthy et al. [78]). It is to be distinguished from the so-called "biological variation" that changes (in expression levels) between experimental subjects (Chen et al. [23]). Studies performed in Marioni et al. [77] indicate that, across *technical replicates* (repeating identical web-lab and sequencing protocols on a single biological sample), only a small proportion ($\sim 0.5\%$) of genes have variations not captured by a simple Poisson model. The variations between *biological replicates*, however, need to be properly accounted for using more flexible statistical models, so that appropriate statistical inference on testing differentially expressed genes can be made (see Section 1.1.4). We will give a general overview of effective variation-modeling approaches in Section 1.2.1.

1.1.3 The Pipeline of RNA-Seq Analysis

Even though the chemistry and procedures substantially differ across HTS platforms, in general a typical RNA-Seq pipeline can be summarized as follows: purified RNA samples are converted to a library of cDNA with attached adaptors, and then sequenced on a HTS platform to produce millions of short sequences (about 30 to 400 base pairs long depending on the DNA sequencing technology used; a.k.a "sequence reads" or just "reads" for short) from one or both ends of the cDNA fragments (called "single-end" or "paired-end" reads, respectively). These reads are aligned to either a reference genome (e.g., genome sequenced in the Human Genome Project) or a reference transcriptome—an important step called *sequence mapping*, or assembled *de novo* without the genomic sequence. The aligned reads are then summarized by counting the number of reads mapped to the genomic feature of interests (e.g., exons or genes). The expression profile is eventually represented by a

Table 1.1: Statistical methodologies and computational tools in the RNA-Seq pipeline (partial): normalization, differential expression (gene-level) and GO analysis.

Analysis Stage	**Method/Category**	**Software**	**Reference**
Normalization	RPKM/FPKM	ERANGE	[84]
		Cufflinks	[113]
	Upper-quantile	Myrna	[19, 62]
	TMM	edgeR	[99]
	Median of fold change	DESeq	[3]
DE (gene-level)	Poisson	DEGseq	[120]
		TSPM	[9]
		PoissonSeq	[71]
	Negative Binomial	edgeR	[102]
		DESeq	[3]
		NBPSeq	[32]
		baySeq	[52]
		EBSeq	[68]
	Beta-Binomial	BBSeq	[128]
	Poisson-Tweedie	tweeDEseq	[36]
	Non-parametric	NOISeq	[112]
		SAMseq	[70]
		NPEBseq	[13]
	Quasi-likelihood	QuasiSeq	[76]
		AMAP.Seq	[107]
	Shrinkage	DSS	[124]
		ShrinkSeq	[116]
	Transformation	limma (voom)	[108, 63]
GO analysis	Wallenius approx.	GOseq	[126]
	logistic regression	GOglm	[83]
	transformation-based	[R codes]	[38]

Abbreviations: RPKM/FPKM, Reads/Fragments Per Kilobase of transcript per Million mapped reads; TMM, Trimmed Mean of M-values; DE, Differential Expression; GO, Gene Ontology.

matrix of read counts having non-negative integer values, where rows are genes (or some other genomic features like exons) and columns are samples. Subsequent steps that rely heavily on statistical analyses include normalization of reads and DE tests (of genes or exon usage). See Oshlack et al. [88] for an overall discussion of the RNA-Seq analysis process. Some downstream analyses, after getting a list of DE genes that have passed multiple-test adjustment, include tests for enriched categories (GO enrichment analysis) and network inference. We summarize in Table 1.1 some statistical methodologies and computational tools developed in recent years for the analysis stages of normalization, differential expression (mostly at the gene-level) and GO analysis. Most of the software packages are part of the Comprehensive R Archive Network (CRAN; R Core Team [94]), or part of the Bioconductor project (Gentleman et al. [40]).

Many statistical methodologies proposed for RNA-Seq data analysis (especially for DE tests) directly use the read count matrix as the starting point, which is a great simplification with less concern to the ambiguity and variability during sequence mapping and count summarization. As the reference genome is never a perfect representation, decisions have to be made when reads arise from a spliced transcriptome rather than a genome, or reads cannot be uniquely assigned to a feature of interest (mostly results from the "multi-reads" issue that read are aligned to multiple locations), among other scenarios (Oshlack et al. [88]). With different aims and scopes, a variety of short-read aligners have been developed (Grant et al. [47]), including general aligners such as `GSNAP` (Wu and Nacu [125]), `Bowtie` (Langmead et al. [61]) and `SHRiMP` (Rumble et al. [104]), and *de novo* annotators such as `TopHat` (Trapnell et al. [114]) and `SpliceMap` (Au et al. [8]). They differ in some evaluation metrics such as mapping accuracy, memory efficiency and the ability to detect splice junctions. In this dissertation, unless otherwise specified, we assume that RNA-Seq reads have been properly aligned under some alignment quality controls without going

into details for any particular alignment algorithm used. For quality controls of RNA-Seq experiments and relevant computational tools, see, for example, `FastQC` and Wang et al. [119] with the `RSeQC` package.

For accurate comparisons of gene expressions between and within samples, normalization is an essential step in both microarray and RNA-Seq data analyses. It is essential to ensure the observed expression differences truly reflect differential expression rather than nuisance experimental or technical effects [97]. Sources of systematic variation present in RNA-Seq experiments can be divided into "*between-sample* differences" and "*within-sample* gene-specific effects". For the former, different sequencing depths (a.k.a. observed library sizes) of the samples make the observed read counts not directly comparable across samples (Mortazavi et al. [84]); for the latter, technical confounders such as gene length (Oshlack and Wakefield [87]) and guanine-cytosine content (GC-content; Pickrell et al. [90]) will influence the measure of a gene's expression level. For absolute quantifications of expression levels and downstream analyses, these within-sample effects need to be properly accounted for. However, they are less of a concern if we test differential expression of individual genes as they will cancel out between samples. See Hansen et al. [51] and Risso et al. [97] for gene-specific normalizations. In the context of DE tests, the between-sample normalization is more essential as it not only involves adjustment for different observed library sizes, but also the (more subtle) apparent reduction or increase in the expression levels of non-DE genes to accommodate the increased or decreased expression of a few truly DE genes (Di et al. [33]). There is no consensus on the most appropriate normalization method, and performances seem to be experiment-dependent. See Bullard et al. [19] for the impact of normalization on DE test results and Dillies et al. [35] for a comprehensive evaluation on a number of popular normalization approaches. We refer to the normalizing factors in more details in Section 1.2.2, and introduce how they can be incorporated into statistical

models there. We note that these normalization methods have an common underlying assumption that only a small proportion of genes are truly differentially expressed. Risso et al. [98] recently published their work on using spike-in sequences as controls to relax this assumption, but that is beyond the scope of our discussions.

1.1.4 Differential Expression in RNA-Seq Data

Because many biological studies primarily aim to profile gene expressions between samples under different environmental or experimental conditions, in this dissertation we restrict our attention to statistical inferences at the *gene-level* where the rows in a summarized read count matrix represent genes. Despite that such gene-level summarization may not recover true differential expression if there are both up-/down-regulated isoforms present in a gene (Trapnell et al. [114]), many statistical methods which start from a count matrix have been developed and they are quite extensible to analyzing other types of count data beyond RNA-Seq, for example, comparative analysis of immunoprecipitated DNA (ChIP-seq, MBD-seq), proteomic spectral counts and metagenomics data (Anders et al. [5]). In fact, with moderate modifications, the statistical theories and methodologies discussed in subsequent chapters are also applicable for analyses at the level of exons or other genomic constructs of interests: Anders et al. [4] is one such attempt that, given a catalog of transcripts, tests for differential exon usage by generalized linear models and accounts for biological variations via the NB distribution (after flattening gene models and forming counting bins from exons); Glaus et al. [42] is another attempt that adopts a Bayesian approach for estimating expressions at the *transcript level*.

Genes with zero or very low read counts are in general of less interest in DE analysis, in the sense that (1) a gene of biological importance needs to be expressed at some minimal

level; (2) little information is available to distinguish between the null and alternative for lowly-expressed genes. Therefore, it is a common practice to specify some *ad hoc* criteria for filtering out genes with very low read counts before performing any statistical analyses. For example, we may require the average read counts for a gene across samples to be greater than five, or require the count-per-million (CPM) value greater than one. We often observe that, in general, genes with low counts have high over-dispersion in the NB setting and are often poorly fitted by conventional NB dispersion models (see Section 1.2.3). Van De Wiel et al. [116] suggested using zero-inflated NB (ZINB) models in such scenarios, but for DE tests in general, we prefer to subset the dataset in advance to achieve more accurate statistical inference and possibly avoid computational issues.

From Table 1.1, we see that the RNA-Seq community is now flooded with statistical methods for DE tests, and many of them adopt the NB distribution to account for the biological variation. However, there is a lack of research on the adequacy of using NB regression and dispersion models on analyzing RNA-Seq data. A simulation-based goodness-of-fit (GOF) test discussed in Chapter 2 is the first attempt to such investigations. Evaluation on the power-robustness trade-offs for DE tests using different NB dispersion methods are closely relevant to GOF tests, but surprisingly, it has received little attention in the literature. We will formally address these topics in Chapters 2 and 3. Before that, in Section 1.2 we will first give an introduction to the statistical methods that are integral to DE tests.

1.1.5 Gene Ontology Enrichment Analysis

In system biology, researchers can gain more biological insights by looking at sets of genes that differ in expressions. Downstream analyses after testing for DE genes aim to high-

light biological processes which are over-represented among those DE genes. To relate the outcome of the DE analysis to biological functions, a widely-used approach is to examine enriched Gene Ontology (GO) categories based on the terms annotated to the genes identified as DE (Ashburner et al. [6], Khatri and Drăghici [58]).

For RNA-Seq data, the value of a read count (for a particular gene) summarized in the count matrix is proportional to its gene length times the true mRNA expression level, under idealized assumptions of no alignment error or sequencing bias (Marioni et al. [77]). Therefore, a longer transcript tends to have more reads mapped to it. Because statistical tests are more powerful in detecting DE from those longer transcripts, even if those shorter transcripts have similar expression levels, they will be at a disadvantage for being declared as DE. Oshlack and Wakefield [87] used three RNA-Seq datasets to empirically demonstrated that such kind of "bias" is *inherent* to the standard RNA-Seq process, regardless of which protocols, normalization or re-scaling methods are used. Such bias does not present in microarrays as the intensity-based measurements are only proportional to the expression level plus features intrinsic to the probe like GC-content.

Statistically speaking, the term "length bias" seems to be inappropriate because essentially the transcript length is a confounding factor as it correlates with both the GO membership and the significance of the DE test. Several adjustment approaches have been proposed to alleviate the effects of gene length on gene-set analysis. Bullard et al. [19] suggested a modification of a DE t-statistic by dividing the square root of gene length. Gao et al. [38] also considered adjustment on each gene's test statistic (at gene level) and the null distribution for the Fisher's exact test (at gene-set level). Young et al. [126] proposed a resampling approach and used the Wallenius approximation. We will address this problem from a different perspective in Chapter 4.

1.2 Statistical Techniques for Modeling Count Data

In this section, we will introduce statistical methodologies for modeling count data in the context of analyzing RNA-Seq experiments. Because of the special characteristics of sequencing data, for example, counts that exhibit extra variations, small sample sizes but a large number of genes (a typical example of "large p, small n" scenario), many existing methods either fail to satisfy the basic assumptions for large-sample asymptotic inference, or require modifications in order to make full use of the data. We now give an overview of the statistical tools that are key components for the following three main chapters, and emphasize how biological factors and research questions mentioned in earlier sections can be modeled and formulated under different statistical frameworks.

1.2.1 Statistical Models for Count Data

Because of the discrete nature of RNA-Seq read counts, it is tempting to assume the Poisson distribution for each read count Y_{ij} of the i^{th} gene and j^{th} sample: $Y_{ij} \sim \text{Poisson}(\mu_{ij})$ with $\text{E}(Y_{ij}) = \text{Var}(Y_{ij}) = \mu_{ij}$, where the variance is completely determined by the mean. In fact, early RNA-Seq studies had shown that the Poisson model was adequate to account for technical variations (Marioni et al. [77]) and some statistical packages were developed based on this model (see, for example, Wang et al. [120] and Auer and Doerge [9]). However, as revealed by many recent RNA-Seq analyses, the one-parameter Poisson model has severe limitations for modeling the variations between biological replicates (Di et al. [31], McCarthy et al. [78]). Since biological samples are essential and required in experiments which aim to provide broader inference scope not limited to the particular samples being studied, appropriate accommodations of such biological variations are considered to be the rule rather than the exception for almost all current RNA-Seq data analyses.

The negative binomial (NB) distribution is a natural extension to the Poisson distribution which includes an extra parameter to account for the extra-Poisson variations (i.e. over-dispersions). It can be derived as a mixture of Poisson distributions in the so-called Gamma-Poisson model. For a random variable Y having an NB distribution with mean μ and dispersion ϕ, the variance $\mathrm{Var}(Y) = \mu + \phi\mu^2$ takes a quadratic form on the mean, and the dispersion parameter ϕ determines the extent to which the variance exceeds the mean. This formulation is often referred to as the "NB2" model, a conventional NB model among a total of 22 varieties of NB regression for modeling count data (see Chapter 8 in Hilbe [53]). Denote $\kappa = 1/\phi$ as the shape parameter, the probability mass function (p.m.f.) of an NB random variable Y under this parameterization is

$$p(Y = y; \mu, \kappa) = \frac{\Gamma(\kappa + y)}{\Gamma(\kappa)\Gamma(1 + y)} \left(\frac{\mu}{\mu + \kappa}\right)^y \left(\frac{\kappa}{\mu + \kappa}\right)^\kappa, \quad y = 0, 1, \cdots \tag{1.1}$$

Quantities such as the likelihood function, the gradient (score function), the observed and expected information matrices can be calculated based on this p.m.f.

Several approaches were proposed in earlier works for testing and estimating the NB dispersion parameter ϕ. For statistical tests of detecting over-dispersion, Dean and Lawless [29] developed two statistics and used score tests for Poisson regression models; a unifying theory of testing for over-dispersion in Poisson and binomial regression models was later given in Dean [28]. For estimation, NB or mixed (random-effects) Poisson models and analyses based on weighted least squares or quasi-likelihoods for count data were discussed in Lawless [64] and Breslow [15] to accommodate over-dispersion. These approaches for testing Poisson against NB models and estimating the dispersion parameter are not directly applicable for analyzing RNA-Seq data, as the asymptotic assumptions and the requirement for large μ's are typically not satisfied in the RNA-Seq context. Therefore, statistical inference on the regression coefficients based on these works will in general be inaccurate

and at best sub-optimal on RNA-Seq data.

Because many RNA-Seq studies have shown that the NB distribution is an effective and flexible tool for modeling count variation, researchers had made more efforts on accurate estimation of the NB dispersion parameter (which is crucial for statistical inference of differential expression), given the limited number of samples available and a large dynamic range of expression levels. Small-sample estimation of the NB dispersion was discussed in Robinson and Smyth [101] using a quantile-adjusted conditional ML (qCML), but it was limited to completely randomized designs with two treatment groups only and did not generalize to the more complex regression settings (see Section 1.2.2). McCarthy et al. [78] advocated the use of adjusted profile likelihood (APL; Cox and Reid [24]) in the generalized linear model (GLM) framework, which has less bias than competitive estimators. Lund et al. [76] discussed the quasi-likelihood (QL) approach with shrunken dispersion estimates, which provided better estimates of false discovery rate (FDR). Empirical Bayes shrinkage for dispersion estimation has gained increasing popularity as it effectively share information across tens of thousands of genes. This technique has been implemented in many software packages including the empirical analysis of digital gene expression data in R (`edgeR`; Robinson et al. [102]), `DESeq` (Anders and Huber [3]) and `DESeq2` (Love et al. [75]). We will provide more details on relevant methods in Chapters 2 and 3.

In the simplest case of two-group comparison, it is also attempting to consider conventional two-sample tests for detecting DE genes. However, it turns out that using parametric discrete distributions for modeling counts has many benefits over a standard two-sample t-test for the difference between the two groups. For untransformed counts, the violations of the normality and homogeneity of variance assumptions make different versions of t-tests far from ideal. Tests on data after applying variance-stabilizing transformations (e.g., square root for the Poisson to stabilize the variance for different means) still have

unsatisfactory performance—for example, the FDR is not maintained at the nominal level (Di et al. [31]). Simulations have also revealed that as the mean decreases and fold change increases, testing with an appropriate discrete distribution (e.g., NB) will provide substantial power gains over transformed data. We will formally introduce the NB regression and dispersion models for RNA-Seq read counts in Section 1.2.3.

1.2.2 Generalized Linear Models and NB Regression

The generalized linear models (GLMs) proposed in Nelder and Wedderburn [85] provide a unified way for handling categorical and continuous response distributions. A GLM consists of two components: the *random component* specifies the distribution assumption on the response y_i given the covariates $\mathbf{x}_i$, and the *systematic component* specifies the link between the expected response and the covariates. For the random component, within the exponential family framework we can re-write the NB p.m.f. (1.1) as

$$p(Y = y; \mu, \kappa) = \exp\left\{ \left[\log(\pi) + (y/\kappa) \log(1-\pi) \right] / (1/\kappa) + \log\left(\frac{\Gamma(\kappa + y)}{\Gamma(\kappa)\Gamma(1+y)} \right) \right\}$$

where $\pi = \kappa/(\mu + \kappa)$. For the scaled response y/κ and dispersion $\phi = 1/\kappa$, we obtain a simple exponential family for *fixed* κ. For the systematic component, instead of using a canonical link function $\eta = \log\left(\frac{\mu}{\mu+1}\right)$ that may cause problem when $\eta \to 0$ ($\mu \to \infty$), it is more common to adopt the log-link $\eta = \log(\mu)$ so that the linear predictor η is not restricted. For a univariate count response, we relate the linear predictor $\eta = \log(\mu)$ with a number of covariates by $X'\beta$, where X and β are p-dimensional vector of known explanatory variables and unknown regression coefficients, respectively. The flexibility of GLMs for accommodating arbitrarily complex designs with multiple treatment conditions and/or blocking variables is essential as the complexity of the study designs for typical

RNA-Seq experiments can vary a lot—from the simple case of comparing gene expressions between two conditions (e.g., stimulated versus unstimulated, or wild-type versus mutant samples), to more complicated designs including additional experimental factors (e.g., drug doses or time points) and additional covariates (e.g., non-biological experimental variation or "batch effects", pairing of tumor and normal tissues, or difference in library type). For multi-factor designs, methods developed within the GLM framework are more superior to their counterparts (say, `Cufflinks`, Trapnell et al. [113]) that cannot handle complex designs such as testing for interaction effects between factors.

We fit a GLM to each of the tens of thousands of genes in the dataset, so here we present the model for a single gene i and repeat the process for the rest of the genes. As discussed in the RNA-Seq pipeline in Section 1.1.3, normalizing read counts between samples involves adjustment for different observed library sizes (denoted by N_j) and the apparent reduction/increase in the expression levels of non-DE genes to accommodate the increased/decreased expression of a few truly DE genes (an adjustment term denoted by R_j). These factors can be modeled in the GLM framework: denote $\pi_{ij} = \frac{\mu_{ij}}{N_j R_j}$ as the relative mean frequency, so that the log mean expression can be decomposed as

$$\log(\mu_{ij}) = \log(N_j) + \log(R_j) + \log(\pi_{ij})$$

with two known additive constants. These quantities $\log(N_j)$ and $\log(R_j)$ which are called *offsets* in the GLM terminology are pre-estimated and treated as known during GLM fitting. The *ad hoc* library size (quantile) adjustment procedure required for the validity of exact NB tests for two-group comparison problems (Robinson and Smyth [100]) is automatically accounted for in this GLM framework. Note that here (and in many applications) the same constant ($N_j R_j$) is assumed for all genes in a sample, but it may be advantageous to calculate between-gene normalization factors s_{ij} to account for some gene-specific sources

of technical biases such as GC-content and gene length (Love et al. [75]). See Hansen et al. [51] and Risso et al. [97] for relevant discussions. This kind of normalization can be treated in the same manner in the GLM framework as well.

For making comparisons between gene expression levels under different conditions j, we have (for the i^{th} gene):

$$\log(\mu_{ij}) = \text{offset} + \sum_{k=1}^{p} \beta_{ik} X_{jk}$$

where in the case of $p = 2$, $X_{j1} = 1$ for all j; $X_{j2} = 1$ if sample j is from group 2 and $X_{j2} = 0$ if sample j is from group 1. This is the popular "treatment contrasts" parameterization due to its easy interpretation in terms of log fold changes (cf. `contr.treatment` in `R`, contrasts each level with the baseline level). Other options (e.g. sum contrasts) are also used. The estimate for β_1 gives the estimate of group 1's overall expression level, while the remaining coefficients $\beta_2, \cdots, \beta_k$ are interpreted as log fold changes against group 1 (the baseline level).

The maximum likelihood estimates of the regression coefficients β in the linear predictor can be obtained by the iteratively re-weighted least squares (IRLS) algorithm that works for all GLMs (McCullagh and Nelder [79]). To numerically solve the maximum likelihood equations, depending on calculation complexity of the observed or the expected information, the Newton-Raphson algorithm or the Fisher scoring method will be used. In the $(k+1)^{th}$ iteration of the Newton-Raphson algorithm, the coefficient β_{k+1} is given by $\beta_k + \frac{U(\beta_k;y)}{\mathcal{I}(\beta_k;y)}$, where $U = \frac{\partial l}{\partial \beta}$ is the score function and $\mathcal{I}$ is the observed information. Fisher scoring replaces $\mathcal{I}(\beta_k; y)$ with the expected information $\mathcal{J}(\beta_k)$. Convergence is achieved if some criterion for estimates over consecutive iterations is satisfied, for example, $|\beta_{k+1} - \beta_k| < \epsilon$, a pre-specified threshold.

For multi-factor design problems, the GLM allows for the possibility of testing any contrast of regression coefficients equals to zero (McCullagh and Nelder [79]), corresponding to

the formulation of different biological questions. In general, hypotheses of regression coefficients in GLMs are tested by asymptotic tests, most notably the Wald test and likelihood ratio tests (LRT) with the asymptotic χ^2 approximation to the likelihood ratio statistic (Wilks [123]). Di et al. [33] incorporates the LRT with higher-order asymptotic adjustments for more accurate inference under small sample size scenarios typical for RNA-Seq studies.

Testing differential expression is essentially equivalent to testing a coefficient or contrast of interest being zero. For the two-group comparison, the null hypothesis $H_0 : \beta_2 = 0$ is tested against either a one-sided or two-sided alternative hypothesis. Under the null hypothesis, twice the log likelihood ratio statistic has an asymptotic χ^2 distribution, and simulation studies have shown that the likelihood ratio test (LRT) controls Type I error rates reasonably well. Di et al. [33] further suggested LRT with higher-order asymptotics (HOA) adjustments. Lund et al. [76] discussed quasi-likelihood (QL) methods by replacing LRTs with QL ratio F-tests for better FDR controls, and the test statistics are based on quasi-dispersion parameter estimates or two variants called QLShrink and QLSpline for sharing information across genes.

1.2.3 NB Dispersion Models

Even though the cost of RNA-Seq experiments is decreasing each year, explicit trade-off exists between having more biological replicates and deeper sequencing. Liu et al. [73] showed that having more biological replications is an effective strategy for statistical power gains and increased accuracy in RNA-Seq studies. However, due to constraints in resources, current RNA-Seq experiments are commonly conducted in duplicates or triplicates in each condition. With such small replication levels, large-sample statistical theories are not

applicable most of the time. Small-sample inference and methods that effectively borrow information across tens of thousands of genes are becoming more important for RNA-Seq data analysis.

The discrete read counts of RNA-Seq data are characterized by strong mean-variance relationships that can be modeled by the NB distribution. However, modeling such variability is challenged by the relatively small number of replicates available in a typical experiment. Earlier works of Robinson and Smyth [100, 101] assumed a common dispersion estimated by all genes combined. Recognizing an evident trend relationship between the dispersion and relative gene expression, researchers have proposed many alternative NB dispersion models. Di et al. [31] (with the `NBPSeq` package) adopted a parametric dispersion-modeling approach by using the over-parameterized NBP distribution (Greene [48], Hilbe [53]), where the log dispersions are modeled as a linear function of the relative mean frequencies (on the log-log scale). A natural extension to NBP is the NBQ model which incorporates an extra quadratic term. Anders and Huber [3] (with the `DESeq`/`DESeq2` packages) suggested fitting a non-parametric curve to capture the dependence so that genewise estimates can be shrunk toward the values predicted by the curve (either a local or parametric regression is used to compute this trend). McCarthy et al. [78] (with the `edgeR` package) introduced a similar "trended" (non-parametric) model in which a different dispersion is estimated for each gene using APL and then modeled as smooth functions of the genewise average read counts. Two variants of the "tagwise" models that adopted empirical Bayes estimations and Cox-Reid approximate conditional likelihood (Cox and Reid [24]) were also introduced in McCarthy et al. [78] (we call them "tagwise-common" and "tagwise-trend"). In the most recent release of the `NBPSeq` package, the NBS approach models the dispersion as a smooth function (a natural cubic spline function) of the preliminary estimates of the log mean relative frequencies, and the "step" approach models the

dispersion as a step (piecewise constant) function. Quasi-likelihood dispersion estimates (either borrowing information across genes or not) were proposed in Lund et al. [76], with implementations in the `QuasiSeq` and `edgeR` packages.

As listed in Table 1.1, various NB dispersion models are flooding the RNA-Seq literature and new methods are expected to be proposed in the future. However, discussions on evaluating the model adequacy and the associated impacts (in terms of power and robustness) on DE tests are rather limited. We address this problem by proposing a simulation-based goodness-of-fit test of NB regression and dispersion models with diagnostic graphics in Chapter 2, with some initial power-robustness evaluations. Chapter 3 further discusses a simple statistic we proposed for quantifying the inadequacy or lack-of-fit of a fitted dispersion model, with more comprehensive power-robustness investigations into several RNA-Seq datasets across different species.

1.3 Dissertation Objectives and Structure

The topics in this dissertation are motivated by practical biological problems in life sciences, where large-scale genomics sequencing data are generated by the state-of-the-art RNA-Seq technology. This dissertation includes three main chapters of manuscripts published or ready for submission. Specifically, in Chapter 2 we discuss goodness-of-fit (GOF) tests and model diagnostics for NB regression and dispersion models for RNA-Seq data analysis. We introduce in Chapter 3 a direct approach for quantifying residual variation in fitted dispersion models, which provides a measure that complements the GOF tests and facilitates realistic power-robustness evaluations; we also investigate into several RNA-Seq datasets across species and use simulation results to show that the magnitude of the residual dispersion variation gives hints on whether we can gain statistical power by a dispersion-modeling

approach. Chapter 4 is about the downstream analysis after DE tests, the Gene Ontology (GO) enrichment analysis, where we propose a simple but effective approach to account for the confounding factor of transcript lengths in RNA-Seq data. Chapter 5 provides a summary of the dissertation and an outline of future work.

2 Goodness-of-Fit Tests and Model Diagnostics for Negative Binomial Regression of RNA Sequencing Data

Gu Mi, Yanming Di and Daniel W. Schafer

Biometrics

International Biometric Society Business Office

1444 I Street, NW, Suite 700

Washington, D.C. 20005, USA

(Current status: revision submitted and under review)

2.1 Abstract

This work is about assessing model adequacy for negative binomial (NB) regression, particularly (1) assessing the adequacy of the NB assumption, and (2) assessing the appropriateness of models for NB dispersion parameters. Tools for the first are appropriate for NB regression generally; those for the second are primarily intended for RNA sequencing (RNA-Seq) data analysis. The typically small number of biological samples and large number of genes in RNA-Seq analysis motivates us to address the trade-offs between robustness and statistical power using NB regression models. One widely-used power-saving strategy, for example, is to assume some commonalities of NB dispersion parameters across genes via simple models relating them to mean expression rates, and many such models have been proposed. As RNA-Seq analysis is becoming ever more popular, it is appropriate to make more thorough investigations into power and robustness of the resulting methods, and into practical tools for model assessment. In this article, we propose simulation-based statistical tests and diagnostic graphics to address model adequacy. We provide simulated and real data examples to illustrate that our proposed methods are effective for judging adequacy of fit of several NB dispersion models. The GOF test results on a small experiment concerning the plant Arabidopsis and a simulation study based on the conditions found in that example indicate that there is a power advantage in capturing the trend (either parametrically or non-parametrically) and that the methods based on an exact trend are robust against noise about the trend of the size found in the example.

Keywords: Dispersion modeling; Goodness-of-fit test; Negative binomial regression; Overdispersion; Residual diagnostics; RNA-Seq

2.2 Introduction

2.2.1 Negative Binomial Models

The negative binomial (NB) model has been widely adopted for regression of count responses because of its convenient implementation and flexible accommodation of extra-Poisson variability. Let Y represent a univariate count response variable and X a p-dimensional vector of known explanatory variables. Then an NB log-linear regression model specifies that the probability distribution of Y is NB with mean μ and dispersion parameter ϕ, with $\log(\mu) = X'\beta$ where β is a p-dimensional vector of unknown regression coefficients.

The NB distribution can be derived as a Poisson-gamma mixture model. For the conventional parameterization (which we refer to as NB2), suppose υ is a gamma-distributed random variable with $\mathrm{E}(\upsilon) = \mu$ and $\mathrm{Var}(\upsilon) = \phi\mu^2$, and that $Y|\upsilon \sim \mathrm{Poisson}(\upsilon)$, then the marginal distribution of Y is NB with mean μ and variance $\mu + \phi\mu^2$ [see, for example, 65]. The NB2 probability mass function (p.m.f.) has the form

$$f(y|\mu,\phi) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\Gamma(y+1)}\left(\frac{\theta}{\mu+\theta}\right)^{\theta}\left(\frac{\mu}{\mu+\theta}\right)^{y}$$

where $\theta = 1/\phi$.

Other NB parameterizations follow from different parameterizations for the gamma mixing distribution. A general form, called NBP (Greene [48], Hilbe [53]), follows from the assumption that the gamma variance is $\phi\mu^{\alpha}$, and has the same form of p.m.f. $f(y|\mu,\phi)$, but with θ replaced by $\phi^{-1}\mu^{2-\alpha}$. In this parameterization, $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \mu + \phi\mu^{\alpha}$. We note that (1) for identically distributed count variables the NBP distribution is over-parameterized, but in a regression setting it offers additional flexibility in mean-variance

modeling, which is useful in the RNA sequencing (RNA-Seq) analysis that follows; and (2) NBP includes the well-known NB1 (α = 1) and NB2 (α = 2) parameterizations as well as others. Greene [48] specified the symbol "P" for our α, which is why this parameterization is called "NBP".

2.2.2 RNA-Seq Analysis and NB Regression

RNA-Seq analysis (Wang et al. [121]) may be performed on biological units from any of the traditional forms of life science study, such as randomized experiments with multiple treatments and covariates, or observational studies with multiple observed explanatory variables. The response variable for each unit is a vector of relative frequencies, which measure expression levels for each of a large number of genes or gene isoforms. Although much of the statistical attention to RNA-Seq analysis has so far focused on the two-group problem – and, therefore, on identification of differentially expressed genes – there is a clear need for regression analysis for identifying differential expression after accounting for other variables, and for identifying patterns of expression and differential expression as a function of explanatory variables.

Future statistical techniques might be derived for the multivariate regression on all genes simultaneously, but the problem is currently tackled by the simpler univariate regression on each gene individually, with appropriate attention to false discovery rate. The response for a single gene is the number of RNA-Seq reads corresponding to that gene (Y) out of a total number of reads for the particular biological unit (s). Although there is evidence that the "technical variability" in Y – meaning the variability in the RNA-Seq technical procedure repeated on a single biological unit – can be described by a Poisson distribution (Marioni et al. [77]), the observed variability from multiple biological units in the same

observational or experimental group is greater than Poisson (see, for example, Anders and Huber [3], Di et al. [31]). The gamma mixture of Poissons, as described in Section 2.2.1, is a conceptually appealing alternative because the gamma mixing represents "biological variability". Practically, the NB model is both flexible and convenient.

The primary statistical challenge involves simultaneous regression fitting for tens of thousands of genes from fairly small numbers of biological samples (e.g. less than twenty). An important power and efficiency issue in this case involves the modeling of the NB2 dispersion parameter ϕ. Five possibilities, for example, are (1) ϕ is constant for all genes; (2) ϕ is allowed to differ between genes but is constant within gene under all conditions; (3) ϕ is allowed to differ for all gene/condition combinations; (4) ϕ is taken to be a function of μ; and (5) ϕ is taken to have a trend as a function of μ, but with some additional between-gene variability. More flexible models are much more likely to fit the data, of course, but at the expense of tens of thousands of nuisance parameters. If a more specific model fits, based on vastly fewer nuisance parameters, it could offer substantial power and efficiency gains (for improved "true discovery" rates of differential expression, for example). Because of the very large number of hypothesis tests performed in a single RNA-Seq study and the very large number of RNA-Seq studies being performed world-wide, even a small improvement in power can have an important impact on the overall rate of scientific learning from the RNA-Seq technology.

Let Y_{ij} denote an RNA-Seq read count for the i^{th} gene ($i = 1, \cdots, m$) of the j^{th} experimental or observational unit ($j = 1, \cdots, n$), and $\mathbf{X}_j$ the associated p-dimensional explanatory variable. Suppose $Y_{ij} \sim \text{NB}(\mu_{ij}, \phi_{ij})$ where μ_{ij} is the mean and ϕ_{ij} is the dispersion parameter in the NB2 parameterization. Suppose also that $\log(\mu_{ij}) = \log(s_j) + \log(R_j) + \log(\pi_{ij})$, with $\pi_{ij} = \exp\left(\mathbf{X}_j'\beta_i\right)$, where s_j is the library size (the number of RNA-Seq reads in the biological sample from unit j), and R_j is an optional normalization factor estimated before-

hand (Anders and Huber [3], Robinson and Oshlack [99], Bullard et al. [19]) and treated as known. In this formulation, π_{ij} is the mean relative frequency of occurrence of RNA-Seq reads associated with gene i, which is taken to be the expression level of gene i associated with observational or experimental unit j.

We label some of the ways to model the nuisance parameters ϕ_{ij} as follows:

#1. Genewise: $\phi_{ij} = \phi_i$ (constant within each gene i across all conditions j), with m parameters for NB dispersion

#2. Common: $\phi_{ij} = \phi$ (constant for all gene/condition combinations), with one parameter for NB dispersion

#3. NBP: $\log(\phi_{ij}) = \alpha_0 + \alpha_1 \log(\pi_{ij})$, equivalent to assuming NBP response distribution discussed in Di et al. [31], with two parameters for NB dispersion

We also introduce here a new approach, in which the dispersion parameter trend is quadratic on the log scale:

#4. NBQ: $\log(\phi_{ij}) = \alpha_0 + \alpha_1 \log(\pi_{ij}) + \alpha_2 [\log(\pi_{ij})]^2$, with three parameters for NB dispersion

An important related method estimates the ϕ_{ij}'s via non-parametric regression:

#5. Non-parameteric: ϕ_{ij} is estimated in a first step as a smooth function of $\log(\hat{\phi}_{ij})$ on $\log(\hat{\mu}_{ij})$, and then treated as known in the second step of regression coefficient inference

In addition, there are variants that use an average of trend and individually-estimated dispersion parameters, based on empirical Bayes considerations [78]:

#6. Tagwise-common: ϕ_{ij} is estimated as a weighted average of the common and genewise estimates, based on empirical Bayes calculations

#7. Tagwise-trend: ϕ_{ij} is estimated as a weighted average of the non-parametric and genewise estimates, based on empirical Bayes calculations

Methods for inference from the common, genewise, non-parametric, tagwise-common, and tagwise-trend approaches are available in the `edgeR` Bioconductor package (Robinson et al. [102], Gentleman et al. [40]). The non-parametric method is also available in the `DESeq` package (Anders and Huber [3]). Both `DESeq` and `edgeR` provide options for non-parametric estimation of the dispersion parameters (the two packages differ in implementation details). The NBP and NBQ approaches are implemented in the `NBPSeq R` package (Di et al. [34], R Core Team [94]).

The details of estimation for these methods are important but are not relevant to the proposed diagnostic tools and so are not discussed here. For the first four models above, however, maximum likelihood (ML) estimation can be used for simultaneous inference of all parameters, and likelihood ratio tests for regression coefficients account for the uncertainty in estimation of the parameters associated with dispersion. The final three approaches, on the other hand, involve a two-step procedure: NB dispersion parameters are estimated in the first step and then treated as known in the second, and the uncertainty in estimating dispersion parameters is typically ignored. Because of this automatic incorporation of dispersion parameter uncertainty, the avoidance of user-required tuning parameters, and the optimality of likelihood ratio tests, ML based on models for NB2 dispersion is appealing. The adequacy of the models for RNA-Seq data is not yet well understood, though. We wish to use the model diagnostic tools proposed in this article to judge the degree of fit of the various models on real RNA-Seq data – particularly the fit of simple parametric models for the trend of $\log(\phi)$ as a function of π and the degree of noise, if any, about this trend,

so that realistic robustness and power studies can follow.

To further clarify this introduction, Figure 2.1 shows a log-log scatter plot of method-of-moments-like estimated NB2 dispersion parameters, $\hat{\phi}$, versus estimated mean relative frequencies, $\hat{\pi}$, for each of 19,623 genes from a single sample of size three of a pilot Arabidopsis RNA-Seq study examined in Di et al. [31]. The curves on the plot are estimated dispersion trends based on the models described above. Polynomial gamma log-linear regression models of $\hat{\phi}$ on $\log(\hat{\pi})$ were used for quick-and-dirty testing and quantification of the trend, as follows. The linear model explains 24.1% of the variability in logged dispersion parameter estimates. A quadratic term (with p-value < 0.0001) explains an additional 7.2% of variability. A cubic term (p-value < 0.0001) in a full cubic model explains less than 0.1% additional variability. This plot and informal analysis suggest that the common ϕ model is inadequate; the trend is primarily, but not entirely, linear; and that a quadratic model captures essentially all of the trend in this particular dataset.

A simple model for trend in NB dispersion parameter ϕ as a function of mean relative frequency π is a good starting point for reducing the number of nuisance parameters, but the evidence of a trend does not imply that the ϕ's fall exactly on the trend; there may be additional variability in ϕ for genes with the same value of π. The main questions we wish to address with diagnostic tools are the following: (1) Does the NB assumption hold for a very rich model (for both regression and dispersion)? (2) What relatively simple models are adequate for describing ϕ as a function of π? (3) Is there evidence of additional biological variability in ϕ between genes having the same value of π?

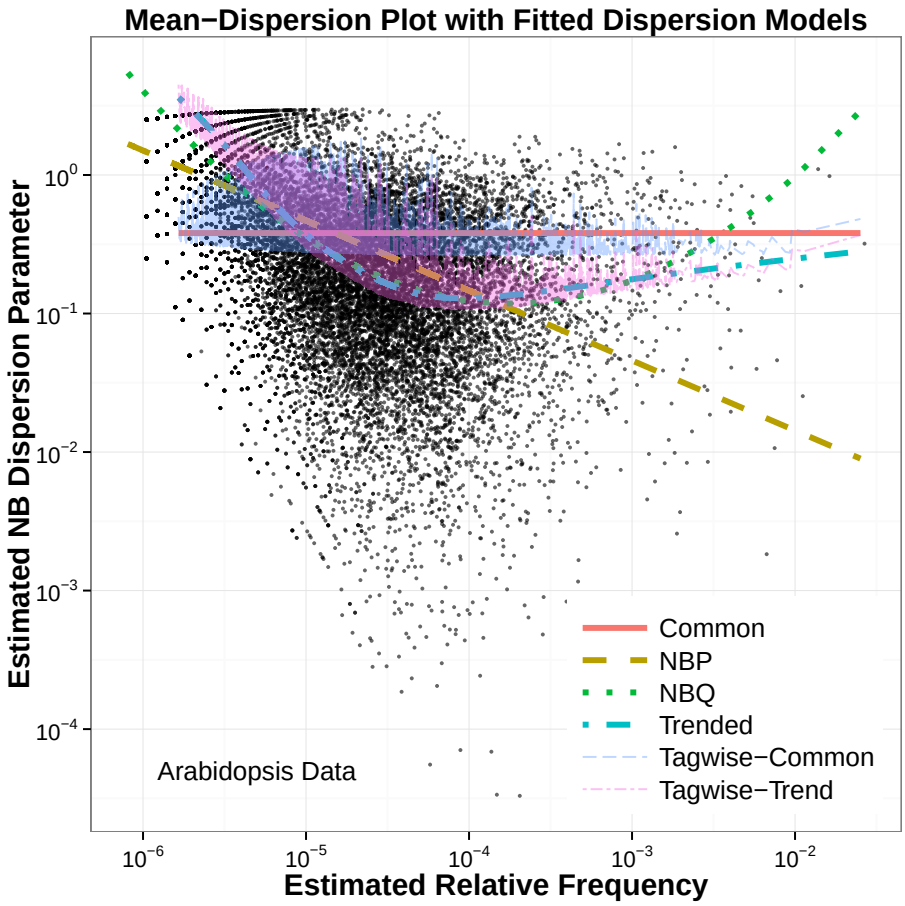

Figure 2.1: The mean-dispersion plot with six fitted dispersion models (common, NBP, NBQ, trended, tagwise-common and tagwise-trend) for the Arabidopsis RNA-Seq dataset (19,623 genes from three biological samples in the mock treatment group). The jagged curves for the tagwise procedures indicate the variability of individual NB2 dispersion parameters about the trend.

2.2.3 Diagnostics and Goodness-of-Fit for NB Regression

Traditional tools for model diagnostics in generalized linear models (GLM), such as deviance and Pearson residuals and goodness-of-fit (GOF) tests, are suitable for binomial and Poisson regression if the means are large, i.e. the adequacy of the normal and χ^2 null distributions for residuals and GOF test statistics, respectively, are justified under central-limit-theorem-like asymptotics rather than large sample size asymptotics (Pierce and Schafer [91]). Such GOF tests are not appropriate for small means (which are typical for the majority of genes in RNA-Seq analysis), and the theory for the null sampling distribution of the residuals and GOF test statistics does not extend to NB regression.

Best et al. [12] extended Anscombe's tests of fit for the NB distribution by using fourth order smooth tests, but these tests don't extend in an obvious way to regression models for non-exponential family response distributions. The test we propose in this paper gives similar results to theirs for i.i.d. samples and can also be used for the procedures that involve non-parametric trend fitting and empirical Bayes averaging. Esnaola et al. [36] proposed a larger family of response distributions for RNA-Seq analysis, which permits the testing of NB as a special case; but we do not believe the approach (validated under extensively replicated experiments) is suitable for the small sample sizes we have in mind here.

In this article, we propose a goodness-of-fit test statistic for NB regression based on Pearson residuals, and the calculation of a p-value using Monte Carlo-estimated null sampling distributions. The same simulations are used to estimate expected ordered residuals for an empirical probability plot. For RNA-Seq diagnostics, the GOF p-values from all genes are examined in a uniform QQ plot and combined via the Fisher's combined probability test (Fisher's method, Fisher [37]).

Similar regression diagnostic approaches that use Monte Carlo or resampling to de-

rive null sampling distributions of diagnostic quantities have been previously proposed for several situations. For ordinary linear regression, Atkinson [7] proposed half normal plots of jackknife residuals. For logistic regression, Landwehr et al. [60] proposed an "empirical probability plot" in which ordered residuals from the observed data are plotted against their expected values (or median values), as computed by Monte Carlo simulations. Their simulation procedure, which resembles parametric bootstrapping, is based on the estimated parameters from the fitted model. Similar graphical displays were adopted as informal checks of various count models. For example, Svetliza and Paula [111] considered normal probability plots for log-linear Poisson, log-linear NB and non-linear NB models. Garay et al. [39] evaluated GOF between zero-inflated Poisson (ZIP) and zero-inflated NB (ZINB) models. Both of these used simulated envelopes in their plots, but with standard normal quantiles (instead of quantiles from simulations) on the x-axis. None of the aforementioned papers provided statistical tests for evaluating model lack-of-fit.

2.2.4 Outline of Goals and Proposed Methods

In Section 2.3.1, we propose Monte Carlo-based GOF tests and graphical diagnostics for univariate NB regression. We demonstrate these tools in Sections 2.3.1 and 2.3.2 on real and simulated datasets. In Section 2.3.3, we use simulations to illustrate Type I error rates of the univariate NB regression GOF tests and power under various alternatives. In Section 2.4, we provide tools for judging NB dispersion models in RNA-Seq analysis by combining GOF test results on a random sample of genes. In Section 2.5 we illustrate the techniques on an Arabidopsis RNA-Seq dataset, and discuss the quantification of dispersion noise. The power and robustness of dispersion models are explored in Section 2.6 by a set of simulations. In Section 2.7 we document the conclusions and further directions for

research.

2.3 Residual QQ Plots and GOF Tests for NB Regression

2.3.1 Residual Plot and Test

We consider univariate NB regression in this section and then return to the RNA-Seq problem – of NB regression for each of many genes – in Section 2.4. For regression data with counted response, we wish to determine whether any NBP model fits and, because of the convenience of NB2 estimation programs, whether the NB2 model fits in particular. We propose two GOF tests and an associated residual plot. The test p-values provide an overall assessment of fit and the plot shows whether a small GOF p-value might be due to a small portion of the data. We use the same notation as in Section 2.2.2, but without the subscript i. In the RNA-Seq context, the methods of this section apply to a single gene. We start with Pearson residuals: $r_j = (y_j - \hat{\mu}_j)/\hat{s}_j$, where $\hat{\mu}_j$ is the estimated NB mean and $\hat{s}_j$ is the estimated NB standard deviation of y_j from the particular model being tested, for $j = 1, \cdots, n$.

We first propose an empirical probability plot of the ordered Pearson residuals $r_{(j)}$ versus the sampling distribution medians for each ordered Pearson residual, $\text{Med}[r_{(j)}]$, assuming the proposed NB regression model is correct. To approximate the medians, we simulate a large number of NB regression datasets of the same size and form as the observed one, using the data estimates as parameters for simulation; fit the same NB regression model to each simulated dataset; extract the ordered Pearson residuals; and retain the sample medians for each ordered residual. This is exactly the Landwehr et al. [60] approach applied to NB regression. Figure 2.2 shows the plot, along with a 95% pointwise prediction envelope (in dashed blue lines) formed from the similarly estimated

2.5^{th} and 97.5^{th} percentiles of the ordered residuals, for the NB2 and NBP regression of 48-year earthquake frequencies on magnitude (i.e. the Gutenberg-Richter Law, for 45 magnitudes from 4.5 to 9.1). Note that the band is formed from prediction intervals for the corresponding sample quantiles (not confidence intervals for their expected values). If the model fits, we would expect about 95% of the ordered residuals to fall within the band. We also superimpose a 95% simultaneous prediction band in solid red using the simulation method discussed in Buja and Rolke [18], so that for 95% of samples *all* ordered residuals should be contained in the red band. See Appendix (I) for details.

We wish to provide a global GOF test to accompany this empirical probability plot of Pearson residuals. A natural starting point is a test based on the Pearson statistic, i.e. the sum of squared Pearson residuals. The classical use of the χ^2 reference distribution is not appropriate here, but the null sampling distribution may be approximated by the Monte Carlo estimate. A p-value can be obtained as the proportion of simulated samples that produce a Pearson statistic as extreme or more extreme than the observed one. This is similar to the approach of Best et al. [12] in its application of parametric bootstrap to obtain a GOF p-value. We have found that our procedure and theirs give very similar results for single samples, but the approach discussed in Best et al. [12] requires specification of higher order moments, which is infeasible for the regression models we have in mind for RNA-Seq analysis.

Since the simulations provide estimated sampling distributions for each ordered residual, a finer test statistic is available as the sum of squared differences of the ordered residuals from their sampling distribution medians. We believed this was a worthwhile test statistic to consider given that we had already obtained approximate sampling distributions for each ordered residual to obtain the Landwehr et al. [60] type diagnostic plot. The test is also related to the SAM graphical procedure of Tusher et al. [115] for identifying differential

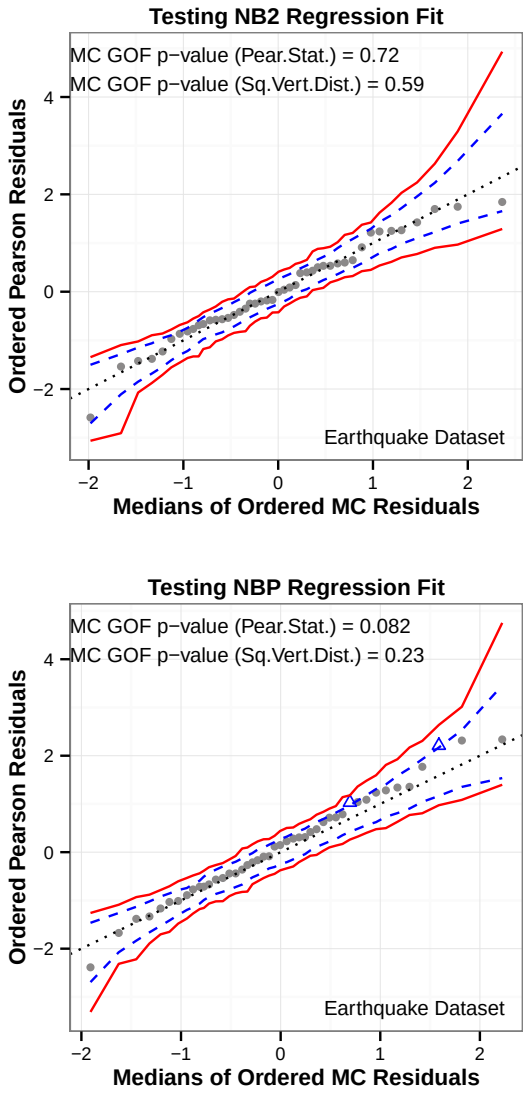

Figure 2.2: Empirical probability plots with GOF test p-values for evaluating NB2 and NBP model fits on the earthquake dataset (sample size: 45), with 95% prediction envelope in dashed blue lines and 95% simultaneous prediction band in solid red lines (based on 999 Monte Carlo simulations). Points outside the prediction envelope are flagged as blue triangles.

gene expression from microarray.

The following algorithm defines the diagnostic empirical probability plot of residuals and the Monte Carlo GOF test p-value based on the second test statistic.

#1. Fit an NB regression model from the data $\mathbf{Y}^{(0)} = (Y_1, \cdots, Y_n)^T$; estimate all unknown dispersion parameters, e.g. $\hat{\alpha}_0, \hat{\alpha}_1, \cdots$, and regression coefficients $\hat{\beta}^{(0)}$; calculate Pearson residuals $\mathbf{r}^{(0)} = (r_1^{(0)}, \cdots, r_n^{(0)})$ and the mean vector $\hat{\mu}^{(0)}$

#2. For $h = 1, \cdots, R$:

   a. Simulate a random vector $\mathbf{Y}^{(h)}$ from NB$(\hat{\mu}^{(0)}, \hat{\alpha}_0, \hat{\alpha}_1, \cdots)$

   b. Compute and retain Pearson residuals $\mathbf{r}^{(h)}$

#3. Find the median, 2.5^{th} and 97.5^{th} percentiles of the Monte Carlo sampling distribution for each ordered residual, denoted by $\tilde{r}_{(j)}^{50}, \tilde{r}_{(j)}^{2.5}$ and $\tilde{r}_{(j)}^{97.5}$, respectively; plot the ordered residuals from the observed data against the Monte Carlo medians; draw a 95% pointwise prediction envelope on the plot by connecting the $\tilde{r}_{(j)}^{2.5}$'s (for lower bound) and the $\tilde{r}_{(j)}^{97.5}$'s (for upper bound).

#4. Compute the sum of squared deviations of ordered residuals from the medians of their sampling distributions $d^{(h)} = \sum_{j=1}^{n} \left(r_{(j)}^{(h)} - \tilde{r}_{(j)}^{50} \right)^2$ for the observed data ($h = 0$) and for the simulated samples ($h = 1, \cdots, R$); compute a Monte Carlo GOF test p-value by

$$P_{\text{1sided}}^{MC} = \frac{\sum_{h=1}^{R} \mathbb{1}\left(d^{(h)} \geq d^{(0)}\right) + 1}{R + 1} \tag{2.1}$$

   where $\mathbb{1}(A)$ is the indicator function equal to 1 if the event A is true and 0 otherwise (Davison [27]).

The Pearson GOF p-value is computed in the same way, but using the sum of squared residuals as the test statistic. The two statistics are visualized on the empirical probability

plots as the sum of squared deviations about the $y = 0$ line and the sum of squared deviations about the $y = x$ dotted line. For this reason, we call the latter statistic the sum of squared vertical distances. The p-values for testing the NB2 and NBP models on the earthquake dataset are shown in Figure 2.2. The suggestive Pearson statistic p-value for goodness-of-fit of the NBP model is due to two outliers, corresponding to the frequencies of earthquakes of magnitudes 7.1 and 7.8. Although NB2 (with variance function $\mu+\phi\mu^2$ and NBP (in which the variance function is estimated to be $\mu+\phi\mu^{2.5}$) produce nearly identical fits, the standard errors (in the denominators of the Pearson residuals) from the NBP fit tend to be smaller for the smaller counts, which is why the NBP, but not the NB2 diagnostic, is detecting some potential lack-of-fit of the simple log-linear model in the region of magnitudes between 7 and 8 (corresponding to relatively small frequencies).

In using Monte Carlo simulation in lieu of theory to obtain the sampling distributions of the ordered residuals and test statistics, it would be ideal, but impossible, to use the *true* parameter values rather than their *estimates* from the data. Nevertheless, we expect the sampling distribution of the *residuals* to be about the same.

We use $R = 999$ Monte Carlo samples so that the (binomial) standard error in p-value estimation is 0.016 for p-values near 0.5, 0.007 for p-values near 0.05, and 0.003 for p-values near 0.01. As pointed out in North et al. [86], adding 1 to both the numerator and the denominator in Equation (2.1) produces a slightly biased estimate of the true p-value but with the correct Type I error rate, in contrast to the unbiased but anti-conservative p-value obtained without adding the 1.

2.3.2 Illustration on Simulated Datasets with Known Response Distributions

In Figure 2.3 we demonstrate the NBP and NB2 empirical probability plots and Monte Carlo GOF test p-values on four simulated regression datasets with known response distributions. The regression structure is taken to be the estimated log-linear model from the earthquake dataset of Figure 2.2, with sample size 45. For the first two scenarios, we generate NBP responses (with variance function $\mu + \phi\mu^{\alpha}$) as follows: (1) NB1, with variance 2μ and (2) NB2, with variance $\mu + 0.1\mu^2$. For the third scenario, we simulate NB2 responses as (2) above, but introduce "outliers": (3) "NB2 + Outliers", by randomly doubling three of the 45 counts. For the last one, we generate a mixture of NB2 distributions with different dispersion parameters: (4) "NB2 + Noise", with conditional variance $\mu + [0.1\exp(G)] \cdot \mu^2$, where $G \sim \mathcal{N}(0, 2^2)$. In this case the response counts are still gamma mixtures of Poissons, but the gamma variances are not constant.

In general, the two tests correctly indicate or fail to indicate lack-of-fit. An exception is that the Pearson test doesn't do as well at detecting the lack-of-fit in the last two columns. As evident in the last column of the empirical probability plots, the ordered Pearson residuals are larger in magnitude than expected in some regions and smaller in others. While the ordered residuals do not seem to behave as a sample from the tested distribution, the sum of their squares is moderated by the combination of small and large magnitudes. We are particularly concerned about this cancellation aspect of the Pearson test in the extension to RNA-Seq data, in which both under- and over-dispersion (relative to the tested model) may be present in subsets of genes. This issue is not relevant to the squared vertical distance test.

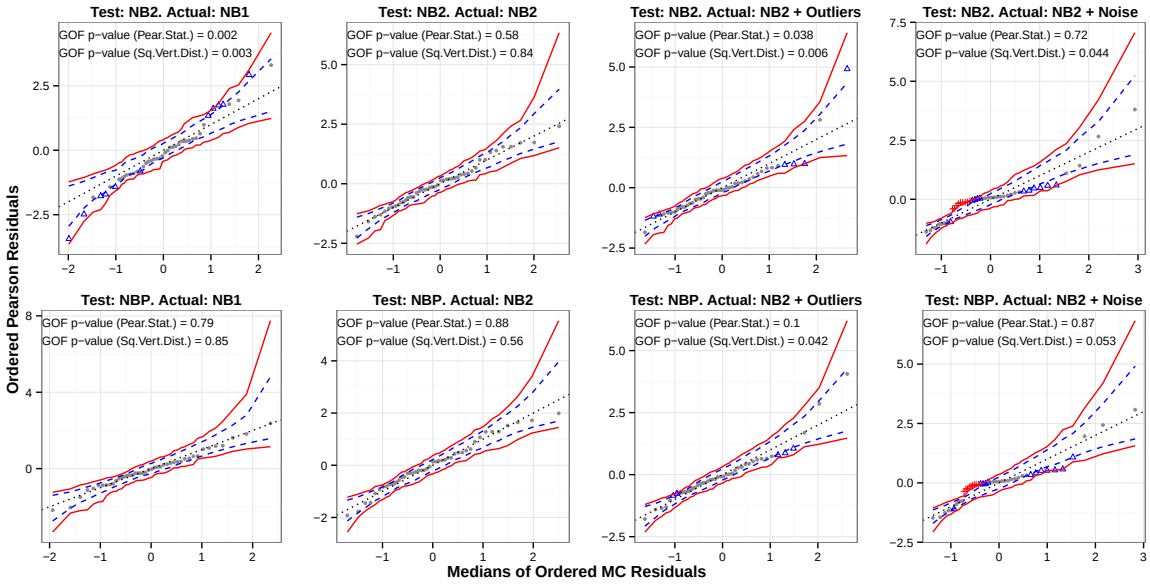

Figure 2.3: Empirical probability plots and GOF p-values for testing NB2 (top row) and NBP (bottom) on four simulated datasets with sample size = 45. The simulated response distributions are (left to right): NB1, NB2, NB2 with outliers and NB2 with random $\mathcal{N}(0, 2^2)$ noise added to $\log(\phi)$. We superimpose 95% prediction envelopes in dashed blue lines and 95% simultaneous prediction bands in solid red lines (based on 999 Monte Carlo simulations). Points outside the prediction envelope but inside the simultaneous confidence bands are flagged as blue triangles, and points outside the simultaneous confidence bands are flagged as red crosses.

2.3.3 Error Rates of GOF Tests in Simulations

Table 2.1: Type I error rates and rejection rates for 0.05-level NB2 and NBP GOF tests based on squared vertical distance ("Sq.Vert.D.") or Pearson statistics ("Pear.Stat."), from 1,000 simulated samples from each of several conditions. The standard error of simulation is approximately 0.007 for the Type I error evaluations, and the maximum standard error of simulation for the power evaluations is approximately 0.016. The simulation conditions are detailed in Appendix (II).

Type I Error Rate Evaluations									
		Sq.Vert.D.				**Pear.Stat.**			
GOF Test For	**Simulated Data n:**	5	10	50	100	5	10	50	100
NB2	NB2	0.045	0.055	0.049	0.054	0.041	0.032	0.051	0.052
NBP	NB1	0.031	0.042	0.057	0.041	0.025	0.034	0.049	0.042
	NB2	0.040	0.034	0.056	0.060	0.044	0.030	0.047	0.044
Reject Rate (Power) Evaluations									
		Sq.Vert.D.				**Pear.Stat.**			
GOF Test For	**Simulated Data n:**	5	10	50	100	5	10	50	100
NB2	NB1	0.17	0.17	0.37	0.55	0.17	0.19	0.48	0.71
	NB2 + Outliers	0.05	0.13	0.45	0.70	0.03	0.15	0.56	0.84
	NB2 + Noise	0.06	0.12	0.33	0.54	0.05	0.08	0.13	0.17
NBP	NB2 + Outliers	0.07	0.18	0.74	0.95	0.08	0.15	0.84	0.99
	NB2 + Noise	0.05	0.06	0.36	0.61	0.05	0.04	0.16	0.22

The top three rows of Table 2.1 show the Monte Carlo Type I error rates for 0.05-level tests using the NB2 and NBP Monte Carlo GOF tests on 1,000 simulated samples of NB1 and NB2 response distributions. The parameter specifications are detailed in the Appendix (II). The standard error of simulation is approximately 0.007. The Type I error rates are smaller than the nominal values for both tests at the small sample sizes. As the sample size increases, the Monte Carlo evidence is consistent with actual Type I error rates matching the nominal values. The severity of the small-sample conservatism is slightly greater for the Pearson test than for the squared vertical distance test.

The bottom five rows of Table 2.1 show estimated statistical power of the NB2 and NBP

Monte Carlo GOF tests under several alternative distributions. In the "NB2 plus noise" alternative, we add random $\mathcal{N}(0,1)$ noise to $\log(\phi)$ as described in Section 2.3.2, which means the data are a mixture of negative binomials with different dispersion parameters ϕ. In the "NB2 plus outliers" alternative, we randomly double 20% of the counts. The details of the generated distributions are provided in Appendix (II). The results do not indicate major power differences between the two tests, but the squared vertical distance test is more powerful in detecting the "NB2 plus noise" alternative.

2.4 Diagnostic Tools for RNA-Seq Modeling

A major step for improving RNA-Seq analysis is the comparative evaluation of the models and methods for incorporating commonalities of NB dispersion parameters within and across genes, as described in Section 2.2.2 and displayed in Figure 2.1. For studying these models on a given RNA-Seq dataset, we propose fitting them, calculating the squared vertical distance NB GOF p-value from Section 2.3.1 for each of a randomly selected sample of genes (i.e. testing the univariate NB regression model fit for each gene individually, using the NB2 dispersion parameter estimated according to the particular dispersion model), drawing a uniform QQ plot of the p-values, and calculating a single p-value using Fisher's method.

Let p_i be the GOF p-value for gene i, based on the parameter estimates from the global model being tested. Fisher's method produces a single GOF p-value by testing the conformity of the p_i's from m genes to a standard uniform distribution, i.e. by comparing the test statistic $X^2 = -2\sum_{i=1}^{m}\log(p_i)$ to the $\chi^2_{(2m)}$ distribution. Although it is possible to base this on all genes, we elect to reduce the computational burden by selecting a random sample m^* genes, and use $m^* = 1,000$ as a computationally tolerable value. We don't

have a direct way to study the suitability of this sample size for testing whether the p-values follow a uniform(0,1) distribution with Fisher's method; but we do have an indirect approach that helps. Let P be the proportion of genes with p-values less than 0.05. A binomial 95% confidence interval for P from a sample of 1,000 genes has half-width 0.0135, so we would be likely to detect lack-of-fit to the uniform(0,1) – if it were represented by a proportion of p-values less than or equal to 0.05 that is different from 0.05 – if the actual proportion is 0.0635 or greater. Although we use Fisher's method rather than this (arbitrary) binomial test, the binomial calculation provides some clarification of the type of departure from the uniform(0,1) that we are likely to detect with a sample of 1,000 genes. The m p-values are not exactly independent, as required for the theory of Fisher's combined test, but are approximately so because the global parameter estimates are based on such a large number of genes.

As we noted in Section 2.3.2, the Pearson statistic can give misleading results if there are combinations of under- and over-dispersion relative to the response distributional model being tested. We have found this problem to be exacerbated in the RNA-Seq setting and so focus only on the squared vertical distance estimator, which does not suffer from the same problem.

Along with the Fisher's combination of p-values, we suggest a uniform QQ plot of individual p-values to help reveal the nature of any lack-of-fit, indicated by a higher than expected proportion of small p-values. The proportion of genes with small p-values may have some effect on the thinking about appropriate models.

2.5 Application to an Arabidopsis RNA-Seq Study

2.5.1 Introduction to the Arabidopsis Study

Arabidopsis thaliana has been intensively studied as a model organism in plant biology. The Arabidopsis data discussed in Di et al. [31] contain RNA-Seq reads that aligned to more than 25,000 genes from two groups of Arabidopsis samples of size three each. The two groups of size three each were derived from plants inoculated with $\Delta hrcC$ of *Pseudomonas syringae* pv tomato DC3000 or 10 mM $MgCl_2$ (mock). The dataset used in this article comes from Di et al. [31], which is a subset of the data described in Cumbie et al. [26].

2.5.2 Goodness-of-Fit Analysis

The following are the GOF p-values (in parentheses) from fitting the seven dispersion models (described in Section 2.2.2) to a random sample of 1,000 genes: common (< 0.0001); NBP (0.04); NBQ (0.94); trended (0.21); genewise (> 0.9999); tagwise-common (> 0.9999) and tagwise-trend (> 0.9999). The corresponding uniform QQ plots of p-values are shown in Figure 2.4.

The p-values greater than 0.9999 for the genewise and tagwise methods, as can be seen in the uniform QQ plots, are due to fewer small p-values than expected from a uniform$(0, 1)$ distribution. The extra large p-value, therefore, cannot be taken as evidence of lack-of-fit and is most likely due to conservatism in the tests from small sample sizes, as evident in Table 2.1, when individual (genewise and tagwise) dispersion parameters are estimated from the small sample. Even a slight degree of conservatism in the individual NB GOF tests can produce a very small Fisher combination statistic when there are so many p-values being combined. The evidence from Table 2.1, and our experience with simulations and

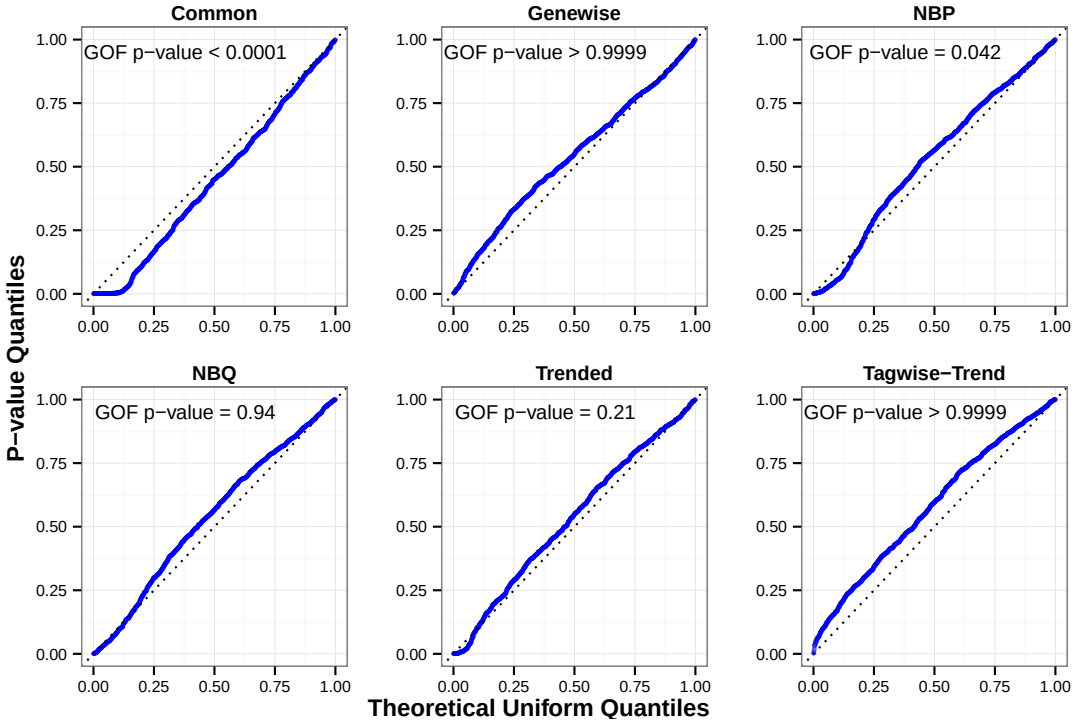

Figure 2.4: Uniform QQ plots of individual GOF p-values for the Arabidopsis dataset (based on a random sample of 1,000 genes from six experimental units in two experimental groups). The tagwise-common model (not shown) has a very similar pattern to the tagwise-trend model.

other RNA-Seq datasets suggests that this conservatism diminishes with increasing sample size.

The consistency of the data with the NBQ trend model and the non-parametric "trended" approach, suggest that models without noise about the the trend may be adequate. We need to exercise caution with this conclusion, though – since the lack of evidence for lack-of-fit does not prove "fit" and that the test may be conservative at this sample size. Nevertheless, the apparent fit of NBQ in conjunction with the evidence of lack-of-fit for NBP is intriguing.

Figure 2.1 shows the mean-dispersion plot (log-log scale) with six fitted dispersion models (common, NBP, NBQ, trended, tagwise-common and tagwise-trend) based on the mock treatment group alone. The genewise estimates are not included since there is no implied trend associated with that model.

2.5.3 Estimating Variability of NB Dispersion

The tagwise approaches allow for individual gene-wise variations of the NB2 dispersion parameter, ϕ, about a specified trend as a function of relative frequency, π (as indicated by the light blue and magenta jagged lines in Figure 2.1) via approximate empirical Bayes shrinkage of individual estimates towards the trend. In preparation for studying robustness of the methods that assume an exact NB dispersion parameter trend, we wish to quantify this degree of individual variability from real RNA-Seq datasets. For this purpose, we consider a simple model for noise about the trend in which $\log(\phi)$ is the sum of a trend and a $\mathcal{N}(0, \sigma^2)$ random variable. As a diagnostic measure to accompany the GOF p-value and uniform QQ plot, we propose an estimate of σ derived from a calculated quantity, "evar", in the `limma` package (Smyth [109]).

The variable *evar* is a sum of squares of a certain kind of residual associated with the empirical Bayes estimation in the tagwise approaches, and is associated with the variability of the gene-wise NB2 dispersion parameter about the trend. To use this to estimate σ in the model that is more convenient for our purposes, we use a calibration approach as follows. For each of several values of σ, we simulate RNA-Seq datasets from the "NB plus noise" model, with parameters taken to be the estimates from the given dataset, but with noise variance σ^2; we obtain the value of *evar* from a tagwise fit to each simulated dataset, fit the regression of *evar* on σ, and use calibration (inverse prediction, see Chapter 7 in Ramsey and Schafer [95]) to obtain an inverse prediction of σ corresponding to the actually observed value of *evar*. The resulting estimate of σ for the Arabidopsis data is 0.82, with 95% inverse prediction interval 0.75 to 0.88. (The calibration graph is available in the Web Supplementary Materials.)

2.6 Power and Robustness Evaluations

2.6.1 Overview

We intend to obtain the following set of diagnostic tools for many RNA-Seq datasets: the uniform QQ plot of individual GOF p-values, the single Fisher-combined GOF p-value, and the estimate of "NB2 dispersion noise", σ, for each of the seven NB dispersion models. That information will provide the basis for a more thorough investigation into robustness and power. As a start to this larger enterprise, we present simulations based on the Arabidopsis example conditions.

2.6.2 Simulation Specifications

For each of 96 sets of conditions, we simulated RNA-Seq read counts for 5,000 genes with total library size of read counts constant at two million, for each of six sampling units belonging to one of two treatment groups of size three. Four variable factors were the percentage of DE genes (at 5% or 20% with equal numbers up- and down-regulated), the fold change (1.2, 1.5, 2.0, or 3.0), the variance of the random $\mathcal{N}(0, \sigma^2)$ noise added to $\log(\phi)$ (with $\sigma = 0, 0.5, 1.0$, or 1.5), and the trend of NB2 dispersion as a function of mean relative frequency, π (with three trends on the log-log scale: linear [NBP] with parameters chosen to match those estimated from the Arabidopsis data, i.e. the dashed dark golden line in Figure 2.1; quadratic [NBQ], with parameters chosen to match those estimated from the Arabidopsis data, i.e. the dotted green curve in Figure 2.1; and "non-parametric" using the estimated "trended" ϕ's from the Arabidopsis data analysis, i.e. the dash-dotted dark cyan curve in Figure 2.1).

We illustrate the three dispersion trends in the Web Supplementary Materials and will just highlight the main findings in the next two sections. A primary emphasis is on characterizing the power advantages of the NBQ parametric approach if its model is in fact correct and the drop off in its power benefit when the quadratic trend isn't quite right or when there is extra variability in NB dispersion parameters about the trend.

2.6.3 Exact Quadratic Trend of $\log(\phi)$ on $\log(\pi)$

We focus on the exploration of the quadratic trend model, in which the NB response variance function is $\mu + \exp[\alpha_0 + \alpha_1 \log(\pi) + \alpha_2 \log(\pi)^2]\mu^2$, as suggested by the informal gamma log-linear analysis of the Arabidopsis data reported in Section 2.2.2 and the conformity of this model to the data based on the diagnostic analysis in Section 2.5.2. The gamma

analysis indicates the trend is quadratic; the diagnostic analysis indicates, additionally, that the data are consistent with this trend without any additional variability about the trend. The estimate of σ, however, is 0.82.

We consider the performance of the tests of differential gene expression based on the seven inferential methods. These simulation conditions match the model upon which the NBQ method is based, so this setting provides an opportunity to examine the power benefit from NBQ when its model is exactly right. We will highlight the particular conditions in which 20% of the genes are truly differentially expressed (half upward and half downward) and with all the fold changes fixed at 3.0 or 1/3.

In this two-sample setting, test performances may be assessed via the false and true positive identification rates for differential expression. Figure 2.5 shows the true positive rate (TPR, which is the same as recall and sensitivity) for all possible values of FDR for the seven methods in this setting. Note that the y-axis values are the actual – not nominal – FDRs, so that the x-axis indicates the power to correctly identify differentially expressed genes if the test is properly calibrated.

For the conditions shown in panel A, the actual FDR corresponding to a nominal FDR of 10% are as follows: genewise: 34.2%, common: 11.2%, NBP: 11.5%, NBQ: 7.3%, trended: 9.0%, tagwise-common: 12.0%, and tagwise-trend: 9.6%, with approximate standard errors of 1.8%. We use the p-value adjustment method discussed in Benjamini and Hochberg [11] to determine the nominal FDR. The plot shows the true detection probability for all possible choices of FDR. For example, in order to achieve a true positive rate (a hit rate) of 50% in this setting, a biologist must be willing to tolerate the following false discovery rates: genewise: 34%, common: 20%, NBP: 20%, tagwise-common: 17%, NBQ: 16%, tagwise-trend: 14%, and trended: 14%. The approximate standard errors of these estimates (corresponding to a true FDR of 0.2) are 1.6%. These simulations do not reveal

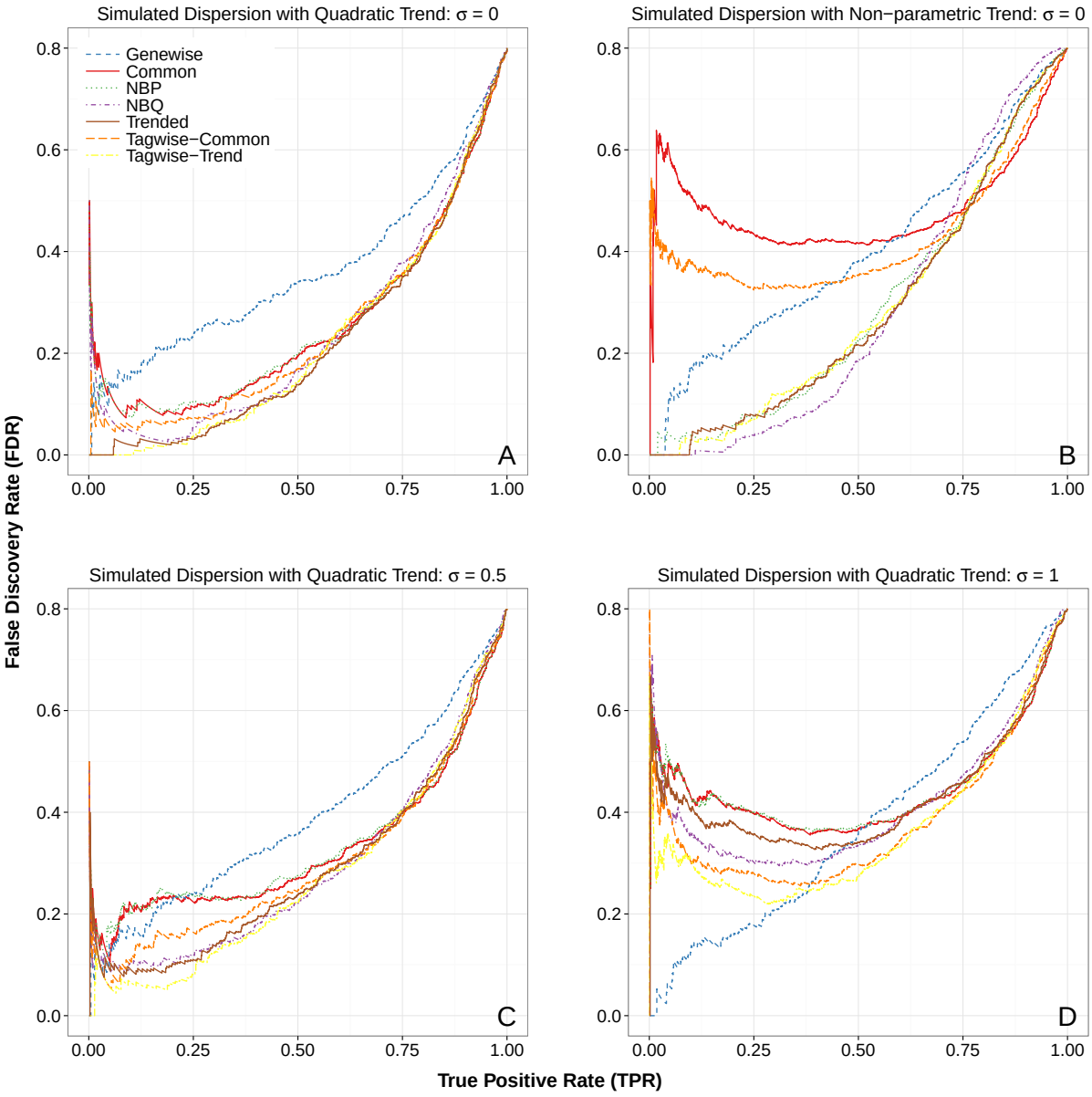

Figure 2.5: Evaluations of power and robustness of seven NB dispersion models. Among the 5,000 genes, we set 10% of DE genes with a fold change of 3.0 and 10% with a fold change of 1/3. The simulated dispersions follow either a quadratic trend (panels A, C and D) or a non-parametric trend (panel B). No noise was added to the simulated dispersions in panels A and B, and $\mathcal{N}(0, \sigma^2)$ noise was added in panel C (σ = 0.5) and in panel D (σ = 1.0).

power differences between NBQ and the two methods that use non-parametric regression to estimate trend, but the estimation of the trend reduces the necessary FDR (to achieve a true discovery rate of 50%) by about half over the genewise approach. The results based on smaller fold-changes and a smaller percentage of differentially expressed genes are less conclusive but show essentially the same patterns.

2.6.4 Robustness of NBQ Parametric Approach

We examine separately the robustness with respect to the form of the trend and the degree of variability of gene-wise ϕ's about the trend. The results for the simulation conditions in which the trend is taken to be the one estimated non-parametrically from the Arabidopsis data (the dash-dotted dark cyan curve in Figure 2.1) and with 20% of genes differentially expressed with fold change of 3.0 or 1/3, is shown in panel B. As in panel A, there is a clear improvement in power due to incorporating the trend, but a less clear distinction among the methods that incorporate the trend.

Panels C and D exhibit the results for the simulation conditions in which the trend is quadratic but normal noise is added to $\log(\phi)$ with standard deviations 0.5 and 1.0. For the smaller value, the trend methods still provide power benefits over the genewise approach. For the larger value, the results are more ambiguous.

2.7 Discussion

In this article, we proposed a simulation-based GOF test and associated graphical displays for assessing NB model adequacy for NB regression, and we showed a way to combine those tests from multiple genes or gene isoforms in RNA-Seq datasets. We believe the results may be useful for ordinary regression with count responses, but our concentration is on the

RNA-Seq setting.

We are interested in the potential power and efficiency gains in inferences from NB regression fits of individual genes when we adopt a global model that reduces the number of NB nuisance parameters. In this article we proposed methodology for judging such models. It is important to understand that there are two kinds of trend models relating the NB2 dispersion parameter to the mean. In one, represented by the NBP approach, the NB2 dispersion parameter is taken to be a simple function of the mean, so that the NB2 dispersion parameter will differ on the same gene for observations in different treatment groups if there are different expression levels in the different groups. For the trended approach and the related non-parametric approach in the `DESeq` package (Anders and Huber [3]), the NB2 dispersion parameter is taken to be constant for all observations on a single gene and that constant dispersion parameter is thought to be a smooth function of the average of means for that gene. It is not theoretically obvious whether the NB2 dispersion parameter should or should not be constant for a gene or, for that matter, whether the observed trend in dispersion parameter as a function of the mean is exact. We intend that the diagnostic analysis, performed on a variety of RNA-Seq datasets, will help provide an empirical clarification. The resolution of model adequacy is not, of course, the final piece of the puzzle. As in data analysis more generally, we do not expect models to fit exactly; we just need them to fit well enough for accurate and efficient inference. The diagnostic tools should help clarify models so that more comprehensive robustness and power studies can be used to compare the usefulness of the various inferential procedures upon which they are based.

The NBP model – in which the log of the NB2 dispersion parameter is a straight line function of the log of the mean – does not fully capture the trend in the RNA-Seq data we have examined. For that reason, we introduced the NBQ to allow for the next

simplest model. We see evident improvement in model fit to the Arabidopsis data when the quadratic term is included. Note that the NBQ model also avoids the need for user-specified tuning parameters (both for nonparametric regression and for empirical Bayes weighting of trend and individual components of dispersion).

Although the results for NBQ on the Arabidopsis data are intriguing, no strong generalizations about model adequacy emerge from the analysis of this single dataset or from the simulations based on the conditions of the dataset, which includes a very small sample size. We are currently applying the diagnostic tools to a variety of RNA-Seq studies on different organisms. In that regard, we believe a useful picture emerges from the following set of diagnostic tools: (a) A plot of estimated NB2 dispersion parameter estimates with various model fits (as in Figure 2.1). (b) The informal gamma log-linear regression analysis associated with that plot, including successive testing of polynomial terms and estimates of the proportion of variation in dispersion parameter estimates explained by polynomial models (as discussed in Section 2.2.2). (c) The NB GOF p-value from the fits to various models, such as the seven models reported in Section 2.5.2. (d) The estimate of σ in the noise model, in which the log of the NB2 dispersion parameter is the sum of a trend component (from NBQ trend or from trend estimated non-parametrically) and an individual component from a $\mathcal{N}(0, \sigma^2)$ distribution (as a measure of "noise" about the trend).

2.8 Supplementary Materials

The earthquake event dataset referenced in Section 2.3.1 and `R` codes for performing all simulations and real data analyses are available with this article at the Biometrics website on Wiley Online Library. The `R` package `NBGOF` (version 0.1.4) is available for download at `github.com/gu-mi/NBGOF`.

Acknowledgements

We thank the Co-Editor, the Associate Editor, and two referees for insightful comments that have substantially improved this article. We also thank Jeff H. Chang for preparing the Arabidopsis dataset, and Sarah C. Emerson for helpful discussions. This article is part of a doctoral dissertation written by the first author, under the supervision of the other two. Work of YD and GM is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM104977. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

3 Power-Robustness Analysis of Statistical Models for RNA Sequencing Data

Gu Mi and Yanming Di

PLOS ONE

Public Library of Science

1160 Battery Street

Koshland Building East, Suite 100

San Francisco, CA 94111, USA

(Current status: in preparation for submission)

3.1 Abstract

RNA-Sequencing (RNA-Seq) has been widely adopted for quantifying gene expression changes in comparative transcriptome analysis. For detecting differentially expressed genes, a variety of statistical methods based on the negative binomial (NB) distribution have been proposed. These methods differ in the ways they handle the NB nuisance parameters (i.e., the dispersion parameters associated with each gene) to save power, such as by using a dispersion model to exploit an apparent relationship between the dispersion parameter and the NB mean. Presumably, dispersion models with fewer parameters will result in greater power if the models are correct, but will produce misleading conclusions if not. This paper investigates this power and robustness trade-off by assessing rates of identifying true differential expression using the various methods under realistic assumptions about NB dispersion parameters. Our results indicate that the relative performances of the different methods are closely related to the level of dispersion variation unexplained by the dispersion model. We propose a simple statistic to quantify the level of residual dispersion variation from a fitted dispersion model and show that the magnitude of this statistic gives hints about whether and how much we can gain statistical power by a dispersion-modeling approach.

3.2 Introduction

Over the last ten years, RNA-Sequencing (RNA-Seq) has become the technology of choice for quantifying gene expression changes in comparative transcriptome analysis [121]. The negative binomial (NB) distribution has been widely used for modeling RNA-Seq read counts [102, 3, 31]. Although early studies have shown that the Poisson model was adequate for modeling RNA-Seq count variation from *technical* replicates [77], many recent RNA-Seq

analyses revealed that RNA-Seq counts from *biological* replicates show significant extra-Poisson variation. The NB distribution can be derived as a mixture of Poisson distributions in the so-called Gamma-Poisson model. For a random variable Y having an NB distribution with mean μ and dispersion ϕ, the variance is given by $\mathrm{Var}(Y) = \mu + \phi\mu^2$, and the dispersion parameter ϕ determines the extent to which the variance exceeds the mean. The square root of ϕ is also termed "biological coefficient of variation" (BCV) in [78].

The dispersion ϕ is a nuisance parameter in tests for differential expression (DE), but correct estimation of ϕ is essential for valid statistical inference. In a typical RNA-Seq experiment, our ability to detect truly DE genes is hampered by the large number of genes, the small sample size, and the need to estimate the dispersion parameters. To ameliorate this difficulty, many different NB dispersion models have been proposed (see the Background section for more details) with a common theme of "pooling information across genes". An NB dispersion model relates the dispersion to some measure of read abundance, a, through a simple parametric or smooth function f with a small number of parameters α (estimated from data):

$$\log(\phi_{ij}) = f(a_{ij}; \alpha), \tag{3.1}$$

where i indexes genes and j indexes biological samples. For example, in [31] we let a be preliminarily estimated mean relative frequencies and let f be a linear or quadratic function of $\log(a)$. This and other dispersion models are motivated by empirical evidence of a trend over all genes—of decreasing size of dispersion parameter with increasing relative frequency of RNA-Seq reads for the genes. By introducing a dispersion model f, one hopes to summarize the dispersion parameters for all genes by a small number of model parameters α and thus drastically reduce the number of nuisance parameters to estimate.

A dispersion-modeling approach as described above can lead to power saving, *if* a

correct or "close enough" model is used. It will be convenient for us to consider a general trend in dispersion parameter, but also allow for variation about the trend, as follows:

$$\log(\phi_{ij}) = f(a_{ij}; \alpha) + \epsilon_{ij}, \tag{3.2}$$

where ϵ represents an individual component in ϕ that is not explained by the trend. Intuitively, the strategy of "pooling information across genes" through a dispersion model f will be most effective if the overall level of residual variation in ϵ is low. In this paper, as an approximation, we model ϵ using a normal distribution $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and quantify the level of variation in ϵ by σ^2. We estimate σ for five real RNA-Seq datasets (from human, mouse, zebrafish, arabidopsis and fruit fly) and then investigate the power and robustness of DE tests when the amount of residual variation in dispersion matches that from the real data. We also explore how the relative performances of different DE test methods will change as the magnitude of σ changes.

3.3 Background

3.3.1 RNA-Seq

In brief, a typical RNA-Seq pipeline can be summarized as follows: purified RNA samples are converted to a library of cDNA with attached adaptors, and then sequenced on an HTS platform to produce millions of short sequences from one or both ends of the cDNA fragments. These reads are aligned to either a reference genome or transcriptome (called sequence mapping), or assembled *de novo* without the genomic sequence. The aligned reads are then summarized by counting the number of reads mapped to the genomic feature of interests (e.g., exons or genes), and the expression profile is eventually represented

by a matrix of read counts (non-negative integers) where rows are genes (or some other genomic features like exons) and columns are samples. Subsequent steps that rely heavily on statistical analyses include normalization of reads and testing DE genes between samples under different environmental or experimental conditions.

3.3.2 NB Regression Models

An NB regression model for describing the mean expression as a function of explanatory variables include the following two components:

1. An NB distribution for the individual RNA-Seq read counts Y_{ij}:

$$Y_{ij} \sim \mathrm{NB}(\mu_{ij}, \phi_{ij}),$$

   where $i = 1, \ldots, m$ indexes genes, $j = 1, \ldots, n$ indexes samples, μ_{ij} is the mean, and ϕ_{ij} is the dispersion parameter such that $\mathrm{Var}(Y_{ij}) = \mu_{ij} + \phi_{ij}\mu_{ij}^2$.

2. A log-linear regression model for the mean μ_{ij} as a function of p explanatory variables X_{jk} $(k = 1, \ldots, p)$:

$$\log(\mu_{ij}) = \log(N_j) + \log(R_j) + \sum_{k=1}^{p} \beta_{ik} X_{jk}. \tag{3.3}$$

These two components resemble a generalized linear model (GLM) [85], but note that the dispersion ϕ_{ij} is unknown (see the next subsection for different dispersion models). The two additive constants, $\log(N_j)$ and $\log(R_j)$, have to do with count normalization: accounting for different observed library sizes (N_j) and the apparent reduction/increase in the expression levels of non-DE genes resulting from the increased/decreased expression of a few truly DE genes [3, 99]. The normalization constants, N_j and R_j, are pre-estimated

and treated as known during GLM fitting. In many applications, the same constant ($N_j R_j$) is assumed for all genes in a sample, but it may be advantageous to introduce between-gene normalization factors to account for some gene-specific sources of technical biases such as GC-content and gene length [75]. Between-gene normalization can be incorporated into the GLM framework as well. See [51, 97, 98] for relevant discussions.

3.3.3 DE Tests

Testing differential expression can often be reduced to testing that one or more of the regression coefficients equals zero. For example, for comparing gene expression levels between two groups, we can let $p = 2$, $X_{j1} = 1$ for all j; $X_{j2} = 1$ if sample j is from group 2 and $X_{j2} = 0$ if sample j is from group 1. Under this parameterization, β_1 corresponds to group 1's relative mean expression level and β_2 corresponds to the log fold change between group 2 and group 1. The null hypothesis is $H_0 : \beta_2 = 0$.

In general NB regression settings, exact tests are not available, but asymptotic tests, such as likelihood ratio test, can be used. Di *et al.* [33, 30] showed that the performance of likelihood ratio test in small sample settings can be improved with higher-order asymptotics (HOA) adjustment. Lund *et al.* [76] discussed quasi-likelihood (QL) methods by replacing likelihood ratio test with QL ratio F-test for better FDR control, where the test statistic is based on quasi-dispersion parameter estimates or two variants called QLShrink and QLSpline for pooling information across genes.

3.3.4 NB Dispersion Models

As mentioned in the Introduction section, many current DE analysis methods use an NB dispersion model to capture the general trend between dispersion and read abundance. The

different DE analysis methods can be put into the following general categories according to the functional form f of the dispersion model and the treatment of individual variation (see Equation (3.2)):

1. Common: Earlier works of Robinson and Smyth [100, 101] assumed a common dispersion model where f is a constant. In other words, $\phi_{ij} = c$ for all i, j.

2. Parametric function: Recognizing an evident trend between the dispersion and relative gene expression, Di *et al.* [31] adopted a parametric NBP model where the log dispersions are modeled as a linear function of the log relative mean frequencies. Referring to Equation (3.1), in an NBP model, $a_{ij} = \pi_{ij} = \frac{\mu_{ij}}{N_j R_j}$ and $f(a_{ij}; \alpha) = \alpha_0 + \alpha_1 \log(\pi_{ij})$. A natural extension to NBP is the NBQ model which incorporates an extra quadratic term:

$$f(a_{ij}; \alpha) = \alpha_0 + \alpha_1 \log(\pi_{ij}) + \alpha_2 \left[\log(\pi_{ij})\right]^2 . \tag{3.4}$$

3. Smooth function: Anders and Huber [3] suggested fitting a non-parametric curve to capture the dispersion-mean dependence. McCarthy *et al.* [78] introduced a similar "trended" (non-parametric) model. `NBPSeq` added an NBS model for non-parametric smooth dispersion model.

The methods above ignore possible individual dispersion variation (i.e., ϵ_{ij} in Equation (3.2)) in subsequent DE tests.

4. Shrinkage methods: McCarthy *et al.* [78] discussed options to use weighted average between genewise dispersion estimates and trended estimates in an empirical Bayes framework (we will call this method "tagwise-trend"). The genewise estimates can also be shrunk towards a common value [100, 101]. Love *et al.* [75] added a shrinkage option in `DESeq2` similar to that implemented in `edgeR`.

5. Quasi-likelihood methods: Lund *et al.* [76] suggested fitting a quasi-likelihood (QL) model by specifying (for gene i and sample j):

$$\mathrm{Var}(Y_{ij}) = \Phi_i V_i(\mu_{ij}), \tag{3.5}$$

   with the (negative binomial) variance function $V_i(\mu_{ij}) = \mu_{ij} + \omega_i \mu_{ij}^2$. Both the NB dispersion parameter (ω_i) and the quasi-likelihood dispersion parameter (Φ_i) are estimated from the data and used to model the variance of the read count Y_{ij}. The advantage of accounting for uncertainty in the modeled variance is brought by the QL-dispersion Φ_i, which is absent in existing NB-based methods. Due to the small number of replicates in RNA-Seq data, the QL approach can be improved by pooling information across genes. This result in two variants of the QL dispersion, "QL-Shrink" and "QLSpline", that differ in the formulation of a scaling factor (either estimated from the distribution of $\hat{\Phi}_i$ or a scale-adjusted spline-based estimate of Φ_i), to which the QL dispersion is shrunken. These QL-based approaches are implemented in the `QuasiSeq` package.

6. Genewise: `NBPSeq` allows for fitting NB regression model and performing DE test to each gene separately without assuming any dispersion model.

In the above, we mainly summarized methods implemented in the R/Bioconductor packages `DESeq`, `DESeq2`, `edgeR`, `NBPSeq` and `QuasiSeq` [94, 40]. They represent the wide range of currently available options. These packages use slightly different predictors (a_{ij} in Equation (3.1)) in their dispersion models, and also use different methods to estimate dispersion models, but these differences are of no primary interest in our power-robustness analysis. As we will see later, the main factor that influences the DE test performance is how the individual dispersion variation is handled.

3.3.5 Other Related Work

Our approach for estimating σ^2 is related to the empirical Bayes approach for estimating ϵ under a normal prior distribution, but our focus here is in estimating σ^2, not the individual ϵ_i's. The issue of dispersion model adequacy has been raised before: the program `edgeR` [102] used an empirical Bayes approach to combine the dispersion estimates from a fitted model with genewise estimates. Lund *et al.* [76] proposed to incorporate the residual dispersion variation using a quasi-likelihood method. Wu *et al.* [124] proposed another empirical Bayes shrinkage estimator for the dispersion parameter which aims to adequately capture the heterogeneity in dispersion among genes. Mi *et al.* [82] proposed goodness-of-fit tests for dispersion models. There are also recent works on comparing the performances of DE tests: Soneson and Delorenzi [110] evaluated 11 tools for their ability to rank truly DE genes ahead of non-DE genes, the Type I error rate and false discovery rate (FDR) controls, and computational times. Landau and Liu [59] discussed dispersion estimation and its impact on DE test performance, mainly focusing on different shrinkage strategies (none, common, tagwise or maximum). Our new contribution in this paper is to explicitly quantify the level of inadequacy of a fitted dispersion model using a simple statistic, and link the magnitude of this statistic directly to the performance of the associated DE test.

3.4 Results

We investigate the power and robustness of DE tests under realistic assumptions about the NB dispersion parameters. We fit the NBQ dispersion model (see Equation (3.4)) to real datasets to capture the general trend in the dispersion-mean dependence. We model the residual variation in dispersion using a normal distribution (see Equation (3.2)) and the level of residual variation is then summarized by a simple quantity, the normal variance

σ^2. Because biological variations are likely to differ across species, and experiments involve varied sources of uncertainty, we choose to analyze five datasets from different species that represent a broad range of characteristics and diversity for typical RNA-Seq experiments. The species include human (*Homo sapiens*), mouse (*Mus musculus*), zebrafish (*Danio rerio*), arabidopsis (*Arabidopsis thaliana*) and fruit fly (*Drosophila melanogaster*). The Methods section includes descriptions of the datasets. For each experiment/dataset, unless otherwise specified we will provide the following results:

1. Mean-dispersion plot with trends estimated from NB dispersion models;
2. Gamma log-linear regression as informal model checking;
3. Estimation of the variance σ^2 of dispersion residuals from a fitted dispersion model;
4. Power-robustness evaluations of DE tests using datasets simulated to mimic real datasets.

The main focus of this paper is on the quantification of the level of residual dispersion variation and power-robustness investigation under realistic settings (3 and 4 above). The diagnostic plots and statistics (1 and 2 above) are useful in routine analysis of RNA-Seq data, and they also help us verify that the NBQ dispersion model largely captures the general trend in the dispersion-mean dependence.

Anders *et al.* [5] suggested removing genes without one read per million (rpm) in at least n of the samples, where n is the size of the smallest group of replicates. We follow a similar criterion but set $n = 1$ in order to keep more (lowly-expressed) genes in study. In R, this is achieved by subsetting the row indices by `rowSums(cpm(data)>1)>=1`. The library size adjustments are computed for genes passing this criterion.

3.4.1 Mean-Dispersion Plots with Estimated Trends from Dispersion Models

Figure 3.1 shows the mean-dispersion plots for the two treatment groups in the human dataset (with sequencing depth of 30 million). In each plot, method-of-moment (MOM) estimates ($\hat{\phi}^{\text{MOM}}$) of the dispersion ϕ for each gene are plotted against estimated relative mean frequencies (on the log-log scales). We also overlaid the trends from five fitted dispersion models representing the wide range of currently available options: common, NBP, NBQ, NBS and trended. We make the following remarks:

#1 The fitted NBP, NBQ, NBS and trended dispersion models all capture the overall decreasing trend in the MOM genewise estimates.

#2 The fitted models agree more in the mid-section of the expression distribution and less in the tails where genes have extremely low or high expression levels. This kind of behavior is common in non-parametric smooth estimates and regression models, and it has some implications on how we design the power simulations later.

#3 Such mean-dispersion plots are informative in checking how different dispersion models may potentially over-/under-estimate the dispersion parameters, which in turn will influence DE test results.

#4 Note that the deviation of the genewise MOM estimates from the fitted dispersion models is *not* the same as the ϵ in Equation (3.2), since this deviation also reflects the additional estimation error due to small sample size.

Mean-dispersion plots for the other four datasets show similar features and are included in Supporting Information S1.

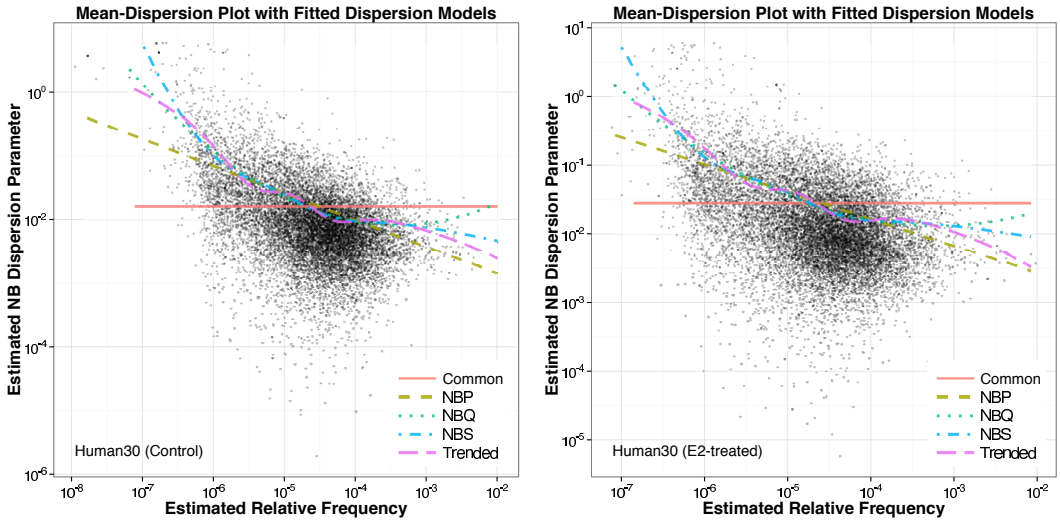

Figure 3.1: Mean-Dispersion Plot of the Human30 RNA-Seq Dataset. The sequencing depth for this dataset is 30 million. The control (E2-treated) group with seven biological replicates is shown on the left (right) panel. Each point on the plots represents one gene with its method-of-moment (MOM) dispersion estimate ($\hat{\phi}^{\text{MOM}}$) on the y-axis and estimated relative mean frequency on the x-axis. The fitted curves for five dispersion models are superimposed on the scatter plot.

3.4.2 Gamma Log-Linear Regression Analysis

As informal model checking, we fit polynomial gamma log-linear regression models of $\hat{\phi}^{\text{MOM}}$ on $\log(\hat{\pi})$. Table 3.1 summarizes the variability in the logged genewise dispersion estimates $\log(\hat{\phi}^{\text{MOM}})$ explained by the linear, quadratic and cubic models (results shown for the control group only and without pre-filtering lowly-expressed genes). The proportion of variation in $\log(\hat{\pi})$ explained by the fitted models varies across species (e.g., for the quadratic fit, it ranges from 31% to 75%) and also depends on sequencing depths. The quadratic regression model improves over the simple linear regression model by explaining an additional 2% to 11% of variation, while adding a cubic term has almost negligible effects.

Table 3.1: Polynomial gamma log-linear regression models of $\hat{\phi}$ on $\log(\hat{\pi})$ (results shown for the control group only).

	Dataset					
Model	Human5	Human30	Mouse	Zebrafish	Arabidopsis	Fruit Fly
Linear	73.09%	72.29%	49.20%	32.08%	36.30%	23.79%
Quadratic	75.15%	74.38%	54.85%	43.02%	41.02%	31.20%
Cubic	75.46%	74.45%	54.55%	43.77%	41.01%	32.74%

3.4.3 Quantification of the Level of Residual Dispersion Variation

As discussed in the Introduction section, we model the dispersion residuals using a normal distribution, $\epsilon = \log(\phi) - \log(\hat{\phi}) \sim \mathcal{N}(0, \sigma^2)$, and thus quantify the level of residual variation using σ^2 or equivalently σ. Using the approach described in the Methods section, we estimated σ from each of the real datasets analyzed after fitting an NBQ dispersion model (see Equation (3.4)). Table 3.2 summarizes the estimates and the corresponding standard errors. The magnitudes of $\hat{\sigma}$ indicate that the fitted dispersion models did not fully explain the total variation in the dispersion. The NBQ dispersion model uses estimated mean relative frequencies ($\hat{\pi}_{ij}$) as predictors, and the results here suggest that there is still substantial individual variation among genes with the same values of $\hat{\pi}_{ij}$.

It is possible to turn the estimate $\hat{\sigma}$ into a goodness-of-fit test for the fitted dispersion model. However, we want to ask whether a dispersion model is useful even when the fitted model shows lack-of-fit. For this purpose, the quantitative measure $\hat{\sigma}$ is more intuitive than a test p-value, since it directly reflects the degree of deviation from the fitted dispersion model. In the next section, we will explore the connection between the magnitude of $\hat{\sigma}$ and the performance of DE tests in terms of power and FDR.

Table 3.2: Level of residual dispersion variation in five real RNA-Seq datasets. The columns are the name of the dataset, the number of samples (control, treatment), the maximum likelihood estimate (MLE) $\hat{\sigma}$, and the standard error (SE) of $\hat{\sigma}$.

Dataset	**#samples**	**MLE** $\hat{\sigma}$	**SE**($\hat{\sigma}$)
Human30	(7, 7)	1.021	0.028
Mouse	(3, 3)	1.228	0.055
Zebrafish	(4, 4)	1.105	0.043
Arabidopsis	(3, 3)	0.956	0.040
Fruit fly	(4, 3)	1.015	0.032

3.4.4 Power-Robustness Evaluations

We compare the power and FDR control of a range of DE test methods on datasets simulated to mimic the five real datasets.

3.4.4.1 Simulation Setup

In our power-robustness analysis, we will choose one representative method from each of the categories summarized in the "Background/NB Dispersion Models" subsection (with R/Bioconductor package names in parentheses): genewise-HOA (`NBPSeq`), common (`edgeR`), NBQ (`NBPSeq`), trended (`edgeR`), tagwise-trend (`edgeR`) and QLSpline (`QuasiSeq`). The common dispersion model is included solely for benchmark purpose as it is over-simplified and not recommended for practical use. The NBQ method represents parametric dispersion models and it generally provides better model fit than the simpler NBP model [82]. The tagwise-trend method represents the empirical Bayes shrinkage methods. The QLSpline method represents quasi-likelihood methods [76]. For testing DE, methods from `edgeR` use likelihood ratio test, methods from `NBPSeq` use likelihood ratio test with HOA adjustment,

and the QLSpline method uses QL ratio F-test.

We simulate two-group comparison datasets that mimic the five real RNA-Seq datasets. From each real dataset, we randomly select $5,000$ genes and fit NB regression models to them (see Equation (3.3) and the "Background/DE Tests" subsection). We generate a new dataset of $5,000$ genes based on fitted models. We specify the mean expression levels based on estimated $\hat{\beta}_{ik}$, with $R_j = 1$ and N_j reflecting the sequencing depth (e.g., $N_j = 2.5 \times 10^7$ for the human dataset and 1.5×10^7 for the mouse dataset). For all genes, we set β_{i1} as the estimated value from the real data. If gene i is designated as DE, we use $\hat{\beta}_{i2}$ estimated from the real data as its log fold change (i.e., we set $\beta_{i2} = \hat{\beta}_{i2}$). For any non-DE gene i', we set $\beta_{i'2} = 0$. In real data analysis, it is unknown which genes are DE. For each dataset, we randomly designate m_1 genes as DE. We consider two levels, 0.1 and 0.2, for the percentage of DE genes ($\pi_1 = m_1/m$). Approximately half of the simulated DE genes are over-expressed and half are under-expressed. Early microarray studies had shown that a smaller proportion of DE genes tend to make it more difficult to control FDR at the nominal level [72].

We specify the dispersion parameters according to Equation (3.2) with the trend part, $f(a_{ij}; \alpha)$, being the fitted NBQ model (fitting Equation (3.4) to real data). The deviation from the trend is controlled by ϵ_i and will be simulated according to a $\mathcal{N}(0, \sigma^2)$ distribution. We want to choose σ^2 to match the real data, but there is some subtlety in how to achieve this: in practice, when fitting the NBQ model, we use the fitted values $\hat{\pi}_{ij}$ as the predictors since true π_{ij} values are not available, but when we simulate counts, the $\hat{\pi}_{ij}$ values are not available. Our solution is to use π_{ij} as predictor in the NBQ model when simulating ϵ, but choose $\sigma = \tilde{\sigma}$ through a *calibration* approach such that if we were to fit the NBQ model to the simulated data later—using the estimated $\hat{\pi}_{ij}$ as predictor, the estimated $\hat{\sigma}$ would match the one estimated from the real data (also using the estimated $\hat{\pi}_{ij}$ as predictor). The

estimated values of $\hat{\sigma}$ from real datasets are summarize in Table 3.2. The calibrated values $\tilde{\sigma}$ and the details about the calibration approach are presented in the Methods section. In our simulations, we will consider different levels of residual dispersion variation and set σ to $\tilde{\sigma}$, $0.5\tilde{\sigma}$ or 0.

There are other factors that may potentially contribute to the difference in DE test performance, such as the presence of outliers, the proportion of up and down-regulated genes, potential correlation between gene expression levels, to just name a few. In this paper, we will focus on the impact of umodeled dispersion variation on DE test performance.

3.4.4.2 Power Evaluation

For power evaluation, we plot true positive rates (TPR) versus false discovery rates (FDR). For a DE test, a true positive (TP) indicates the test correctly identifies a DE gene; a false positive (FP) indicates the test incorrectly identifies a non-DE gene as DE; and a false negative (FN) indicates the test incorrectly declares a DE gene as non-DE. The TPR and FDR are defined as: $\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ and $\mathrm{FDR} = \mathrm{FP}/(\mathrm{TP}+\mathrm{FP})$. A TPR-FDR curve contains equivalent information as a precision-recall curve, but focuses on the relationship between TPR (power) and FDR. The power of a DE test depends on alternative hypothesis and will likely vary between genes. TPR reflects the average power of a test to detect truly DE genes in the simulated dataset. If we compare the TPR of the tests at the same FDR level, we are essentially comparing the size-corrected power.

The left column of Figure 3.2 shows the TPR-FDR plots for the six tests performed on each of the five datasets simulated to mimic the five real datasets. In particular, the simulated datasets have the same level of residual dispersion variation σ^2 as estimated from the five real datasets. A better method will have its TPR-FDR curve closer to the lower-

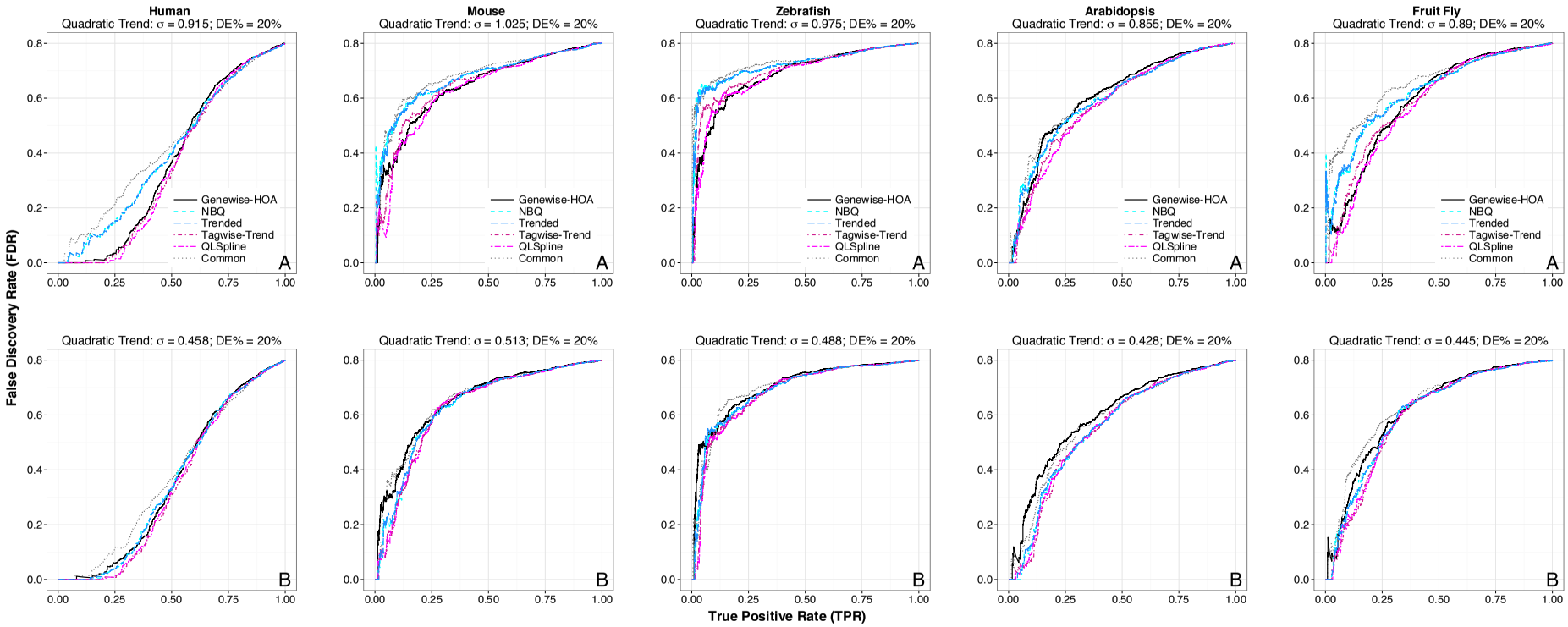

Figure 3.2: True Positive Rate (TPR) vs. False Discovery Rate (FDR). The x-axis is the TPR (which is the same as recall and sensitivity) and the y-axis is the FDR (which is the same as one minus precision). The percentage of DE is specified at 20% in all scenarios. We specify σ at the estimated value ($\tilde{\sigma}$) in panels labeled with A (first row), and half the estimated value ($0.5\tilde{\sigma}$) in panels labeled with B (second row). Each column shows the results for the following datasets (left to right): human, mouse, zebrafish, arabidopsis, and fruit fly. The FDR values are highly variable when TPR is close to 0, since the denominator TP + FP is close to 0. When comparing the performance, we suggest inspecting regions with TPR not close to 0.

right corner, indicating a lower FDR for achieving a fixed power, or a higher power for a fixed tolerable FDR. For four of the datasets, the QLSpline, tagwise-trend and genewise methods outperform the NBQ, trended and common dispersion models, with the common model being the worst. For the simulation dataset based on the Arabidopsis real dataset, no test is dominate at all FDR levels.

It is somewhat surprising that the performance of the simple genewise method is comparable to the best methods in all cases. This indicates that if the level of residual dispersion variation is as high as the estimated (see Table 3.2), the potential power saving through dispersion modeling is quite limited.

The relative performance of the tests will change if the level of residual dispersion variation (σ^2) changes. The right column of Figure 3.2 shows the TPR-FDR plots when σ is simulated to be half the estimated values ($\sigma = 0.5\tilde{\sigma}$). The performance of the NBQ and trended methods has much improved and is better than the genewise method in three of the datasets (the ones based on mouse, zebrafish and arabidopsis). When we further reduced σ to 0 in our simulations, all methods outperformed the genewise approach. The QLSpline and tagwise-trend methods managed to perform consistently well as we vary the magnitude of σ.

3.4.4.3 FDR and Type-I Error

In practice, the Benjemini-Hochberg method [11] is commonly used to control the FDR of DE tests. In Table 3.3, we compare the *actual* FDR of the different DE tests based on the simulation results when the *nominal* FDR is set to 10% using the Benjemini-Hochberg method. The results were based on the datasets simulated to mimic the human dataset, where we varied the percentage of DE genes (10% and 20%) and we varied σ from estimated

value ($\sigma = \tilde{\sigma}$), to half the estimated value ($\sigma = 0.5\tilde{\sigma}$), and then to 0. The QLSpline and genewise methods have good controls on FDR in all cases, and are conservative in some cases. The FDRs from the tagwise-trend method are below the nominal level when the percentage of DE is 20%, but are above the nominal levels when the percentage of DE is 10%. For the NBQ and trended methods, the FDR control improves as the residual dispersion variation decreases and as the percentage of truly DE genes increases. The common method does not have good control of FDR in all scenarios.

Table 3.3: Actual FDR for a nominal FDR of 0.1. The best results are highlighted in bold, and the second best results are highlighted with underlines. We consider three levels of σ: at the estimated value ($\sigma = \tilde{\sigma}$), half the estimated value ($\sigma = 0.5\tilde{\sigma}$), and no variation ($\sigma = 0$).

		Actual FDR for 10% Nominal FDR					
Variation	%DE	Genewise	QLSpline	Tagwise-Trend	NBQ	Trended	Common
$\tilde{\sigma}$	10%	**9.72%**	<u>7.84%</u>	12.5%	28.3%	29.3%	39.6%
	20%	**7.90%**	6.57%	<u>6.76%</u>	17.3%	18.2%	25.3%
$0.5\tilde{\sigma}$	10%	<u>11.8%</u>	**10.5%**	14.2%	13.3%	14.4%	38.3%
	20%	**10.7%**	8.03%	<u>11.5%</u>	13.4%	14.1%	23.9%
0	10%	11.2%	**10.2%**	16.2%	<u>10.6%</u>	11.9%	26.6%
	20%	7.19%	5.61%	**9.88%**	6.86%	<u>7.46%</u>	17.4%

The FDR control is closely related to the test p-values. Figure 3.3 shows the histograms of p-values computed for the non-DE genes in one of the datasets used for the FDR comparison above (20% DE and $\sigma = \tilde{\sigma}$). The histograms from the genewise and QLSpline methods are more close to uniform. For the common, NBQ and trended methods, the histograms are asymmetric v-shaped: there is an overabundance of small p-values as compared to a uniform distribution, but the histograms also indicate that these tests are conservative for many genes. Similar patterns have been observed for other dispersion-modeling methods

by Lund *et al.* in [76]. The tagwise-trend method produces conservative p-values.

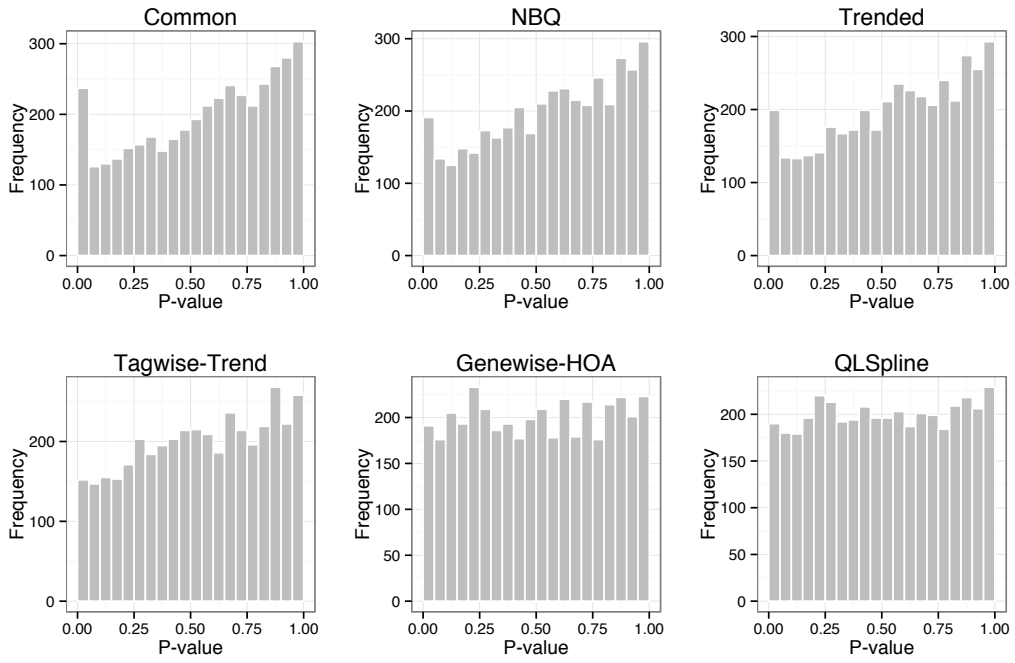

Figure 3.3: P-value histograms of non-DE genes from six dispersion methods. The dispersions are simulated with residual dispersion variation estimated from the human dataset ($\sigma = \tilde{\sigma}$). Out of a total of 5,000 genes, 80% are non-DE.

Figure 3.4 shows similar histogram comparisons when σ was reduced to half the estimated value ($0.5\tilde{\sigma}$). The null p-value histograms from the NBQ and trended methods have improved and are closer to the uniform distribution. The tagwise-trend method produces a slight overabundance of small p-values. The common method is still unsatisfactory.

3.5 Conclusion and Discussion

We quantified the residual dispersion variation in five real RNA-Seq datasets. Using simulations, we compared the performance—in terms of power and FDR/Type-I error control—of

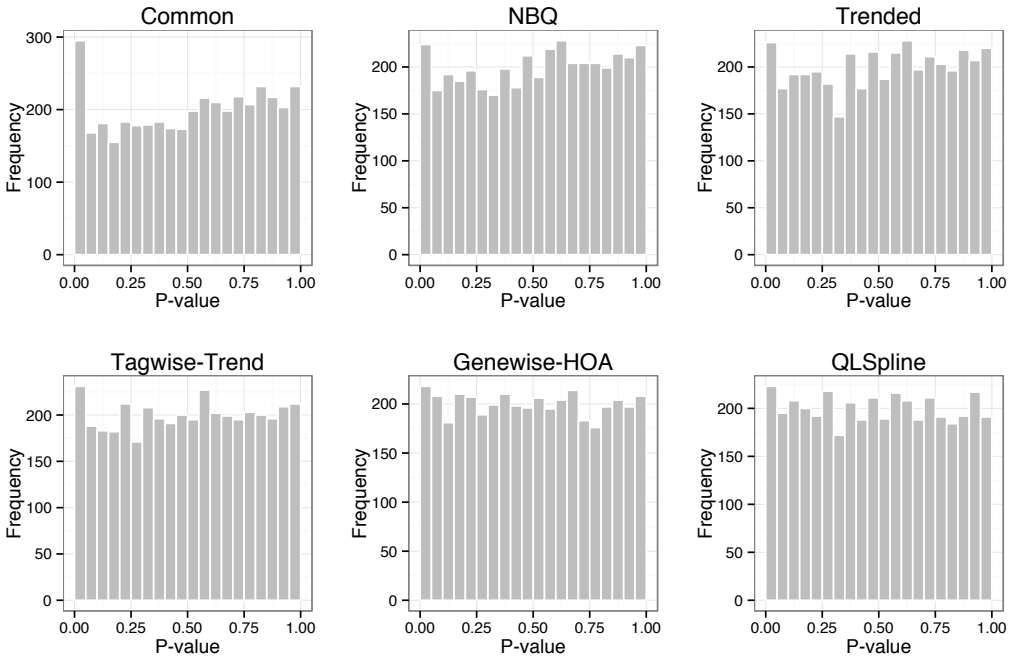

Figure 3.4: P-value histograms of non-DE genes from six dispersion methods. The dispersions are simulated with half the value of the residual dispersion variation estimated from the human dataset ($\sigma = 0.5\tilde{\sigma}$). Out of a total of 5,000 genes, 80% are non-DE.

six representative DE tests based on different dispersion methods. We demonstrated that the level of residual dispersion variation is a crucial factor in determining the performance of DE tests. When the residual dispersion variation is as high as we estimated from the five real datasets, methods such as NBQ and trended, which ignore possible residual dispersion variation, fail to control Type-I errors and give suboptimal power. The QLSpline and tagwise-trend methods have similar size-corrected power, but the tagwise-trend method gives conservative FDR. QLSpline and tagwise-trend both account for individual dispersion variation, but QLSpline also makes degrees-of-freedom adjustment to address the uncertainty in estimated NB dispersions. For other DE test methods that use a dispersion model, we recommend using degrees-of-freedom adjustment to improve robustness.

The genewise method does not rely on a dispersion model, and it uses an HOA technique to improve small-sample performance of the likelihood ratio test. The genewise method has good Type-I error and FDR control in all simulations. The power of the genewise method is comparable to that of the QLSpline and tagwise-trend. This indicates that when the level of dispersion variation is high, the power saving available through dispersion modeling is limited.

Reducing the level of dispersion variation boosts the performance of DE tests that use a dispersion model. One may attempt to improve the dispersion model by considering different functional forms of the trend and/or including additional predictors. We plan to explore such possibilities in our future research. It is not well understood what factors contribute to the count and dispersion variation in an RNA-Seq experiment: possible factors to consider include transcript length, GC-content, and so on.

One notable difference between the genewise method and a dispersion-modeling method is that the former detects more DE genes with small fold changes, while a method using a dispersion model tends to detect DE genes with large fold changes. Figure 3.5 illustrates

this point using MA plots. This is because current dispersion models often assume the dispersion is the same for genes with similar mean levels (those genes having the same x-values). Under such assumptions, large fold changes tend to correspond to more significant test results. The behaviors of the tagwise-trend and the QLSpline methods are intermediate between the genewise method and a dispersion-modeling method such as the trended model.

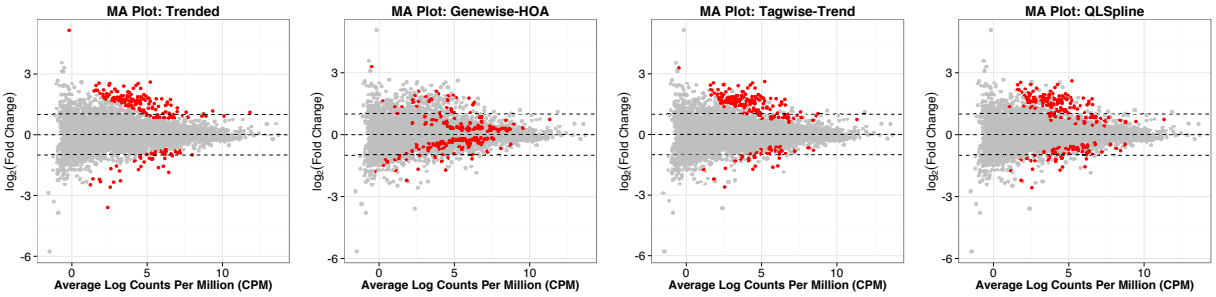

Figure 3.5: MA plot of the mouse dataset for the trended, genewise, tagwise-trend and QLSpline approaches. Predictive log fold changes (posterior Bayesian estimators of the true log fold changes) for NB GLMs are calculated (the "M" values) and shown on the y-axis. Averages of log counts per million (CPM) are shown on the x-axis (the "A" values).

We used a $\mathcal{N}(0, \sigma^2)$ distribution to model the residual dispersion variation ϵ_i (see Equation (3.2)). We believe this is a reasonable starting point. A similar assumption is made in [124] and the authors presented simple diagnostic plots to show the normality assumption is reasonable. In future, we will explore more formal methods for model checking and consider the possibility that σ may vary with some other variables, such as the mean level. However, the conclusion that the performance of the DE tests depend on the level of the residual dispersion variation should remain valid.

3.6 Methods

3.6.1 Description of RNA-Seq Datasets

Experiment information for all species and the raw/processed data are available at the Gene Expression Omnibus (GEO) of the National Center for Biotechnology Information (NCBI). Table 3.4 gives a brief summary of the datasets analyzed in this paper, including the dataset names in the `SeqDisp` R package we develop (see the Software Information section), the SRA accessions that provides all the metadata describing a particular study (see the NCBI website for different accession types), and published references. See Supporting Information S1, "Access to the Datasets" section for more details.

Table 3.4: Summary of RNA-Seq datasets analyzed in this article.

Organism	**Name in `SeqDisp`**	**SRA Accession**	**References**
Homo Sapiens	`human30/human5`	SRP031476	[74]
Mus Musculus	`mouse`	SRP022850	[17]
Danio Rerio	`zebrafish`	SRP017511	[118]
Arabidopsis Thaliana	`arabidopsis`	SRP013873	[55]
Drosophila Melanogaster	`fruit.fly`	SRP001537	[16, 56]

3.6.1.1 Human RNA-Seq Data

The *Homo sapiens* (human) RNA-Seq experiment was discussed in [74], and information for this experiment, the raw and processed data are available at NCBI GEO under accession number GSE51403. Liu *et al.* [74] focused more on the technical side of RNA-Seq experiments by investigating the trade-offs between sequencing depth (where a higher depth generates more informational reads) and biological replications. Seven biological replicates

of both control and E2-treated MCF7 cells were sequenced, and the RNA-Seq reads in each sample were down-sampled to generate datasets of different depths (a total of seven depths from 2.5M to 30M). We include datasets from two sequencing depths (5M and 30M) in our R package, but mainly focus on the dataset with 30M sequencing depth for analyses. See [74] and NCBI GSE51403 for detailed descriptions of the dataset.

3.6.1.2 Mouse RNA-Seq Data

The *Mus musculus* (mouse) RNA-Seq experiment was discussed in [17], and information for this experiment and the raw data are available at NCBI GEO under accession number GSE47019. The raw data are downloaded from NCBI Sequence Read Archive (SRA), and processed using the pipeline described in [5]. According to the experiment summary, the goals of this study were "to compare signals from competent and abnormal human embryos impacted differently on the expression of endometrial receptivity genes in mouse uteri. Uterine mRNA profiles of 21-day old mice were generated by deep sequencing, in triplicate." We summarized the samples of "Control Salker", "Developmentally competent embryo conditioned media Salker" (abbreviated as DCECM) and "Arrested embryo conditioned media Salker" (abbreviated as AECM) into the `mouse` dataset in the `SeqDisp` R package. We only considered the control and DCECM groups in the analyses.

3.6.1.3 Zebrafish RNA-Seq Data

The *Danio rerio* (zebrafish) RNA-Seq experiment was discussed in [118], and information for this experiment and the raw data are available at NCBI GEO under accession number GSE42846. According to the experiment summary, the authors "used the zebrafish embryo model to study the innate immune response against *Staphylococcus epidermidis*, by

injecting *S. epidermidis* into the yolk at 2 hpf and took samples at 5 days post injection." Four biological replicates are prepared for the control group (Non-injected 5 DPI) and for the treatment group (S. epi mcherry O-47 5 DPI).

3.6.1.4 Arabidopsis RNA-Seq Data

The *Arabidopsis thaliana* (arabidopsis) RNA-Seq experiment was discussed in [55], and information for this experiment and the raw data are available at NCBI GEO under accession number GSE38879. According to the experiment summary, "the goal of this study is to identify the targets of RVE8, a MYB-like transcription factor involved in the circadian clock in Arabidopsis, and study reveals that RVE8 is a master regulator of circadian gene expression in Arabidopsis." The overall design includes transgenic line rve8-1 RVE8::RVE8:GR and rve8-1 treated with DEX or mock with three biological replicates each, for a total of 12 samples. Our analyses only focused on the RVE8:GR_mock control group, and the RVE8:GR_DEX treatment group.

3.6.1.5 Fruit Fly RNA-Seq Data

The *Drosophila melanogaster* (fruit fly) RNA-Seq experiment was discussed in [16], and information for this experiment and the raw data are available at NCBI GEO under accession numbers GSM461176 to GSM461181. The dataset `fruit.fly` in our `SeqDisp` package is directly obtained from the `pasilla` Bioconductor package [56], which provides per-exon and per-gene read counts computed for selected genes in [16]. It can also be accessed from `data(pasillaGenes)` once `pasilla` is loaded. The dataset contains three and four biological replicates of the knockdown and the untreated control, respectively. See the package vignette for more information.

3.6.2 Methodological Details for Quantifying Dispersion Residual Variation

In the RNA-Seq context, we use Y_{ij} to denote the read count for gene i in sample j, where $i = 1, \cdots, m$ and $j = 1, \cdots, n$. We model a single read count as negative binomial with mean μ_{ij} and dispersion ϕ_{ij}:

$$Y_{ij} \sim \text{NB}(\mu_{ij}, \phi_{ij}).$$

We further assume a log-linear model for μ_{ij}, i.e. $\log(\mu_{ij}) = \text{offset} + X'\beta_i$, and a parametric distribution as the prior distribution for the dispersion parameter ϕ_{ij}:

$$\log(\phi_{ij}) = \log(\phi_{ij}^0) + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The prior mean, $\log(\phi_{ij}^0)$, is estimated from some preliminary model (e.g. NBQ or a smooth fit like NBS) and is treated as known. Denote $\theta_{ij} = \log(\phi_{ij})$ and $\theta_{ij}^0 = \log(\phi_{ij}^0)$, so that $\theta_{ij} = \theta_{ij}^0 + \epsilon_i$. Across all m genes, we assume that ϵ_i's are independent, and denote the prior distribution of ϵ_i by $\pi(\epsilon_i|\sigma^2)$. We obtain the joint likelihood function of the unknown parameters (σ^2, β) as:

$$L(\sigma^2, \beta) = \prod_{i=1}^{m} \int \left[\prod_{j=1}^{n} \Pr(y_{ij}|\theta_{ij} = \theta_{ij}^0 + \epsilon_i, \beta_i) \right] \pi(\epsilon_i|\sigma^2) d\epsilon_i. \tag{3.6}$$

To estimate the hyperparameter σ^2 (and the collection of all NB regression coefficients β), one can use $(\hat{\sigma}^2, \hat{\beta}) = \arg\max_{\sigma^2, \beta} L(\sigma^2, \beta)$, but this approach can be quite difficult computationally. For easier evaluations, suppose we know σ^2 (i.e. fix θ_i) and rewrite the likelihood (Equation (3.6)) as

$$L(\sigma^2, \beta) = L_{\sigma^2}(\beta), \tag{3.7}$$

where β is varying. To estimate β, we maximize $L_{\sigma^2}(\beta)$ with respect to β, i.e. the MLE $\hat{\beta}_{\sigma^2} = \arg\max_\beta L_{\sigma^2}(\beta)$. Note that for each σ^2 we have a new curve $L_{\sigma^2}(\beta)$ over β. To estimate σ^2, we evaluate the maximum $L_{\sigma^2}(\beta)$ over β and choose the σ^2 that yields the maximum of all these curves. In other words, we have

$$\hat{\sigma}^2 = \arg\max_{\sigma^2} L_{\sigma^2}(\hat{\beta}). \tag{3.8}$$

By profiling out the nuisance parameter β, the likelihood $L_{\sigma^2}(\hat{\beta})$ is completely in terms of the parameter of interest σ^2. Therefore, the term $\Pr(y_{ij}|\theta_{ij} = \theta^0_{ij} + \epsilon_i, \beta_i)$ in Equation (3.6) is replaced by the profile likelihood of ϵ_i:

$$L_p(\epsilon) = \Pr(y_{ij}|\theta_{ij} = \theta^0_{ij} + \epsilon_i, \hat{\beta}_i(\theta_i)). \tag{3.9}$$

Let $l_p(\epsilon_i)$ be the log profile likelihood of ϵ_i so that $L_p(\epsilon_i) = \exp\{l_p(\epsilon_i)\}$. Define the integrated likelihood of σ^2 as

$$L_{int}(\sigma^2) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp\left(l_p(\epsilon_i; y_i) - \frac{\epsilon_i^2}{2\sigma^2}\right) d\epsilon_i. \tag{3.10}$$

The task is to evaluate $L_{int}(\sigma^2)$, which will be discussed next.

3.6.2.1 Laplace Approximation of Posterior Distributions

For simple models such as the generalized linear model, the Laplace approximation of integrals works quite well. This analytical approach is more efficient in terms of computation compared to the Markov Chain Monte Carlo (MCMC) approach that converges to the true posterior. It only has to find the posterior mode (i.e. the maximum *a posteriori*, MAP) estimator without exploring the whole posterior distribution. The performance of

the Laplace approximation gets better with (1) increasing sample size, as the posterior is asymptotically normally distributed; (2) re-parameterization of the bounded parameter (e.g. logarithm for the dispersion ϕ) to make its support on the entire real line.

Since for any j, $\epsilon_i = \theta_{ij} - \theta_{ij}^0$, we drop the sample index j and denote

$$\theta_i^* = \arg\max g(\theta_i; y_i) = \arg\max \left[l_p(\theta_i; y_i) - \frac{(\theta_i - \theta_i^0)^2}{2\sigma^2} \right],$$

so that

$$g'(\theta_i; y_i)|_{\theta_i = \theta_i^*} = l_p'(\theta_i^*; y_i) - \frac{\theta_i^* - \theta_i^0}{\sigma^2} = 0.$$

We can approximate the integral $\int \exp\{g(\theta_i; y_i)\}\, d\theta_i$ by the Laplace approximation, so that the approximate of the integrated likelihood (Equation (3.10)) is

$$L_{int}(\sigma^2) = \prod_{i=1}^{m} \left(1 - \sigma^2 l_p''(\theta_i^*)\right)^{-1/2} \cdot \exp\left(l_p(\theta_i^*; y_i) - \frac{(\theta_i^* - \theta_i^0)^2}{2\sigma^2} \right). \tag{3.11}$$

By optimizing (Equation (3.11)) we obtain the variance of the prior distribution of $\log(\phi)$.

To reduce the estimation bias from maximizing the (unadjusted) profile likelihood $l_p(\theta)$, we may use the adjusted profile likelihood (APL) discussed in [24]:

$$l_a(\theta) = l_p(\theta, \hat{\beta}_\theta) - \frac{1}{2} \log\det\left\{ j_{\beta\beta}\left(\hat{\beta}_\theta; \theta\right)\right\}, \tag{3.12}$$

where in the adjustment term, $j_{\beta\beta}$ is the observed information matrix for estimating $\hat{\beta}_\theta$, where $\hat{\beta}_\theta$ maximizes the constrained likelihood $l(\theta, \beta)$ for fixed θ (i.e. fixed ϕ). Currently, this APL approach is not implemented.

3.6.2.2 Simulation Studies

To evaluate the estimation accuracy for σ, we perform a set of simulations using the human RNA-Seq dataset as the "template" in order to preserve observed relationships between the dispersions and gene-specific mean counts. We simulate 5,000 genes with a single group of seven replicates: the mean structure μ is randomly generated from the log-normal distribution with mean 8.5 and standard deviation 1.5 (both on the log scale and the values are chosen to mimic the real dataset); the dispersion parameters are estimated from the real dataset. We compare $\hat{\sigma}$ with true σ specified at eight levels that are within a reasonable range for typical RNA-Seq data: 0.1, 0.3, 0.5, 0.7, 0.9, 1.2, 1.5 and 2.0. At each level of σ we evaluate three times using different random number seeds for generating $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Figure 3.6 shows the simulation results for the NB2 (left panel) and NBQ (right panel) models where the underlying dispersion model is refitted (NBP and NBS have similar results not shown here). We highlight the median value (out of three repetitions) in solid blue point at each σ level, and ideally these points should follow the $y = x$ reference line.

As discussed in the "Results/Power-Robustness Evaluations/Simulation Setup" subsection, we want to choose a σ that matches the residual dispersion variation in real data. This is achieved by a calibration approach which determines the $\sigma = \tilde{\sigma}$ to be used in simulations. The results are summarized in Table 3.5. The calibrated $\tilde{\sigma}$'s are essentially obtained from a calibration plot (included in Supporting Information S1) that can be generated as follows: first, at each of the eight true σ values, we obtain one estimated $\hat{\sigma}$ (we actually repeat three times using different random number seeds and take the median value highlighted in solid blue); second, we fit a quadratic curve to the eight points and calculate the associated 95% prediction interval; third, we calculate $\hat{\sigma}_{\text{real}}$ from the real dataset and plot the value as a horizontal line; fourth, we find the intersection of the horizontal line ($\hat{\sigma}_{\text{real}}$) with the fitted

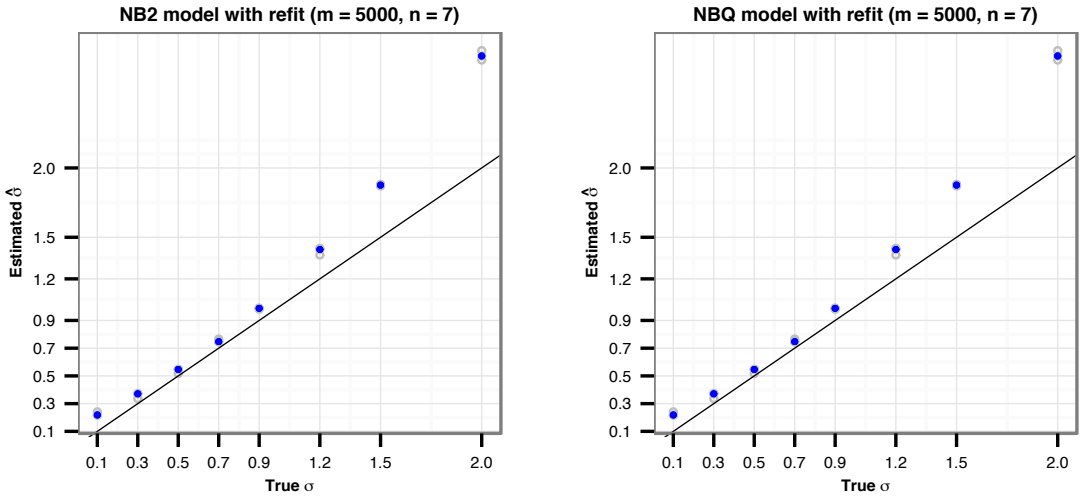

Figure 3.6: Simulation studies for empirical Bayes estimation of the residual variation in NB dispersions. We refit the underlying NB dispersion model and estimate σ three times at each level of the true σ's. The underlying dispersion models are NB2 (left panel) and NBQ (right panel).

curve to determine the $\tilde{\sigma}$ we will use in subsequent power-robustness simulations; the last step is to find the 95% calibration interval (CI), as determined jointly by the horizontal line ($\hat{\sigma}_{\text{real}}$) and the 95% prediction interval in the second step.

Table 3.5: Calibrated $\tilde{\sigma}$ to be used for power-robustness simulations for each dataset. We also report the 95% calibration interval (CI) for each point estimate.

Dataset	**#samples**	**Calibrated $\tilde{\sigma}$**	**95% CI**
Human30	(7, 7)	0.915	(0.90, 0.93)
Mouse	(3, 3)	1.025	(0.99, 1.06)
Zebrafish	(4, 4)	0.975	(0.94, 1.02)
Arabidopsis	(3, 3)	0.855	(0.80, 0.92)
Fruit fly	(4, 3)	0.890	(0.86, 0.93)

3.6.3 Software Information

The datasets (raw read count table) analyzed in this article are available from the `SeqDisp` package (version 0.1.4) for the cross-platform computing environment R (version $\geq$ 3.0.0). The R codes for performing the simulations are also included in the `SeqDisp` package (version 0.1.4) to be released at https://github.com/gu-mi/SeqDisp.

3.7 Acknowledgments

We thank Daniel W. Schafer, Sarah C. Emerson, Yuan Jiang and Jeff H. Chang for helpful discussions. This article is part of a doctoral dissertation written by the first author, under the supervision of YD and DWS. Work of YD and GM is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM104977. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

4 Length Bias Correction in Gene Ontology Enrichment Analysis Using Logistic Regression

Gu Mi, Yanming Di, Sarah Emerson, Jason S. Cumbie and Jeff H. Chang

PLOS ONE

Public Library of Science

1160 Battery Street

Koshland Building East, Suite 100

San Francisco, CA 94111, USA

PLOS ONE 7(10): e46128 (October 2012)

4.1 Abstract

When assessing differential gene expression from RNA sequencing data, commonly used statistical tests tend to have greater power to detect differential expression of genes encoding longer transcripts. This phenomenon, called "length bias", will influence subsequent analyses such as Gene Ontology enrichment analysis. In the presence of length bias, Gene Ontology categories that include longer genes are more likely to be identified as enriched. These categories, however, are not necessarily biologically more relevant.

We show that one can effectively adjust for length bias in Gene Ontology analysis by including transcript length as a covariate in a logistic regression model. The logistic regression model makes the statistical issue underlying length bias more transparent: transcript length becomes a confounding factor when it correlates with both the Gene Ontology membership and the significance of the differential expression test. The inclusion of the transcript length as a covariate allows one to investigate the direct correlation between the Gene Ontology membership and the significance of testing differential expression, conditional on the transcript length. We present both real and simulated data examples to show that the logistic regression approach is simple, effective, and flexible.

4.2 Introduction

RNA sequencing (RNA-Seq) has the potential to enable simultaneous measurement of expression for all genes expressed in a cell. Statistical tests (Robinson et al. [102], Anders and Huber [3], Di et al. [31]) further enable assessment of differential expression (DE) of individual genes under different environmental or experimental conditions. To relate the outcome of the DE analysis to biological functions, a widely-used approach is to examine enriched Gene Ontology (GO) categories based on the terms annotated to the genes

identified as DE (Khatri and Drăghici [58], Ashburner et al. [6]). GO uses a structured vocabulary to describe functional categories of gene products. Genes annotated with the same GO term form a gene category and share a common biological function. The enrichment of a GO term among DE genes can be used to indicate the association of biological functions to variations in experimental conditions.

To quantify the enrichment of a GO term, one can dichotomize results of DE analysis and cross-classify the genes according to whether they are indicated as DE and whether they are annotated to the specific GO term. The level of enrichment can then be assessed using contingency-table-based tests such as the Fisher's exact test or the chi-square test (for a summary, see Sherman et al. [106]). Table 4.1 shows an example of testing the enrichment of the GO term GO:0005575 among DE genes in a prostate cancer dataset (see the Results section for more details). One unappealing feature of the contingency-table-based approach is that the numbers of DE and non-DE genes and, in turn, the GO enrichment test result depend on the p-value cut-off for declaring a gene as DE. In Table 4.1, genes with DE test p-values less than 0.05 are declared as DE. Figure 4.1 shows that Fisher's exact test p-values vary with DE test p-value cut-offs.

Table 4.1: A typical two-by-two contingency table for testing enrichment of a GO category.

	D	$\bar{D}$	Sum
C	1962	9803	11765
$\bar{C}$	118	709	827
Sum	2080	10512	12592

C: in category; $\bar{C}$: not in category;
D: DE genes; $\bar{D}$: non-DE genes.

Logistic regression is an alternative GO enrichment analysis approach that does not require dichotomizing DE test results. For each gene i, let the binary variable y_i indicate the

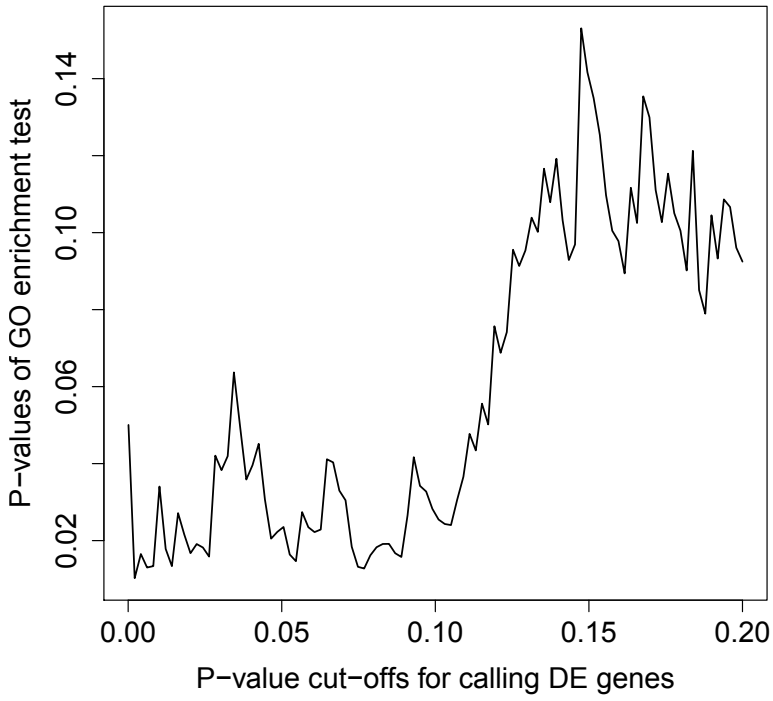

Figure 4.1: Influence of DE testing p-value thresholds on the determination of enriched categories. The p-value cut-off for calling DE genes (x-axis) influences the p-value of subsequent GO enrichment test (y-axis). Therefore, subjective decisions on declaring DE genes will make subsequent enrichment results rather unstable.

presence ($y_i = 1$) or absence ($y_i = 0$) of the gene in the GO category. Denote $\pi_i = \Pr(y_i = 1)$, and let x_i measure the significance of the DE test result (e.g., transformation of the DE test p-value). The logistic regression (McCulloch and Neuhaus [80])

$$\text{logit}\left[\pi_i\right] = \log\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \beta_1 x_i \tag{4.1}$$

relates the log odds of a gene belonging to a GO category to the significance of DE tests. A significant positive β_1 indicates that the odds of a gene belonging to this particular category increase as the significance of DE increases. Sartor *et al.* implemented the logistic regression model in the software LRpath and applied it to enrichment analyses for microarray expression data (Sartor et al. [105]). The logistic regression approach is more flexible than the contingency table approach. First, it is easy to include covariates in a logistic regression setting to adjust for potential confounding factors—such as gene length. Second, logistic regression allows the use of continuous measures of DE test significance, which conveys more information than dichotomizing DE test results and avoids the nuisance of choosing an arbitrary cut-off for DE test p-values.

One statistical concern with tests for enriched GO terms, particularly those based on analysis of RNA-Seq datasets, is that transcript length can be a confounding factor if it correlates with both the GO membership and the DE test significance. In regards to the latter, many existing DE tests have greater statistical power to detect DE for genes with more reads mapped to them (Robinson et al. [102], Anders and Huber [3], Di et al. [31]). Since genes with longer transcripts will have more reads mapped to them than an equally expressed shorter gene, the statistical *power* of these DE tests will depend on transcript lengths. Oshlack *et al.* refer to the dependence of DE test power on transcript length as length bias (Oshlack and Wakefield [87]). In the presence of length bias, subsequent GO enrichment analysis will have the potential to identify GO categories with a higher

proportion of longer genes. These categories are not necessarily biologically more relevant.

Young et al. [126] compensated for potential length bias by developing a weighted resampling strategy based on contingency tables. The basic idea is to estimate the DE test power as a smooth function of the transcript length and resample genes with weights inversely proportional to the estimated power of the DE test. For computational efficiency, the resampling method can be approximated by a test based on the Wallenius non-central hypergeometric distribution. If no length bias is present, the Wallenius approximation reduces to the Fisher's exact test, which is based on a central hypergeometric distribution. Their method is implemented in the Bioconductor package `goseq`. Gao *et al.* proposed a similar method where a different weighting function is used to compute the non-central parameter of the Wallenius distribution (Gao et al. [38]).

The dependencies between the GO terms should also be considered in the statistical assessment for enrichment of GO categories. GO terms are organized as a directed acyclic graph (DAG). In this DAG, parent terms describe more general functional categories than their child terms (Rhee et al. [96]) and each child term can have multiple parent terms. Because of this relationship, a gene, when annotated with a GO term, is automatically annotated to the term's parent terms. Furthermore, any gene has the potential for being annotated with multiple GO terms. Three distinct categories, *biological process* (BP), *cellular component* (CC) and *molecular function* (MF), describe the most general biological functions each having potentially thousands of annotated genes, while some of the very specialized categories (e.g. *inner membrane complex*) may even have no gene products annotated to them. When a GO term describing a general biological function is identified as enriched among DE genes, all its offspring – GO terms describing more specific functions – tend to be enriched as well. As a result, if we rank the GO terms according to enrichment test p-values, we tend to see many specific terms at the top of the list, which may result in

potentially misleading interpretation of the data. To address this issue, Alexa et al. [2] and Grossmann et al. [49] proposed "local" GO enrichment tests that incorporate the parent-child relationship among GO terms. The basic idea is to examine the relative enrichment of a GO term among genes that are offspring of the parents of the GO term being tested. If a GO term is enriched only because its parents are, the local GO enrichment test will not identify it as enriched. The `topGO` Bioconductor package (Alexa and Rahnenführer [1]) implements the method proposed in Alexa et al. [2], and the Ontologizer2 software (Bauer et al. [10]) implements methods proposed in Grossmann et al. [49].

In this paper, we describe our development of GOglm, a logistic regression model that effectively adjusts for length bias by including transcript length as a covariate. This inclusion allows one to investigate the direct correlation between the GO membership and the DE test significance conditional on the transcript length. We analyzed two public RNA-Seq datasets and simulated data to show that in comparison to the GOseq approach, the GOglm method for length bias corrections is equally effective, but confers the advantages of being more simple, transparent and flexible. We also show that the flexibility of GOglm allows one to address dependences in the GO terms.

4.3 Results

4.3.1 Length Bias Correction Using Logistic Regression

We propose to adjust for length bias using logistic regression. The method is simple and effective, allows the use of continuous measures of DE test significance, and is flexible to incorporate the parent-child relationship among GO terms.

Correcting length bias using logistic regression is straightforward. One includes a mea-

sure of gene length, l_i, as a covariate in the logistic regression model:

$$\text{logit}\left[\pi_i\right] = \beta_0 + \beta_1 x_i + \beta_2 l_i, \tag{4.2}$$

where i indexes genes, π_i is the probability of a gene belonging to the specified GO category, and x_i measures the significance of the DE test result. In GOglm, a gene's length is defined as the median length of all its corresponding mature transcripts.

The logistic regression model is easy to interpret and makes the underlying statistical issue more transparent: the fundamental cause of length bias is that the transcript length becomes a confounding factor when it correlates with both the GO membership and the DE test significance. When we include transcript length as a covariate, the coefficient β_1 now captures the correlation between the log odds of being in the specified GO category and the DE test significance – conditional on gene lengths. A significant result from the hypothesis test $H_0 : \beta_1 = 0$ indicates that the GO membership is correlated with the DE test significance even after adjusting for length bias.

The logistic regression method is more flexible than the GOseq approach. The logistic regression can be used in the contingency table setting by letting x_i be a binary variable indicating whether gene i is DE or non-DE. In the data examples below, we will show that the logistic regression method is equally effective in accounting for length bias as the GOseq approach. But more generally, the logistic regression model can use continuous measures of DE significance as explained earlier. In GOglm, one option is to use $x_i = \log(1 - \log(\text{DE test } p\text{-value}))$ as the continuous measure of DE. The inner log transformation helps us focus attention on the order of magnitude change in p-values. The outer log transformation will down weigh influence from extremely small p-values. There are other ways to construct the significance measure. We discussed earlier that p-values can be dichotomized, but that will incur loss of information. One can construct the significance

measure based on test statistic values (usually, the p-value is a monotone function of the test statistic). Other measures such as log fold change can also be used. We compared the performance of using different measures of DE in the section Simulation Studies.

Logistic regression is flexible to incorporate the parent-child relationship. To test for local enrichment, one fits the logistic regression using a subset of genes. For example, if the set of all genes annotated to the direct parents of the GO term is used to fit the regression model, the results will be similar to the parent-child union approach in Ontologizer2. In the following section on RNA-Seq data examples, we analyze the Arabidopsis data and compare results between GOglm and Ontologizer2.

Most statistical software includes efficient programs for fitting logistic regression models. We implement GOglm in R (R Core Team [93]) and fit the logistic regression model using the `glm` function with quasi-binomial error distribution in order to account for potential over-dispersions. The function `glm` uses the iteratively weighted least squares method (equivalent to Newton-Raphson algorithm for logistic regression) for estimation of regression coefficients and the Wald test for hypothesis tests of regression coefficients.

When a GO category contains very few (e.g., fewer than 5) annotated genes, the enrichment test p-value can be volatile and the statistical evidence can be unreliable. In GOglm, users have the option to exclude GO categories with too few genes when ranking the enrichment test results. Other GO analysis software such as `topGO` (Alexa and Rahnenführer [1]) and LRpath (Sartor et al. [105]) also allow users to filter out GO categories with small sizes.

4.3.2 RNA-Seq Data Examples

We present results to demonstrate the effectiveness and flexibility of the GOglm logistic regression method for GO enrichment analysis. First, we present results from GO enrichment analysis of the prostate cancer data (Li et al. [69]). Young et al. [126] used this dataset to validate their GOseq method. We compare the performance of GOglm and GOseq and demonstrate that the GOglm method effectively accounts for length bias. Second, we use GOglm to perform local enrichment test on an Arabidopsis dataset (Di et al. [31]) by examining a GO term's relative enrichment in the context of its direct parent(s). The results are compared to results derived from Ontologizer2 (Grossmann et al. [49]).

4.3.2.1 Prostate Cancer Data Example

The prostate cancer data of Li et al. [69] contain RNA-Seq reads that aligned to 49506 genes from three untreated and four treated cancer samples. It has been demonstrated that the GOseq method is effective in correcting length bias in this dataset (Young et al. [126]). We will use this dataset to compare the performance of GOglm and GOseq.

We used edgeR (Robinson et al. [102]) with a common dispersion estimate to obtain DE test p-values. The GOseq method was designed for contingency tables, so for the purpose of comparison, the p-values were dichotomized – a gene was called DE if the FDR adjusted p-value (i.e. q-value) was less than 0.05. The same DE results were then used by GOglm and GOseq for GO enrichment analyses.

As discussed in the Introduction section, length bias becomes a confounding factor when it correlates with both the response (GO category membership) and the predictor variable (DE test significance). To illustrate the prevalence of the correlation between GO category and gene length, we considered 4249 categories having at least 10 annotated genes and

asked what proportion of these categories showed significant correlation with gene length. For each GO category, we tested the correlation between category membership and gene length by comparing lengths of genes (on log scale) within and out of the category using Welch's two-sample t-test. We found the proportion to be 44.7%, indicating that significant correlation between gene length and GO category membership is prevalent among GO categories.

In Figure 4.2, we compare GO enrichment test p-values from the logistic regression method in GOglm and from the Wallenius method in GOseq. There is strong correlation (0.854) between the two sets of p-values, especially for small p-values. In Table 4.2, we list the most enriched GO terms identified by GOglm, and all of them also rank highly on GOseq's top list.

Figure 4.3 demonstrates the effect of length bias corrections. We first ranked the GO categories according to one of the GO enrichment tests: GOglm, GOseq, or the Fisher's exact test. We then divided GO categories into 300 GO groups according to the average gene length in each category and computed the average GO enrichment rank in each group. In Figure 4.3, we plot the average GO enrichment ranks against the average gene lengths in the 300 GO groups. Panel B shows that the ranks based on the Fisher's exact test – which was not adjusted for length bias – tend to be more biased towards GO categories with greater average gene lengths. This trend is less pronounced in panels A and C, where data were analyzed using GOglm and GOseq, respectively.

4.3.2.2 Arabidopsis Data Example

The Arabidopsis data contain RNA-Seq reads that aligned to more than 25000 genes from two groups of Arabidopsis samples of size three each. The two groups were derived from

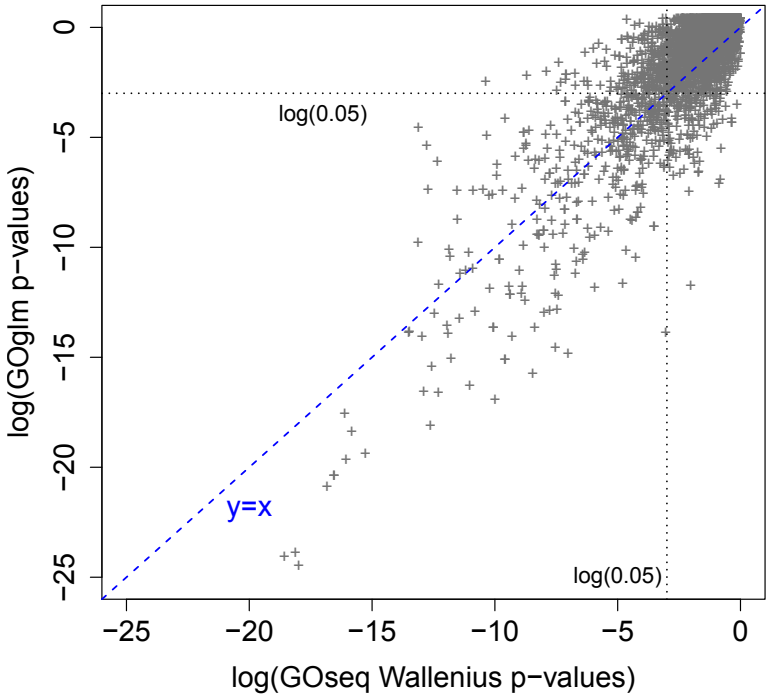

Figure 4.2: Comparison of p-values (on log scale) between GOseq Wallenius and GOglm. Among 3966 GO terms in the prostate cancer dataset, GOseq Wallenius and GOglm detected 492 and 486 enriched categories, respectively. Each plus sign denotes one category.

Table 4.2: Top 10 enriched categories of the prostate cancer dataset as ranked by GOglm.

Accession	Term	Onto[1]	p-value	GOglm[2]	GOseq[3]	Leng[4]	Anno[5]
GO:0005737	cytoplasm	CC	2.30e-13	1	3	2972	6731
GO:0007049	cell cycle	BP	9.19e-12	2	2	3260	1044
GO:0000278	mitotic cell cycle	BP	1.92e-11	3	1	3288	607
GO:0022402	cell cycle process	BP	1.12e-10	4	4	3289	792
GO:0044444	cytoplasmic part	CC	1.41e-10	5	6	2882	4872
GO:0022403	cell cycle phase	BP	2.16e-9	6	8	3226	653
GO:0000087	M phase of mitotic cell cycle	BP	3.59e-9	7	9	3506	303
GO:0000280	nuclear division	BP	5.68e-9	8	12	3555	295
GO:0007067	mitosis	BP	5.68e-9	9	13	3555	295
GO:0048285	organelle fission	BP	8.40e-9	10	16	3516	306

[1]BP, biological process; CC, cellular component.
[2]GOglm: category ranks by GOglm.
[3]GOseq: category ranks by GOseq Wallenius.
[4]Leng: median gene length (in base pair) within a category.
[5]Anno: number of annotated genes within a category.

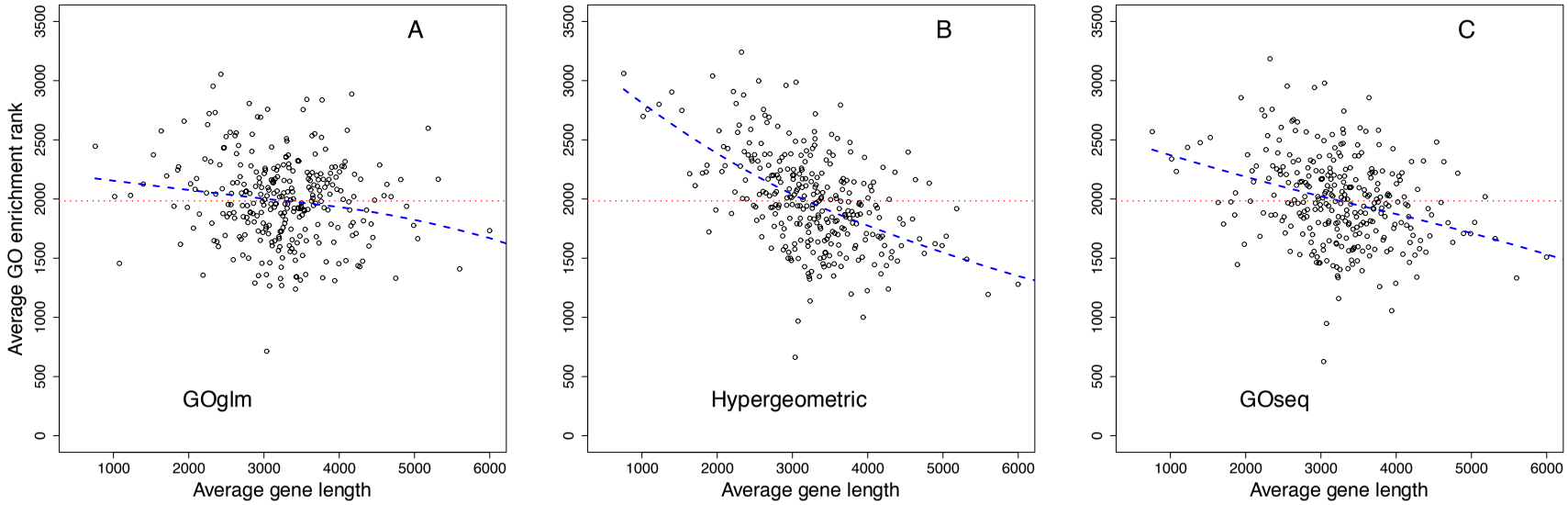

Figure 4.3: The effect of length bias corrections. GO categories are divided into 300 GO groups based on the average gene length in each category. In each plot, the x-axis represents the average gene length and the y-axis represents the average GO enrichment rank in each of the 300 GO groups. The Fisher's exact test (panel B) did not correct for length bias and the enrichment analysis based on this test tended to favor GO categories with longer average lengths. This is reflected as an obvious downward trend in panel B. The downward trend is less pronounced in panels A and C, where GOglm (panel A) and GOseq Wallenius (panel C) were used to adjust for length bias. A horizontal line has been added to each plot to facilitate visual comparison.

plants inoculated with $\Delta hrcC$ of *Pseudomonas syringae* pv tomato DC3000 or 10 mM $MgCl_2$ (mock). Di et al. [31] performed DE test on this dataset using the NBP negative binomial model. Cumbie et al. [26] performed local enrichment analysis on this data using the GORich tool of GENE-Counter. The dataset used in this paper comes directly from Di et al. [31], which is a subset of the data described in Cumbie et al. [26].

Here we used the logistic regression in GOglm to perform the local enrichment analysis and compared the results to those derived from Ontologizer2. We focused on up-regulated genes only. Using the R package `NBPSeq` (Di et al. [32]), we performed one-sided DE test to detect up-regulated genes. The logistic regression method in GOglm took continuous measures of DE based on the DE test p-values. In Ontologizer2, DE genes were determined by two criteria: the log fold-change greater than 0 (up-regulated) and the DE testing p-value less than 0.05. In the local enrichment tests, we tested the relative enrichment of a GO term relative to all genes that were offspring of the direct parents of the GO term being tested, which corresponded to the parent-child union (PCU) option in Ontologizer2.

From a total of 3851 categories, GOglm and Ontologizer2 detected 358 and 483 enriched categories respectively (using a p-value cut-off of 0.1 in the enrichment tests). Among the 358 categories identified by GOglm, 201 (56.1%) categories were also declared as enriched by Ontologizer2. The two methods also show high consistency if we focus on the rankings (instead of p-values) of the GO categories. Figure 4.4 shows the proportion of overlapping categories when the same number of top-ranked categories are selected using each method. Table 4.3 lists some of the most relevant categories that were declared as enriched by both GOglm and Ontologizer2. Categories such as plant defense, differential expression in response to pathogens, wounding, and/or stresses, signal perception, transduction, secretion or modification of plant cell wall were expected. For the complete ranking lists, see the Supporting Information (Table S1 and Table S2).

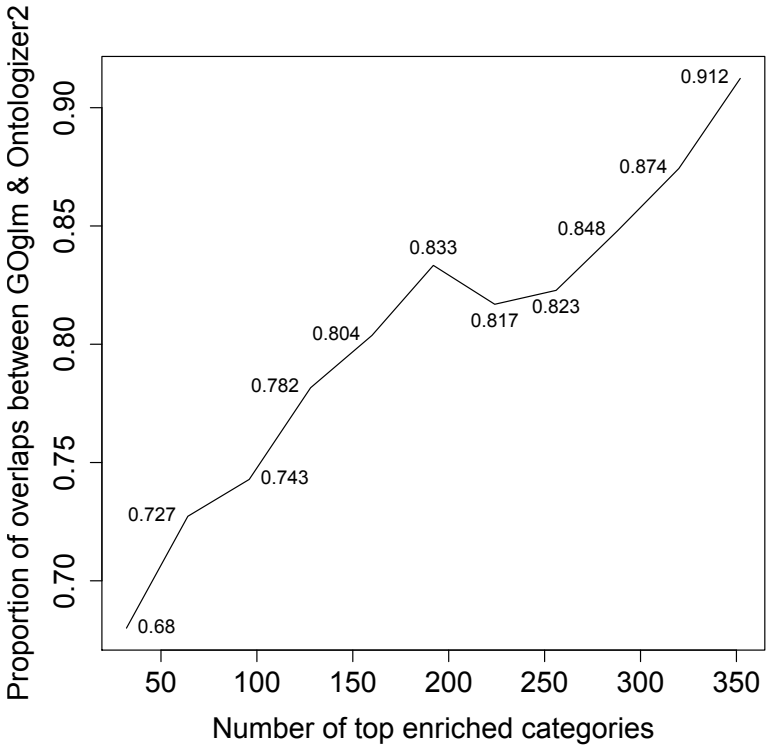

Figure 4.4: Proportion of overlapping categories by GOglm and Ontologizer2 (PCU). The proportion of overlapping categories (y-axis) when the same number (x-axis) of top-ranked categories are selected using GOglm and Ontologizer2 (PCU). As more enriched categories were included, there were more overlaps (enriched categories in common) between the two approaches as seen by the increasing trend and the percentages.

Table 4.3: Partial list of enriched categories identified by GOglm in the Arabidopsis dataset. Top 358 and top 483 categories are declared as enriched by GOglm and Ontologizer2 (PCU), respectively.

Accession	Term	Onto[1]	p-value	GOglm[2]	Ontgz[3]	Leng[4]	Anno[5]
GO:0050896	response to stimulus	BP	1.20e-19	4	1	1791	4062
GO:0009607	response to biotic stimulus	BP	9.66e-15	6	4	1809	643
GO:0009611	response to wounding	BP	9.66e-13	11	6	1801	164
GO:0045730	respiratory burst	BP	1.78e-7	28	93	1833	6
GO:0009753	response to jasmonic acid stimulus	BP	1.65e-6	37	18	1604	169
GO:0005886	plasma membrane	CC	3.28e-6	41	28	1996	1763
GO:0006952	defense response	BP	5.90e-5	58	39	1940	763
GO:0052482	defense response by cell wall thickening	BP	3.44e-4	69	119	2537	15
GO:0004568	chitinase activity	MF	6.50e-4	78	45	1104	17
GO:0009867	jasmonic acid mediated signaling pathway	BP	1.21e-3	90	150	1711	44

[1]BP, biological process; CC, cellular component; MF, molecular function.
[2]GOglm: category ranks by GOglm's local enrichment test.
[3]Ontgz: category ranks by Ontologizer2 (PCU).
[4]Leng: median gene length (in base pair) within a category.
[5]Anno: number of annotated genes within a category.

4.3.3 Simulation Studies

4.3.3.1 Simulation I

We developed a simulated dataset to further clarify the cause of length bias and demonstrate the effectiveness of our method (GOglm) in correcting the length bias. This dataset consisted of 10426 genes that were binned into 40 non-overlapping categories of different sizes. These 40 categories were simulated such that they varied in the average length of genes, but none of the categories was significantly enriched with DE genes. (See Materials and Methods for further details on the simulation setup.)

To demonstrate the efficacy of our method, we analyzed this simulated dataset using three different methods: 1) a simple logistic regression without length bias corrections (see equation (4.1)); 2) the GOglm method: logistic regression using log gene length as one additional covariate (see equation (4.2)); 3) the GOseq method.

In the absence of any length bias corrections, the simple logistic regression method identified a higher than expected number of categories with small p-values (false positives; Figure 4.5). In contrast, with length bias corrections, both the GOglm and GOseq methods identified the correct proportions of small p-values as expected under this simulation. A further examination comparing the scatter plots of enrichment test statistic values versus median gene lengths across gene categories provided additional support that the correction methods addressed the problem of length bias (Figure 4.6). These scatter plots clearly revealed that the simple logistic method in the absence of any length bias corrections favored categories with longer genes whereas the GOglm method correctly adjusted for the length bias.

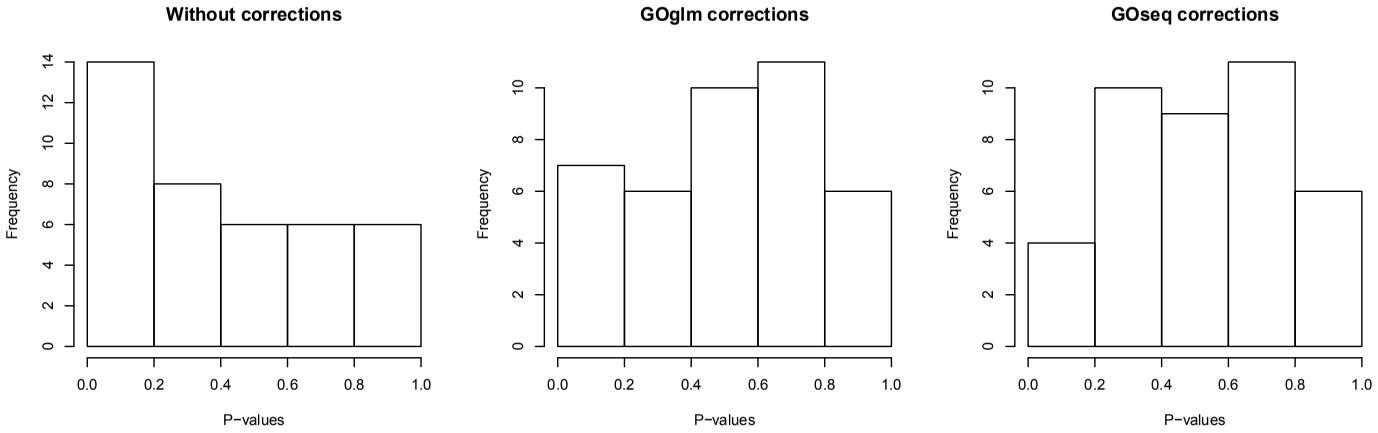

Figure 4.5: Histograms of enrichment test p-values from three enrichment analysis methods: logistic regression without length bias corrections, GOglm, and GOseq. The left panel (no length bias corrections) shows a more than expected proportion of small p-values (false positives). The GOglm (middle panel) and the GOseq (right panel) both gave correct p-value distributions expected under the simulation setting.

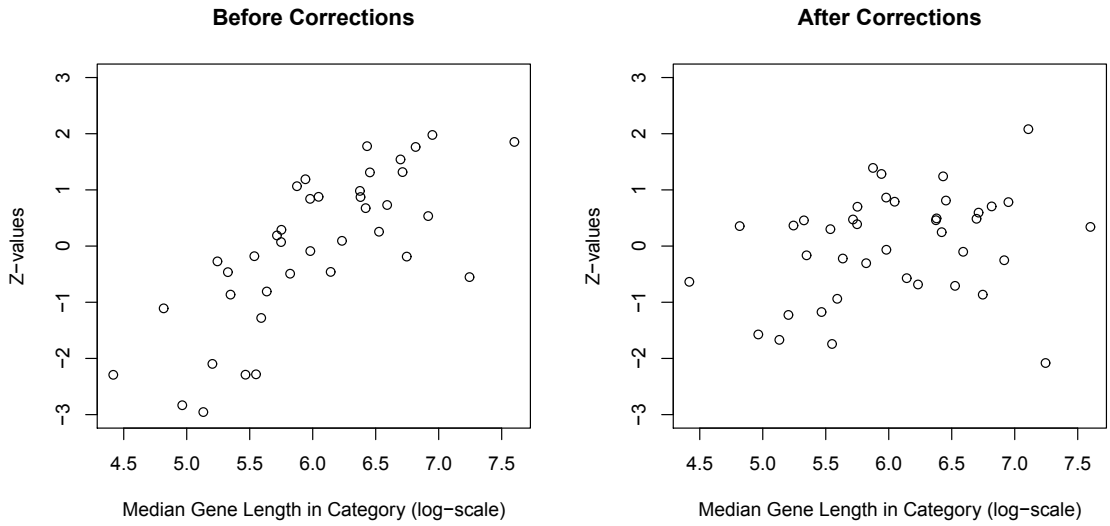

Figure 4.6: Scatter plots of the enrichment test statistic value against median gene length in category (log scale) before and after length bias corrections. Before length bias corrections, the enrichment test statistic value tends to increase with median gene length in category (left panel). After length bias corrections using GOglm, the trend is no longer visible (right panel).

4.3.3.2 Simulation II

We use a second simulation to highlight an important difference between GOglm and GOseq. We also compared the performance of using different measures of DE test significance or different parameterizations of gene length in GOglm.

Here, we simulated six categories known to be enriched with equal proportions (15%) of DE genes (versus on average 10% of DE genes in the other 34 categories). However, we varied the degree to which the genes were differentially expressed: the DE genes in category one has the highest fold change whereas those in category six had the lowest fold change, but were still DE.

We compared the performance of the same three enrichment analysis approaches as in the previous example. However, with the GOglm method, we tested different measures of DE significance: untransformed DE test p-value, log fold change, log-transformed DE test p-value, $\log(1 - \log(\text{DE test } p\text{-value}))$, and dichotomized DE p-values (cut-off used was 0.05). We also examined the use of either lengths or log-transformed lengths as the covariate. We repeated the simulation 10 times and summarized the average ranks given to the six known enriched categories by the different approaches and parameterizations (Table 4.4). Based on the simulation results, we make the following conclusions:

1. The GOglm method with $-\log(p\text{-values})$ and log-transformed lengths (column 5) as covariates yielded the best performance: the known enriched categories were ranked highly and the ranking order reflected the degree of DE. When untransformed lengths were used with $-\log(p\text{-values})$ (column 6), the performance was very similar, but the average ranks given to the top two enriched categories were closer, indicating these two categories were less distinguishable under this setting.

2. The GOglm method with log fold change and log-transformed lengths (column 4) as

Table 4.4: Average ranks of the six known enriched categories by different enrichment tests (over 10 simulations).

	Uncorrected	**GOglm**							**GOseq**
Significance[1]	$-\log(p)$	$-p$	$\log_2$FC	$-\log(p)$	$-\log(p)$	d-log	d-log	0/1	0/1
Length[2]	none	$\log(l)$	$\log(l)$	$\log(l)$	l	$\log(l)$	l	$\log(l)$	$s(l)$
1	2.2	9.0	1.8	1.7	1.8	2.1	2.2	3.9	3.6
2	2.8	7.0	2.4	2.0	1.9	2.5	2.4	2.0	2.1
3	2.9	7.5	2.7	2.6	2.6	2.9	2.9	2.5	2.3
4	6.5	7.0	4.1	4.1	4.1	4.1	4.2	4.0	3.9
5	6.4	6.4	5.4	5.5	5.4	5.2	5.2	5.4	4.8
6	14.2	14.0	12.1	12.3	12.6	12.1	12.0	12.7	13.4

[1]Significance: measures of DE significance. The measures used include minus untransformed DE test p-value, $\log_2$ fold change, minus log-transformed DE test p-value, $\log(1-\log(\text{DE test } p\text{-value}))$ (d-log in the table header), and dichotomized DE p-values (cut-off used was 0.05).
[2]Length: whether to use log-transformed lengths [$\log(l)$] or original lengths (l); a smooth function $s(l)$ of length is used by GOseq for length corrections (the so-called probability weighting function, PWF).

covariates also performed well. Note that conditional on the mean level (which is proportional to gene length under this simulation setting), the p-value is a monotone function of the log fold change. We should also be aware that fold changes are not available in regression settings (e.g., when investigating the dependence of expression level on a continuous covariate).

3. The GOglm method using dichotomized p-values perform very similar to the GOseq method (columns 9 and 10). Both methods ranked the known enriched categories highly, but the ranking orders did not reflect the degree of DE. This is as expected, since using dichotomized p-values retains only information on whether a gene is DE, not on how much the gene is DE. This can be viewed as a feature, rather than a drawback, of using dichotomized p-values, since sometimes it might be of interest to identify categories with a large number of DE genes rather than categories with a few extremely DE genes. Nevertheless, users should be aware of this interesting difference between using continuous p-values and using dichotomized p-values.

4. Using the $\log(1-\log(p))$ transformation (columns 7 and 8) ranked the top 3 enriched categories slightly lower than using the log-transformed p-values (columns 5 and 6). In our real data examples, we used the $\log(1-\log(p))$ transformation to down weigh very extreme p-values. This became unnecessary for the simulated datasets, since we did not simulate very large fold changes. Again, we view this as a feature rather than a drawback. In practice, we recommend the initial use of scatter plots to help determine which transformations of p-values and gene lengths are most suitable. One strength of the logistic regression method is that it allows different choices.

5. When length bias was not corrected (column 2), the logistic regression method still ranked the three enriched categories with the highest degree of DE highly, but the

average ranks given to categories 4 and 5 were lower. This indicated that some non-enriched categories sometimes received higher rankings. As we now understand, this can happen due to length bias.

6. Use of untransformed p-values (column 3) gave very poor performance. The degree of DE significance was not reflected well on the original scale of untransformed p-values.

4.4 Discussion

The GOglm method discussed in this paper provides a simple and effective tool for GO enrichment analysis with simultaneous length bias corrections. The use of continuous measures of DE avoids the subjective specification of p-value thresholds that is inevitable for any contingency-table-based approaches. The GOglm is also applicable to local enrichment analyses which account for the parent-child relationship of the GO structure.

GO facilitates comprehensive and systematic functional explorations based on currently known biological knowledge. However, the complexity inherent in the ontology structure introduces statistical challenges from several different sources. Here we list a few issues that warrant deeper consideration beyond the scope of this article. Researchers should be aware of the potential influences these issues may have on inference before embracing any statistical tools available for enrichment analysis.

4.4.1 Filtering Procedures

Not all genes under study had GO annotations available, and in this paper we restricted attention to study genes annotated with at least one GO term. In a GO enrichment analysis, researchers in general aim to determine which categories are more enriched relative

to others in a database of all possible GO terms, so the selection is competitive in nature among the GO categories. Because genes without annotations tend to have relatively low expression and small variability across conditions, researchers may prefer to exclude them for limited discriminatory power and focus instead on the more studied genes. In the prostate cancer dataset, 10102 genes were without annotations, among which only 4.6% were called DE. In contrast, 16.9% of the 12592 genes having annotations were called DE.

Alternatively, researchers may choose to retain those genes without annotations in the study by assuming that they belong to a single "pseudo-category" and treating them as background genes. We note that analyses by GOseq in Young et al. [126] and the `goseq` Bioconductor package did not exclude genes without annotations, so the final category list was slightly different than the list produced in this study.

4.4.2 Challenges in GO Enrichment Analysis

4.4.2.1 Multiple Testing Correction

Because of the graph structure of GO terms, enrichment tests for different GO categories can be correlated, and therefore multiple testing corrections in GO analyses cannot be simply addressed by calculating ordinary false discovery rates (FDR) (Goeman and Mansmann [43]). In our analysis, we did not adjust for multiplicity partially because we were more interested in relative ranks of categories instead of absolute magnitude of gene set p-values.

Commonly used multiple testing corrections include the Holm-Bonferroni's correction (Holm [54]), Benjamini-Hochberg's FDR (Benjamini and Hochberg [11]), resampling-based p-value adjustment (Westfall and Young [122]), bootstrap and Monte-Carlo simulation approaches. Khatri and Drăghici [58] gave an overview of these corrections with an emphasis

on their performances under different total numbers of functional categories and dependency levels [58]. Graph-structured tests have been discussed in Goeman and Mansmann [43] where the authors proposed a "focus level" method for DAG, and a hierarchical approach suitable for tree structures is proposed in Meinshausen [81]. Both methods are implemented in the Bioconductor package `globaltest` (Goeman and Oosting [44], Goeman et al. [46, 45]). Han et al. [50] also discussed false discovery control when test statistics are correlated. Research on multiple corrections suitable for complex structures like GO is still in high demand.

4.4.2.2 Other Sources of Bias

In addition to length bias corrections, we can also implement GOglm in other situations such as corrections for total read counts of RNA-Seq data or corrections for intensity levels in microarray (Young et al. [126]) in a similar manner. To obtain more accurate gene expression levels, Zheng et al. [127] proposed a generalized-additive-model-based approach for simultaneous bias corrections from different sources, including gene lengths, GC content and dinucleotide frequencies. We believe that such bias corrections performed in regression settings are promising, though the choice of appropriate regression tools depends on the underlying problem of interest.

4.4.2.3 Annotation Quality

Even if the aforementioned statistical problems are properly addressed, the issue of annotation quality still exists as a non-statistical problem. Rhee et al. [96] reported 14 types of evidence supporting the association of GO identifiers to gene identifiers, with different levels of experimental validation. Less than 5% of all annotations have been manually

checked – which is considered as a reliable source of information. Over 95% of annotations, however, are indirectly derived (i.e. inferred from electronic annotation), leading to higher inaccuracy than those manually curated annotations (see Table 1 of Rhee et al. [96]). Therefore, more efforts from biologists are needed to improve the annotation quality.

4.4.3 Fundamental Assumption

One fundamental assumption underlying length bias corrections is that there is no biological cause for longer genes to be more differentially expressed, on average, than shorter genes. This assumption is not statistically verifiable, but both Oshlack and Wakefield [87] and Young et al. [126] cited evidence from microarray studies to support this assumption. If there is a biological cause that violates this assumption, then its effect will not be fully detected in GO enrichment analysis if length bias is corrected.

4.5 Conclusion

We discussed a simple and effective method for length bias corrections in the GO enrichment analysis using logistic regression. We validated its effectiveness by analyzing real and simulated RNA-Seq datasets. We also compared its performance with alternative enrichment methods (e.g. GOseq) and examined the difference between the two approaches via simulations. Explicitly modeling the gene length as a covariate in the logistic regression framework helps to reduce length bias and enables flexible implementations and straightforward interpretations. The use of continuous measures of DE avoids the subjective specification of p-value thresholds. The method is flexible and applicable to local enrichment analyses which account for the parent-child relationship of the GO structure. making it a promising tool for enrichment analyses under different scenarios.

4.6 Materials and Methods

4.6.1 Preprocessing of the Prostate Cancer Dataset

The prostate cancer data (Li et al. [69]) consist of seven samples: three from mock treated prostate cancer cells and four from treated cancer cells. The data originally consisted of 49605 genes as annotated using Ensembl gene ID (ensGene) and NCBI Build 36.3 (hg18). In order to directly compare GOglm to GOseq, we did not attempt to remap the reads to the genome. We fetched gene lengths and mapped gene identifiers to GO terms from available annotation Bioconductor packages. For example, we used the `org.Hs.egGO2ALLEGS` R object in the Bioconductor annotation package `org.Hs.eg.db` to obtain mappings between a given GO category and all its annotated Entrez Gene identifiers (Carlson et al. [22]). Gene length information is also accessible from the Ensembl Project online. Descriptions of data preparations can be found in the additional file of Young et al. [126].

The 49506 genes under study were first filtered by excluding genes with no fold changes. Among the remaining 22743 genes, 12592 genes had GO annotations for a total of 13956 unique GO categories. In our GO enrichment analyses, we excluded 9162 categories (~69.8%) with fewer than 10 annotated genes from the final enrichment ranking list. Statistical and biological considerations for gene subsetting were mentioned in the Results and Discussion sections, respectively.

4.6.2 Preprocessing of the Arabidopsis Dataset

The `org.At.tairGO2ALLTAIRS` R object in the Bioconductor annotation package `org.At.tair.db` provides mappings between a GO category and all its annotated TAIR identifiers (Carlson et al. [20]). We subsetted our dataset from a total of 6916 GO terms available in the

database.

Testing a GO term's relative enrichment requires knowledge of this term's direct parental term(s), and this information is available from the `GO.db` Bioconductor package (Carlson et al. [21]). We found in our Arabidopsis dataset 3993 BP, 601 CC and 2322 MF terms each having at least one annotated gene.

We excluded 2039 genes that had zero read counts in all samples, and discarded an additional 909 genes with zero fold change. We further excluded 2504 genes without annotations, so that the original 26222 genes were subsetted into 20770 genes associated with 6916 unique categories. Median transcript lengths for all genes were available. In this example, we excluded categories with fewer than 4 genes (~50%) and focused on 3851 categories for enrichment analysis.

Assessment of Differential Gene Expression

For ease of comparison with published results, in analyzing the prostate cancer data (Li et al. [69]), we used edgeR (Robinson et al. [102]) with a common dispersion estimate to obtain DE test p-values. For the Arabidopsis dataset, we used NBPSeq (Di et al. [31]) to obtain DE test p-values. EdgeR and NBPSeq are both based on negative binomial models for RNA-Seq read frequencies. The negative binomial model captures potential extra-Poisson variation in RNA-Seq read frequencies between independent biological samples using a dispersion parameter. Other methods based on negative binomial model include the tagwise or trend options in edgeR, or the DESeq approach discussed in Anders and Huber [3]. All of these methods use the same exact NB test (Robinson and Smyth [100]) for assessing DE, but differ in how they estimate the dispersion parameter as a function of the mean frequency.

4.6.3 Simulation I

We simulated a 10426 gene dataset. The genes were binned into 40 non-overlapping categories with the number of genes ranging from 101 to 1252. To keep the simulation simple and focused on the issue of length bias, we simulated non-overlapping categories to avoid correlated test statistics. In addition, moderate sizes of simulated categories provide adequate statistical power in the enrichment analysis. Gene lengths (on the log scale) were simulated according to a normal distribution with mean 6 and standard deviation 0.7 (simulated lengths ranged from 20 to 5867 base pairs). We assigned genes with the shortest lengths to the first category, genes with the second shortest lengths to the second category, and so on. The last category therefore contained the longest genes. After assigning genes to categories, we added additional noise to the log gene length according a normal distribution with mean 0 and standard deviation 0.2 so that there was overlap between gene length distributions in different gene categories. This additional step was useful to avoid potential convergence issues in the logistic regression.

We simulated RNA-Seq read counts according to negative binomial distributions for 12 biological samples divided into two groups, each of size 6. We specified the expected values μ of read counts to be proportional to simulated gene lengths and the dispersion parameter ϕ as a function of the mean $1.5\mu^{-0.5}$. This dispersion model mimicked the one estimated for the Arabidopsis data in Cumbie et al. [26]. We randomly designated 20% of all genes as DE. The proportion of DE genes in each category varied from 14% and 26% due to chance variations, but was independent of the gene length distribution. For the DE genes, the mean values of read counts in one of the groups (randomly decided) was less than that in the other group, and the expected log fold change (base 2) between the two groups was 0.5. We performed DE tests using NBPSeq. The resulting p-values were transformed into a significance measure using $\log(1 - \log(p))$ and used as one covariate in

the logistic regression for testing category enrichment.

4.6.4 Simulation II

We simulated 40 gene categories with the same category sizes and gene length distributions as in the first simulation. The baseline mean levels of simulated RNA-Seq reads were again proportional to simulated gene lengths. However, this time we simulated six categories to be enriched with DE genes. In each of these six categories, there were 15% of DE genes. The remaining 34 categories had 10% of DE genes (randomly simulated among all genes, so the actual number of DE genes in each of these 34 categories followed a binomial distribution with probability 0.1). We assigned a constant $\log_2$ fold change of 0.5 to DE genes in the non-enriched categories, but the $\log_2$ fold changes of DE genes in the six enriched categories ranged from 0.5 to 1.0 at an increment of 0.1, resulting in varying degrees of DE among these categories.

4.6.5 Software Information

The R codes implementing GOglm are available at the first author's website: http://people.oregonstate.edu/~mig/Site/Research.html. GOglm conforms to the definition of Open Source as defined by the Open Source Initiative. We have licensed GOglm under the GNU General Public License.

4.7 Acknowledgments

We thank Dr. Gordon Smyth and Matthew Young for their insightful comments on the prostate cancer data analysis and gene filtering criteria.

5 General Conclusions

5.1 Summary of the Dissertation

As RNA-Seq has rapidly become the *de facto* technique in transcriptome profiling, appropriate statistical analyses on the high-throughput genomics data play a key role in making meaningful biological conclusions. Although some insights from the microarray domain are still helpful in RNA-Seq data analysis (at least conceptually), existing statistical methodologies developed for microarrays need to be revised or completely refined to accommodate some unique characteristics in RNA-Seq data. In this dissertation, broadly speaking, we focused on two stages in the RNA-Seq pipeline: evaluating the adequacy of negative binomial regression/dispersion models for testing differentially expressed genes, and adjusting for gene length bias in the subsequent downstream Gene Ontology enrichment analysis. We also developed computational tools in R for implementing the proposed methodologies and they have been actively maintained.

The negative binomial model has been widely adopted in the RNA-Seq community for modeling read counts variability. However, no confirming answer has been given as to whether the NB assumption is valid, and research on the adequacy of different NB dispersion models with varied complexity is also very limited. To fill this gap, we proposed simulation-based goodness-of-fit tests with diagnostic plots for checking NB models' adequacy. We adopted the parametric bootstrap technique for constructing empirical quantiles and obtaining the associated Monte Carlo p-value for a single gene, and used Fisher's combined probability test to evaluate the adequacy of NB dispersion models for multiple genes. We have not found any inadequacies with the NB distribution so far in our investigations

of real datasets, and NB dispersion models with different complexity offered varied levels of lack-of-fit. The results served as a starting point for a more comprehensive evaluation on power-robustness trade-offs for different NB dispersion models.

Given the goodness-of-fit test results (a p-value indicative of model fit), we further developed a more direct approach for checking model adequacy by quantifying the residual variation on top of an underlying NB dispersion trend (fitted dispersion model). The results were then used to simulate realistic RNA-Seq datasets for power-robustness analyses, i.e., given the noise level estimated from the dataset and other controlling factors, we compared several popular NB dispersion models by measures of true positives and false discoveries, as well as their performances in terms of gene rankings. Given that the RNA-Seq community is currently flooded with many approaches for modeling NB dispersions, our proposed effective measure for quantifying residual variations sheds light on whether we may have power gains by using dispersion-modeling approaches. Real-data-based simulation studies (across several species) also provided benchmarking investigations into the power and robustness properties of NB dispersion models, which will benefit both users and developers of these novel methods for DE tests.

From standard RNA-Seq protocols we know that a long gene is sequenced more often than a short gene, even though there is no difference between the number of mRNA molecules. Unlike in microarray data analysis, this results in between-gene differences arising from purely technical reasons. Hence, the transcript length in RNA-Seq studies becomes an important factor that must be accounted for in certain analysis stages (e.g., gene-specific normalizations). For the downstream analysis of testing enriched GO categories which aims to relate the outcome of DE analysis to biological functions, the transcript length becomes a confounding factor as it correlates with both the GO membership and the significance of the DE test. We proposed to adjust for such bias using the logistic

regression and incorporating the length as a covariate. The use of continuous measure of DE also avoided the subjective specification of p-value thresholds adopted by contingency-table-based approaches. After adjusting for the length covariate, we see from real and simulated datasets that enriched categories no longer favor longer transcripts.

5.2 Future Work

Although RNA-Seq is the current state-of-the-art technology for transcriptome studies, technical biases introduced from library preparation protocols, sequencing artifacts and sources of variation from unmeasured or unmodeled factors are still considerable. Subsequent analyses are inevitably influenced by such detrimental effects from these heterogeneous factors. We believe that although some of the known technical biases can be reduced as the technology advances in the future, there are still pressing demands on novel methods which can provide improved biological accuracy and reproducibility.

Avoiding confounding between real effects and experimental artifacts is essential when designing RNA-Seq experiments (e.g., preparing replicated samples several days apart and/or by different personnel). Surrogate variable analysis (SVA; Leek and Storey [66]) techniques have been proposed as alternatives to explicitly modeling batch effects, and are well-received in analyzing microarray experiments. The `sva` package (Leek et al. [67]) can estimate surrogate variables regardless of whether the batch effects are known or unknown (for removing known batch effects using ComBat, see Johnson et al. [57]). Any subsequent analysis can include the estimated surrogated variables as covariates in the usual way. However, because SVA is designed for symmetrically distributed data, it cannot be directly used for RNA-Seq read counts: some transformations are required to convert counts into continuous-scaled data, for example, using the "voom" transformation from the `limma`

package. A linear model with the primary variable and the set of surrogate variables can be created, and the rest of the analyses follow the standard SVA procedures. There is no previous work on applying SVA-like methods to RNA-Seq data, and we believe it is a relevant alternative approach for reducing the dispersion noise (c.f. Chapter 3) and accounting for the length bias in GO tests (c.f. Chapter 4).

Our current approach for quantifying residual variation in fitted models assumes a common σ estimated for all genes combined. Some preliminary results have indicated that dispersion noise levels are likely to vary across different expression levels. Therefore, as a straightforward extension to the methods discussed in Chapter 3, a refined approach would first stratify genes into K groups based on their relative frequencies, and then estimate σ_k in each group for $k = 1, \cdots, K$. This allows the possibility of fitting the most appropriate dispersion model on genes within each region, so we may strike a better balance between power gains (do not fit gene-wisely when a simpler model is sufficient) and robustness (more flexible model for regions having larger variability). Based on what we've learned about the pros and cons and model adequacy for each dispersion method (c.f. Chapter 2), this approach seeks to borrow their strengths to achieve more appropriate dispersion estimates after taking power and robustness into considerations.

The computational implementations for the methodologies discussed in each chapter are in the form of three separated R packages, which facilitate producing automated analysis pipeline but may not be user-friendly to a broader audience. Open-source web-based interactive applets are becoming more popular nowadays, with client-side applications taking advantage of end-users' computing powers and providing them with greater control of the data exploration and analysis processes. We have made some initial attempts on interactive visualization using the `Shiny` web framework (RStudio and Inc. [103]), and future relevant works include development of dynamic visualization and analytical tools that not

only automate the established analysis pipeline but also stimulate new biological insights.

Bibliography

[1] Adrian Alexa and Jörg Rahnenführer. Gene set enrichment analysis with topgo, 2009.

[2] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13):1600–1607, 2006.

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

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

[5] Simon Anders, Davis J McCarthy, Yunshun Chen, Michal Okoniewski, Gordon K Smyth, Wolfgang Huber, and Mark D Robinson. Count-based differential expression analysis of rna sequencing data using r and bioconductor. *Nature Protocols*, 8(9): 1765–1786, 2013.

[6] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25 (1):25–29, 2000.

[7] Anthony C Atkinson. Two graphical displays for outlying and influential observations in regression. *Biometrika*, 68(1):13–20, 1981.

[8] Kin Fai Au, Hui Jiang, Lan Lin, Yi Xing, and Wing Hung Wong. Detection of splice junctions from paired-end rna-seq data by splicemap. *Nucleic acids research*, 38(14): 4570–4578, 2010.

[9] Paul L Auer and Rebecca W Doerge. A two-stage poisson model for testing rna-seq data. *Statistical applications in genetics and molecular biology*, 10(1):1–26, 2011.

[10] Sebastian Bauer, Steffen Grossmann, Martin Vingron, and Peter N Robinson. Ontologizer 2.0a multifunctional tool for go term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1651, 2008.

[11] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[12] Donald J Best, John CW Rayner, and Olivier Thas. Anscombe's tests of fit for the negative binomial distribution. *Journal of Statistical Theory and Practice*, 3(3): 555–565, 2009.

[13] Yingtao Bi and Ramana V Davuluri. Npebseq: nonparametric empirical bayesian-based procedure for differential expression analysis of rna-seq data. *BMC bioinformatics*, 14(1):262, 2013.

[14] Mark S Boguski, Carolyn M Tolstoshev, Douglas E Bassett Jr, et al. Gene discovery in dbest. *Science*, 265(5181):1993–1994, 1994.

[15] Norman E Breslow. Extra-poisson variation in log-linear models. *Applied statistics*, pages 38–44, 1984.

[16] 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 Research*, 21(2):193–202, 2011.

[17] Jan J Brosens, Madhuri S Salker, Gijs Teklenburg, Jaya Nautiyal, Scarlett Salter, Emma S Lucas, Jennifer H Steel, Mark Christian, Yi-Wah Chan, and Carolien M Boomsma. Uterine selection of human embryos at implantation. *Scientific reports*, 4, 2014.

[18] Andreas Buja and Wolfgang Rolke. Calibration for simultaneity:(re) sampling methods for simultaneous inference with applications to function estimation and functional data. *Tomado De: http://charma. uprm. edu/˜ rolke/simulinf. pdf*, 20, 2003.

[19] James Bullard, Elizabeth Purdom, Kasper Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):94, 2010.

[20] Marc Carlson, Seth Falcon, Herve Pages, and Nianhua Li. *org.Hs.eg.db: Genome wide annotation for Human*, . R package version 2.6.4.

[21] Marc Carlson, Seth Falcon, Herve Pages, and Nianhua Li. *GO.db: A set of annotation maps describing the entire Gene Ontology*, . R package version 2.6.1.

[22] Marc Carlson, Seth Falcon, Herve Pages, and Nianhua Li. *org.At.tair.db: Genome wide annotation for Arabidopsis*, . R package version 2.6.4.

[23] Yunshun Chen, Aaron TL Lun, and Gordon K Smyth. Differential expression analysis of complex rna-seq experiments using edger. 2014.

[24] David R Cox and Nancy Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B. Methodological*, 49(1): 1–39, 1987.

[25] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[26] Jason S Cumbie, Jeffrey A Kimbrel, Yanming Di, Daniel W Schafer, Larry J Wilhelm, Samuel E Fox, Christopher M Sullivan, Aron D Curzon, James C Carrington, Todd C Mockler, et al. Gene-counter: a computational pipeline for the analysis of rna-seq data for gene expression differences. *PLoS One*, 6(10):e25279, 2011.

[27] Anthony C Davison. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

[28] Charmaine B Dean. Testing for overdispersion in poisson and binomial regression models. *Journal of the American Statistical Association*, 87(418):451–457, 1992.

[29] Charmaine B Dean and Jerald F Lawless. Tests for detecting overdispersion in poisson regression models. *Journal of the American Statistical Association*, 84(406):467–472, 1989.

[30] Yanming Di. Single-gene negative binomial regression models for rna-seq data with higher-order asymptotic inference. *Statistics and Its Interface*. in press.

[31] Yanming Di, Daniel W Schafer, Jason S Cumbie, and Jeff H Chang. The nbp negative binomial model for assessing differential gene expression from rna-seq. *Statistical Applications in Genetics and Molecular Biology*, 10(1), 2011.

[32] Yanming Di, Daniel W Schafer, with contributions from Jason S Cumbie, and Jeff H Chang. *NBPSeq: Negative Binomial Models for RNA-Sequencing Data*, 2011. URL `http://CRAN.R-project.org/package=NBPSeq`. R package version 0.1.4.

[33] Yanming Di, Sarah C Emerson, Daniel W Schafer, Jeffrey A Kimbrel, and Jeff H Chang. Higher order asymptotics for negative binomial regression inferences from rna-sequencing data. *Statistical applications in genetics and molecular biology*, 12(1): 49–70, 2013.

[34] Yanming Di, Daniel W Schafer, with contributions from Jason S Cumbie, and Jeff H Chang. *NBPSeq: Negative Binomial Models for RNA-Sequencing Data*, 2013. R package version 0.2.3.

[35] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6): 671–683, 2013.

[36] Mikel Esnaola, Pedro Puig, David Gonzalez, Robert Castelo, and Juan R Gonzalez. A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated rna-seq experiments. *BMC bioinformatics*, 14(1):254, 2013.

[37] Sir Ronald A Fisher. *Statistical methods for research workers*, volume 14. Oliver and Boyd Edinburgh, 1970.

[38] Liyan Gao, Zhide Fang, Kui Zhang, Degui Zhi, and Xiangqin Cui. Length bias correction for rna-seq data in gene set analyses. *Bioinformatics*, 27(5):662–669, 2011.

[39] Aldo M Garay, Elizabeth M Hashimoto, Edwin MM Ortega, and Víctor H Lachos. On estimation and influence diagnostics for zero-inflated negative binomial regression models. *Computational Statistics & Data Analysis*, 55(3):1304–1318, 2011.

[40] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

[41] Daniela S Gerhard, Lukas Wagner, Elise A Feingold, Carolyn M Shenmen, Lynette H Grouse, Greg Schuler, Steven L Klein, Susan Old, Rebekah Rasooly, Peter Good, et al. The status, quality, and expansion of the nih full-length cdna project: the mammalian gene collection (mgc). *Genome research*, 14(10B):2121–2127, 2004.

[42] Peter Glaus, Antti Honkela, and Magnus Rattray. Identifying differentially expressed transcripts from rna-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728, 2012.

[43] Jelle J Goeman and Ulrich Mansmann. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics*, 24(4):537–544, 2008.

[44] Jelle J Goeman and Jan Oosting. *Globaltest R package*, 2012. R package version 5.10.0.

[45] Jelle J Goeman, Sara A Van De Geer, Floor De Kort, and Hans C Van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[46] Jelle J Goeman, Sara A Van De Geer, and Hans C Van Houwelingen. Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):477–493, 2006.

[47] Gregory R Grant, Michael H Farkas, Angel D Pizarro, Nicholas F Lahens, Jonathan Schug, Brian P Brunk, Christian J Stoeckert, John B Hogenesch, and Eric A Pierce. Comparative analysis of rna-seq alignment algorithms and the rna-seq unified mapper (rum). *Bioinformatics*, 27(18):2518–2528, 2011.

[48] William Greene. Functional forms for the negative binomial model for count data. *Economics Letters*, 99(3):585–590, 2008.

[49] Steffen Grossmann, Sebastian Bauer, Peter N Robinson, and Martin Vingron. Improved detection of overrepresentation of gene-ontology annotations with parent–child analysis. *Bioinformatics*, 23(22):3024–3031, 2007.

[50] Xu Han, Weijie Gu, and Jianqing Fan. Control of the false discovery rate under arbitrary covariance dependence. *arXiv preprint arXiv:1012.4397*, 2010.

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

[52] Thomas J Hardcastle and Krystyna A Kelly. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11(1):422, 2010.

[53] Joseph M Hilbe. *Negative binomial regression*. Cambridge University Press, 2011.

[54] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[55] Polly Yingshan Hsu, Upendra K Devisetty, and Stacey L Harmer. Accurate timekeeping is controlled by a cycling activator in arabidopsis. *ELife*, 2, 2013.

[56] Wolfgang Huber and Alejandro Reyes. *pasilla: Data package with per-exon and per-gene read counts of RNA-seq samples of Pasilla knock-down by Brooks et al., Genome Research 2011.* R package version 0.2.16.

[57] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[58] Purvesh Khatri and Sorin Drăghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005.

[59] William Michael Landau and Peng Liu. Dispersion estimation and its effect on test performance in rna-seq data analysis: A simulation-based comparison of methods. *PloS one*, 8(12):e81415, 2013.

[60] James M Landwehr, Daryl Pregibon, and Anne C Shoemaker. Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, 79(385):61–71, 1984.

[61] Ben Langmead, Cole Trapnell, Mihai Pop, Steven L Salzberg, et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[62] Ben Langmead, Kasper D Hansen, Jeffrey T Leek, et al. Cloud-scale rna-sequencing differential expression analysis with myrna. *Genome Biol*, 11(8):R83, 2010.

[63] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. Voom! precision weights unlock linear model analysis tools for rna-seq read counts. *Preprint 2013*, 2013.

[64] Jerald F Lawless. Negative binomial and mixed poisson regression. *Canadian Journal of Statistics*, 15(3):209–225, 1987.

[65] Jerald F Lawless. Regression methods for poisson process data. *Journal of the American Statistical Association*, 82(399):808–815, 1987.

[66] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

[67] Jeffrey T Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey. *sva: Surrogate Variable Analysis*. R package version 3.10.0.

[68] Ning Leng, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart MG Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8):1035–1043, 2013.

[69] Hairi Li, Michael T Lovci, Young-Soo Kwon, Michael G Rosenfeld, Xiang-Dong Fu, and Gene W Yeo. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proceedings of the National Academy of Sciences*, 105(51):20179–20184, 2008.

[70] Jun Li and Robert Tibshirani. Finding consistent patterns: A nonparametric approach for identifying differential expression in rna-seq data. *Statistical methods in medical research*, 22(5):519–536, 2013.

[71] Jun Li, Daniela M Witten, Iain M Johnstone, and Robert Tibshirani. Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, 13 (3):523–538, 2012.

[72] Shuying S Li, Jeannette Bigler, Johanna W Lampe, John D Potter, and Ziding Feng. Fdr-controlling testing procedures and sample size determination for microarrays. *Statistics in medicine*, 24(15):2267–2280, 2005.

[73] Yuwen Liu, Jie Zhou, and Kevin P White. Rna-seq differential expression studies: more sequence, or more replication? *Bioinformatics*, page btt688, 2013.

[74] Yuwen Liu, Jie Zhou, and Kevin P White. Rna-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30(3):301–304, 2014.

[75] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *bioRxiv*, 2014.

[76] Steven P Lund, Dan Nettleton, Davis J McCarthy, Gordon K Smyth, et al. Detecting differential expression in rna-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical applications in genetics and molecular biology*, 11(5): 8, 2012.

[77] 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 research*, 18(9):1509–1517, 2008.

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

[79] Peter McCullagh and John A Nelder. *Generalized Linear Models*, volume 37. CRC Press, 1989.

[80] Charles E McCulloch and John M Neuhaus. *Generalized linear mixed models*. Wiley Online Library, 2001.

[81] Nicolai Meinshausen. Hierarchical testing of variable importance. *Biometrika*, 95(2): 265–278, 2008.

[82] Gu Mi, Yanming Di, and Daniel W Schafer. Goodness-of-fit tests and model diagnostics for negative binomial regression of rna sequencing data. revision submitted.

[83] Gu Mi, Yanming Di, Sarah Emerson, Jason S Cumbie, and Jeff H Chang. Length bias correction in gene ontology enrichment analysis using logistic regression. *PloS one*, 7(10):e46128, 2012.

[84] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[85] J.A. Nelder and R.W.M. Wedderburn. *Generalized linear models*. Number 135. Royal Statistical Society, 1972.

[86] Bernard V North, David Curtis, and Pak C Sham. A note on the calculation of empirical p values from monte carlo procedures. *American journal of human genetics*, 71(2):439, 2002.

[87] Alicia Oshlack and Matthew J Wakefield. Transcript length bias in rna-seq data confounds systems biology. *Biol Direct*, 4(1):14, 2009.

[88] Alicia Oshlack, Mark D Robinson, and Matthew D Young. From rna-seq reads to differential expression results. *Genome Biol*, 11(12):220, 2010.

[89] Fatih Ozsolak and Patrice M Milos. Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87–98, 2011.

[90] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, 2010.

[91] Donald A Pierce and Daniel W Schafer. Residuals in generalized linear models. *Journal of the American Statistical Association*, 81(396):977–986, 1986.

[92] Michael A Quail, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. *BMC genomics*, 13(1):341, 2012.

[93] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`.

[94] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

[95] Fred L Ramsey and Daniel W Schafer. *The Statistical Sleuth: A Course in Methods of Data Analysis: A Course in Methods of Data Analysis*. Cengage Learning, 2012.

[96] Seung Yon Rhee, Valerie Wood, Kara Dolinski, and Sorin Draghici. Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7):509–515, 2008.

[97] Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. Gc-content normalization for rna-seq data. *BMC bioinformatics*, 12(1):480, 2011.

[98] Dvaide Risso, John Ngai, Terence P. Speed, and Sandrine Dudoit. The role of spike-in standards in the normalization of rna-seq.

[99] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11(3):R25, 2010.

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

[101] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2008.

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

[103] RStudio and Inc. *shiny: Web Application Framework for R*, 2014. URL `http://CRAN.R-project.org/package=shiny`. R package version 0.9.1.

[104] Stephen M Rumble, Phil Lacroute, Adrian V Dalca, Marc Fiume, Arend Sidow, and Michael Brudno. Shrimp: accurate mapping of short color-space reads. *PLoS computational biology*, 5(5):e1000386, 2009.

[105] Maureen A Sartor, George D Leikauf, and Mario Medvedovic. Lrpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, 25(2):211–217, 2009.

[106] Brad T Sherman, Richard A Lempicki, et al. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2009.

[107] Yaqing Si and Peng Liu. An optimal test with maximum average power while controlling fdr with application to rna-seq data. *Biometrics*, 69(3):594–605, 2013.

[108] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(1):3, 2004.

[109] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[110] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics*, 14(1):91, 2013.

[111] Carolina F Svetliza and Gilberto A Paula. Diagnostics in nonlinear negative binomial models. *Communications in Statistics-Theory and Methods*, 32(6):1227–1250, 2003.

[112] Sonia Tarazona, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. Differential expression in rna-seq: a matter of depth. *Genome research*, 21 (12):2213–2223, 2011.

[113] 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. *Nature biotechnology*, 28(5):511–515, 2010.

[114] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562–578, 2012.

[115] Virginia G Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[116] Mark A Van De Wiel, Gwenaël GR 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, 2013.

[117] Victor E Velculescu, Lin Zhang, Bert Vogelstein, Kenneth W Kinzler, et al. Serial analysis of gene expression. *Science*, 270(5235):484–487, 1995.

[118] Wouter J Veneman, Oliver W Stockhammer, Leonie De Boer, Sebastian AJ Zaat, Annemarie H Meijer, and Herman P Spaink. A zebrafish high throughput screening system used for staphylococcus epidermidis infection marker discovery. *BMC genomics*, 14(1):255, 2013.

[119] Liguo Wang, Shengqin Wang, and Wei Li. Rseqc: quality control of rna-seq experiments. *Bioinformatics*, 28(16):2184–2185, 2012.

[120] Likun Wang, Zhixing Feng, Xi Wang, Xiaowo Wang, and Xuegong Zhang. Degseq: an r package for identifying differentially expressed genes from rna-seq data. *Bioinformatics*, 26(1):136–138, 2010.

[121] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[122] Peter H Westfall and S. Stanley Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. Wiley-Interscience, 1993.

[123] Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

[124] 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, 2013.

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

[126] Matthew D Young, Matthew J Wakefield, Gordon K Smyth, and Alicia Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*, 11 (2):R14, 2010.

[127] Wei Zheng, Lisa M Chung, and Hongyu Zhao. Bias detection and correction in rna-sequencing data. *BMC bioinformatics*, 12(1):290, 2011.

[128] Yihui Zhou, Kai Xia, and Fred A Wright. A powerful and flexible approach to the analysis of rna sequence count data. *Bioinformatics (Oxford, England)*, 27(19): 2672–2678, 2011.

APPENDIX

A Appendix (I): Description of the Earthquake Event Dataset

We consider an earthquake event dataset as an illustration on residual QQ plots and GOF tests for NB regression in Section 2.3.1. The dataset (provided in the `NBGOF` package) contains the frequencies of all earthquakes of a given magnitude (reported to one decimal place) for magnitudes from 4.5 to 9.1, that occurred between January 1, 1964 to December 31, 2012 (Source: Composite Earthquake Catalog, Advanced National Seismic System, Northern California Earthquake Data Center (NCEDC), http://quake.geo.berkeley.edu/cnss/). The empirical probability plots with GOF test results, based on a log-linear regression of mean number of earthquakes on magnitude, are shown in Figure 2.2. Neither the NB2 nor NBP model shows lack-of-fit.

B Appendix (II): Parameter Specifications (Univariate Simulations)

We specify the parameters in the simulation studies in Section 2.3.3 (Table 2.1) as follows: the mean is determined by $\mu = \exp(X'\beta)$ with the coefficient $\beta = (15, -1.5)$. The design matrix X takes an intercept and a covariate equally spaced from 4 to 8 of length $n = 10, 50$ and 100. The resulting mean levels approximately range from 20 to 8100.

For the NB2 model fit: the NB2 responses are simulated under $\alpha_0 = \log(0.1)$, $\alpha_1 = 0$ and $\phi = 0.1$. The NB1 responses are obtained by simulating NB2 with $\alpha_0 = \log(0.5)$, $\alpha_1 = -1$, and $\phi = 0.5/\mu$. The "NB2 plus outliers" responses are simulated with the same dispersions as in NB2, except we randomly double 20% of responses (as outliers). The "NB2 plus noise" responses are simulated with the same dispersions as in NB2, except ϕ is specified as $0.5/\mu \cdot \exp(G)$, where $G \sim \mathcal{N}(0, 1)$. The α in the variance $\mu + \phi\mu^\alpha$ is determined by $\alpha = \alpha_1 + 2$.

For the NBP model fit: the specifications are almost the same as in the NB2 model fit above, except for the simulated NB2 data, we use $\alpha_0 = \log(0.05)$ so that $\phi = 0.05$.

C Appendix (III): Supporting Information S1 for Chapter 3

C.1 Access to the Datasets

Information for all the datasets we analyzed in this article can be accessed from the NCBI website, using the GEO DataSets Advanced Search Builder. To obtain all the relevant information for an interested species (e.g., experiment descriptions, raw/processed data files, protocols and publications, etc.), we search in the "Organism" box and restrict the scope of "expression profiling by high throughput sequencing" in the "Filter" box.

The following datasets in the `SeqDisp` package contain read counts for all the samples in the original experiments: `human5`, `human30`, `mouse`, `zebrafish` and `arabidopsis`. For the `fruit.fly` dataset, as indicated in the `pasilla` package vignette, we only include read counts for seven samples. See SI Table C.1 for the accession numbers and sequencing platforms for each of the datasets.

Table C.1: Additional information for RNA-Seq datasets analyzed in this article.

Organism	**Accession Number**	**Platform**
Homo Sapiens	GSM1244809 – GSM1244816	Illumina HiSeq 2000
Mus Musculus	GSM1143032 – GSM1143040	Illumina HiSeq 2000
Danio Rerio	GSM1051294 – GSM1051301	Illumina HiSeq 2000
Arabidopsis Thaliana	GSM951349 – GSM951360	Illumina HiSeq 2000
Drosophila Melanogaster	GSM461176 – GSM461181	Illumina Genome Analyzer II

C.2 Supplementary Figure for the Calibration Plot of σ

We show in SI Figure C.1 a calibration plot from the mouse dataset analyzed in the Results section. We choose the true σ value at eight levels: 0.5, 0.7, 0.8, 0.9, 1.0. 1.1, 1.2 and 1.5, and simulate the dispersion ϕ_{ij} according to

$$\log(\phi_{ij}) = \log(\phi_{ij}^{\mathrm{NBQ}}) + \epsilon_i = \alpha_0 + \alpha_1 \log(\pi_{ij}) + \alpha_2[\log(\pi_{ij})]^2 + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ is the residual variation on top of a quadratic dispersion model with the parameters $\alpha_i, i = 0, 1, 2$, estimated from the mouse dataset. At each level of σ, we use the algorithm discussed in the Methods section to obtain three $\hat{\sigma}$'s (by using different random generating seeds), plot them in gray and highlight the median value in solid blue. We then fit a quadratic curve to the eight points (i.e., median values), with a 95% prediction interval superimposed in dashed curves. The $\hat{\sigma}$ estimated from the mouse dataset is also calculated, and the value is shown as a horizontal solid line. The intersection of the fitted quadratic curve and the horizontal line (the solid red point) has its x coordinate being the calibrated $\hat{\sigma}$. Similarly, the intersections between the upper/lower bound of the 95% prediction interval with the horizontal line determine the associated 95% calibration interval (CI) for the calibrated $\hat{\sigma}$.

We only include the calibration plot for the mouse dataset as an illustration. Similar plots for the rest of the datasets are not shown but are available upon request.

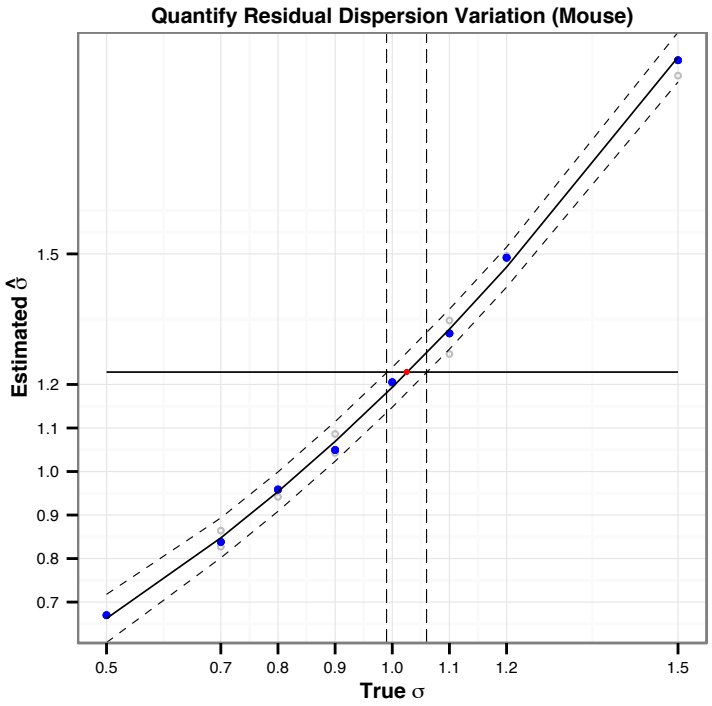

Figure C.1: The calibration approach for estimating residual dispersion variation σ in the mouse dataset. The simulated dispersions follow a quadratic trend (NBQ) which are estimated from the mouse dataset (two groups), subset to 5,000 genes.

C.3 Supplementary Figures for Mean-Dispersion Plots

In the main text, we showed the mean-dispersion plot in Figure 1 for the human dataset (with sequencing depth of 30 million). Here we provide such mean-dispersion plots for the mouse (SI Figure C.2), zebrafish (SI Figure C.3), arabidopsis (SI Figure C.4) and fruit fly (SI Figure C.5) datasets. The left panels are for the control groups, and the right panels are for the treatment groups.

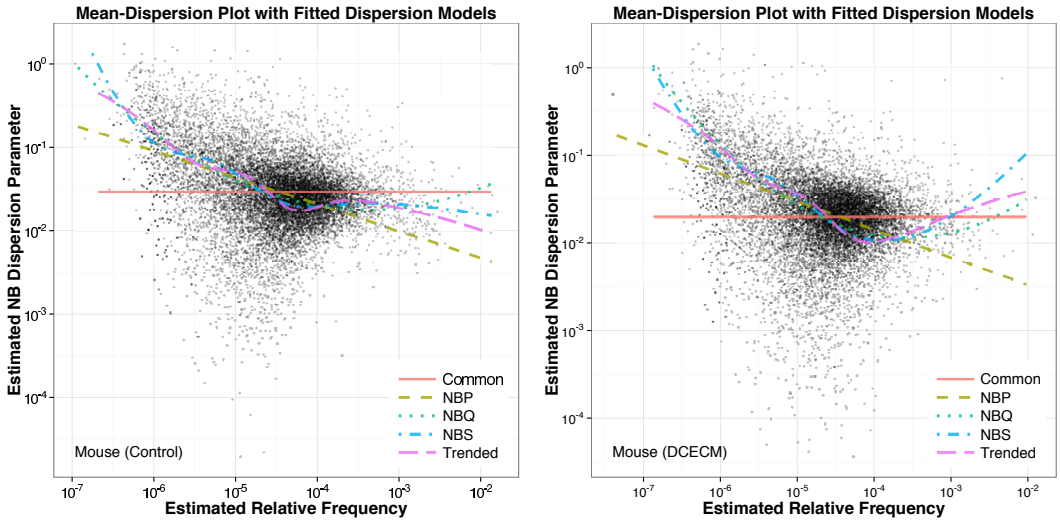

Figure C.2: Mean-Dispersion Plot of the Mouse RNA-Seq Dataset. The control (DCECM) group with three biological replicates is shown on the left (right) panel.

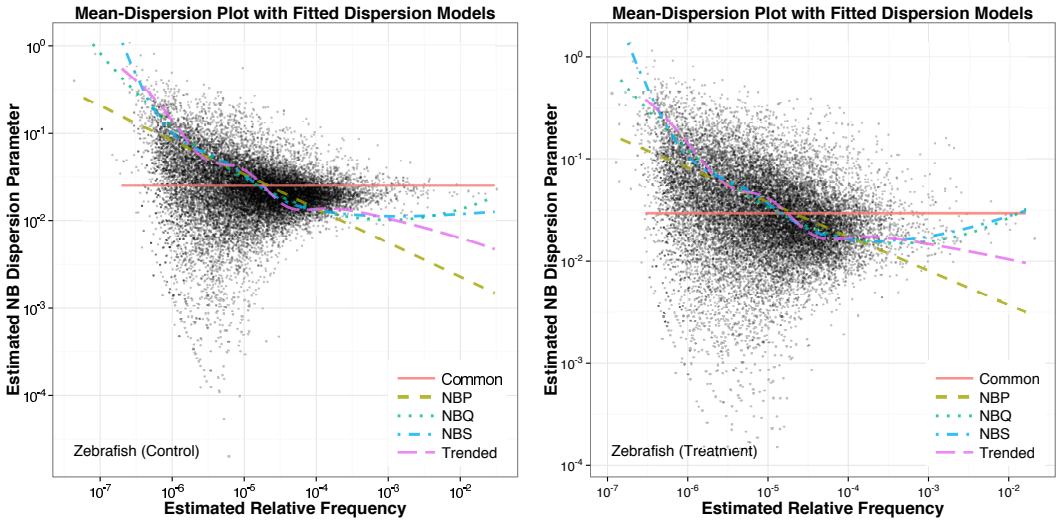

Figure C.3: Mean-Dispersion Plot of the Zebrafish RNA-Seq Dataset. The control (treatment) group with four biological replicates is shown on the left (right) panel.

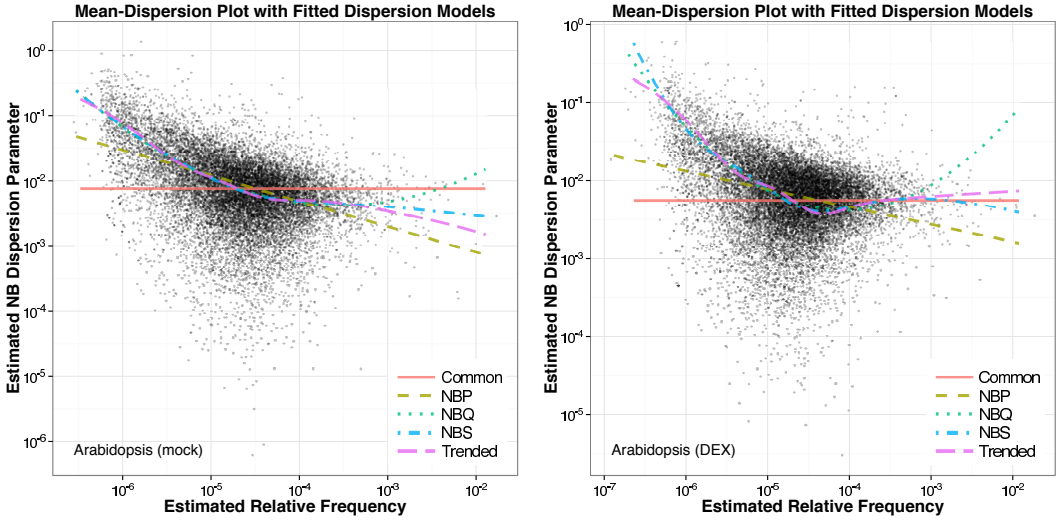

Figure C.4: Mean-Dispersion Plot of the Arabidopsis RNA-Seq Dataset. The mock (DEX) group with three biological replicates is shown on the left (right) panel.

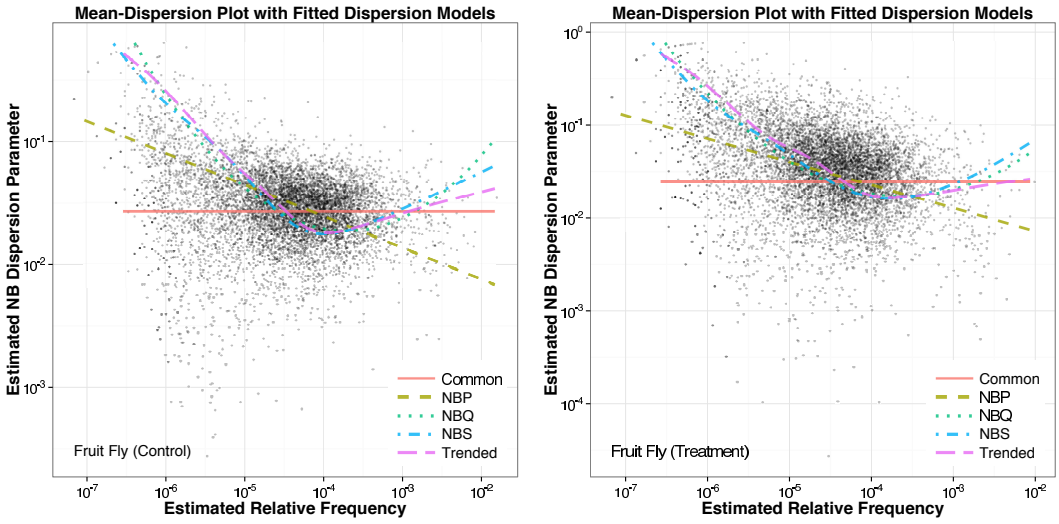

Figure C.5: Mean-Dispersion Plot of the Fruit Fly RNA-Seq Dataset. The untreated control (knockdown treatment) group with four (three) biological replicates is shown on the left (right) panel.

D Appendix (IV): Supporting Information for Chapter 4

Link to Table S1: Complete GO ranking list by GOglm.

`doi:10.1371/journal.pone.0046128.s001`.

Link to Table S2: Complete GO ranking list by Ontologizer2 (PCU).

`doi:10.1371/journal.pone.0046128.s002`.