Global Analysis of RNA-Seq Experiment: Multiple Data Sets & Multiple Genes

By

Bin Zhuo

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy in Statistics

Presented June 22, 2016
Commencement June 2016

AN ABSTRACT OF THE DISSERTATION OF

Bin Zhuo for the degree of Doctor of Philosophy in Statistics presented on June 22, 2016.

Title:
Global Analysis of RNA-Seq Experiment: Multiple Data Sets & Multiple Genes

Abstract approved:

Yanming Di

This is the abstract for my honors thesis. I'm going to start here.

Key Words: keyword1, keyword2, keyword3

Corresponding e-mail address: zhuob@oregonstate.edu

Global Analysis of RNA-Seq Experiment: Multiple Data Sets & Multiple Genes

By

Bin Zhuo

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy in Statistics

Presented June 22, 2016
Commencement June 2016

Doctor of Philosophy in Statistics dissertation of Bin Zhuo presented on June 22, 2016

APPROVED:

Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my project 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.

Bin Zhuo, Author

Contents

List of Figures

List of Tables

1. Introduction

1.1. Biological question of interest

1.1.1. Background

Gene is a piece of DNA that encodes a functional RNA or protein product, and is the basic physical and functional unit of heredity. The process by which genes are used to synthesize functional gene products is called *gene expression*. A gene is considered to be expressed in a cell or group of cells when a gene product is detected. These products can be transcribed messenger RNA (mRNA) and proteins for protein coding genes, or functional RNA species such as transfer RNA (tRNA) or small nuclear RNA (snRNA) for non-protein coding genes. Since the information encoded in a gene is first transcribed into RNA molecules, which is then used to make functional gene products, the RNAs transcribed in a certain condition reflect the current state of the cell.

Why do people do expression analysis?

In a typical gene expression experiment, researchers are usually interested in comparing expression levels of one or more genes from different sources. Factors for comparison could be *before vs after* effect in a drug treatment, *tumor vs normal* tissues in clinical study, or *wild type vs mutant* strains in plant research. Another important factor is the time-course, where cells/tissues at different stages are sampled with the purpose of discovering temporal pattern of gene expression. There are many other types of experiment, each with specific factors of interest to be studied.

What tools do people use to measure gene expression?

The expression levels of a gene can be measured using techniques such as complementary DNA (cDNA) libraries, microarray analysis, RNA fingerprinting by arbitrary primed PCR (RAP-PCR), expressed sequence tag (EST) sequencing, serial analysis of gene expression (SAGE), and RNA sequencing (RNA-Seq) (see [18] for a review).

RNA-Seq, also known as *whole transcriptome shotgun sequencing* [**?**], is a next-generation sequencing (NGS) technology used to uncover the presence and quality of RNA in a biological sample. It is rapidly becoming technology of choice for transcriptome profiling over the past few years. The standard procedure of an RNA-Seq experiment runs as follows [31]: first, the RNAs in the biological sample are fragmented and reverse-transcribed into cDNAs; second, the cDNA fragments are amplified and sequenced in a high-throughput sequencing platform (e.g., Illumine 3000, `http://www.illumina.com`) to generate (up to) hundreds of millions of reads; third, those reads are mapped to a reference genome or a reference transcriptome. It is the number of reads aligned to each gene (referred to as "read count") on the reference genome/transcriptome that quantifies the genes' expression levels.

pros and cons about RNA-Seq

RNA-Seq technology offers several key advantages over other methods [108], the most important of which are that it does not require prior knowledge of an organism for detecting transcripts, and that it is sensitive to genes expressed at either low or higher levels and thus provides higher dynamic range. The sequencing of RNA allows researchers to study the entire transcriptome of a species using only small amount of RNA. It has been demonstrated that a coordinated effort between RNA-Seq and real time PCR (RT-PCR) is one of the most effective ways to identify new exons [45]. However, one major challenge of this technique is data processing: RNA-Seq experiment produces a huge amount of reads (up to hundreds of millions per sample) and obtaining the expression profiles requires fast read mapping tools as well as a lot of computing resource [58**?**].

A workflow of pre-processing RNA-Seq data

Preprocessing RNA-Seq data consists of two main steps: 1) mapping reads to the reference genome/transcriptome, and 2) summarizing read counts at given genomic feature (e.g., exon, gene or transcript) level. Read mapping is the first computational, and usually, the most time-consuming step in RNA-Seq data analysis. Currently,

there are many alignment tools available, for example, `Bowtie` [57, 58], `BWA` [61, 62], `Subread` [92] and `STAR` [25]. In all situations, an index of either the reference genome/transcriptome or the reads is built at the beginning using hash tables or Burrows-Wheeler transform (BWT) [16]. The index allows fast retrieval of the set of positions in the reference sequence where the reads are more likely to align. Once those positions are decided, alignment is performed in the candidate regions. The precision and speed of the alignment is mainly determined by the algorithm used in the alignment tool (see [43] or [63] for a review). After the reads have been aligned, the numbers of reads mapped to each unit of a specified genomic feature are counted, giving the estimate of the corresponding expression levels. This can be done using `HTSeq` [2] or `featureCounts` [64], among other options. Finally, a read count matrix is obtained in which each row represents a genomic feature unit and each column corresponds to a biological sample.

In this research work, we assemble an in-house pipeline to process RNA-Seq data sets based on the R [83] platform. This pipeline, modified from a standard procedure given by Anders et al. [4], is designed to work for sequencing data available at the *National Center for Biotechnology Information* (NCBI, `http://www.ncbi.nlm.nih.gov/`). It uses the SRA (Sequence Read Archive) Toolkit [60] to convert SRA files to FASTQ files, the `Subread` aligner [65] to map reads to the reference genome, and then the `featureCounts` [64] to summarize counts (see Figure 1 for the work flow). We will use this pipeline to process multiple RNA-Seq data sets that are needed in Chapter 2.

1.1.2. Statistical issues

The statistical analysis beginning from the read count matrix consists of three major parts: 1) normalization—adjusting for sources of bias between samples; 2) differential expression (DE) analysis—testing whether the expression levels of a gene are associated with treatment or experimental variables; and 3) gene set test—detecting which

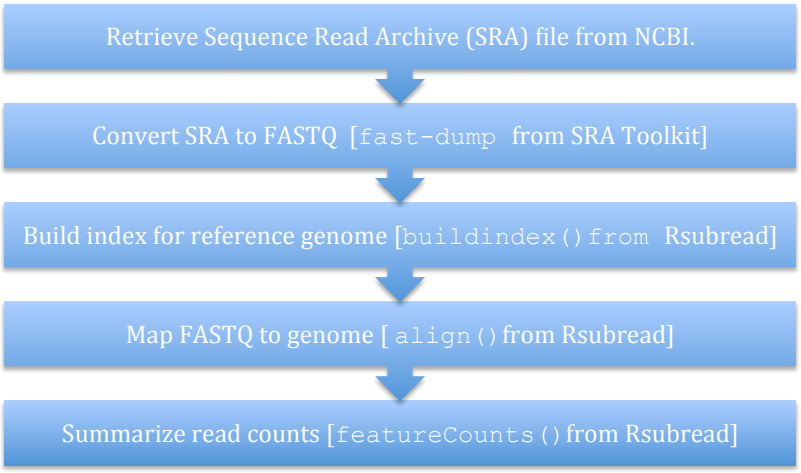

Figure 1: Work flow of data preprocessing: from raw reads sequencing data to read counts. Raw data in this workflow are retrieved from the NCBI respository. Data processing is carried out based on two softwares—the SRA Toolkit [60] and the Rsubread aligner [65].

biological pathways are over-represented with DE genes.

Normalization

Despite the optimistic claim that RNA-Seq does not need sophisticated normalization [108], many works have shown that normalization of count data is highly desirable to account for various sources of bias between samples before accessing differential expression [3, 24, 42, 87, 88, 89]. Normalization is needed for adjusting differences in sequencing depths or library sizes (total number of mapped reads for each biological sample) due to chance variation in sample preparation. In DE analysis, gene expression levels are often estimated from relative read frequencies. Therefore, normalization is also needed to account for the apparent reduction or increase in relative read frequencies of non-differentially expressed genes simply to accommodate the increased or decreased relative frequencies of truly DE genes. Currently there are many normalization methods available, such as the trimmed mean of M-values (TMM) [89], the DESeq normalization [3], and remove unwanted variation (RUV) [87].

DE analysis

Identification of DE genes is the key task in many gene expression studies. DE analysis uncovers the association between expression levels of a gene and the covariates of interest. The covariates could either be categorical (e.g., treatment/control status, cell types), or continuous (e.g., reagent concentration, time). For example, to understand the effect of a drug, one might ask which genes are *up-regulated* (increased expression levels) or *down-regulated* (decreased expression levels) between treatment and control groups? Finding these genes will help researchers to understand the cause of a disease and to develop effective medicine. In recent years, many statistical tools have been developed for DE detection (methods review can be found in [84, 91, 95]) in RNA-Seq experiments. Most of those approaches are based on Poisson [70, 107] or Negative Binomial (NB) distribution [3, 22, 78, 90, 113] because RNA-Seq expression data are present in the form of counts. The NB distribution based models are more popular for their flexibility to deal with *over-dispersion* (a.k.a. extra-Poisson variation) that are often observed in RNA-Seq expression data.

Gene set test

DE analysis evaluates each individual gene separately, but it fails to provide insights into biological mechanisms since genes may be correlated and function together. For this reason, *gene set test* is a frequently used technique that enables researchers to examine an ensemble of genes simultaneously and thus improves interpretability of DE results. Gene set test is the assessment of the association between a set of DE genes, which are significantly correlated with treatment or experimental design variables, and a prior set of genes, which are biologically related. Depending on the definition of the null hypothesis, there are two types of gene set test: the *self-contained* test and the *competitive* test [37]. A self-contained test examines a set of genes by a fixed standard without reference to other genes in the genome (see, for example, [39, 38, 49, 104, 111]). A competitive test compares DE genes in the test set to those

not in the test set [102, 112, 114]. The competitive gene set test is much more popular among genomic literatures [34, 37].

1.1.3. Questions for this thesis

In this thesis, we focus on three aspects of gene expression analysis: identifying stably expressed genes from multiple RNA-Seq data sets (Chapter 2); estimating correlations between test statistics via sample correlations [NEED TO REVISE] (Chapter 3); and adjusting for correlations in competitive gene set test (Chapter 4).

1.1.3.1 Identifying stably expressed genes

Many of the current normalization methods, for example, TMM [89] and DESeq [3] normalizations, assume that the majority of genes are not DE within the experiment under investigation. However, this assumption could be violated for some experiments where over 50% of the genes' expression levels are altered by the treatments [68, 110]. The consequence with such assumption can be alleviated if one could identify a set of stably expressed genes whose expression levels are stable across different experimental conditions. This motivates us to identify such a set of genes by exploring a large number of existing RNA-Seq data sets.

In microarray studies, there have been many attempts to find reference genes for normalization. Traditionally, the *house-keeping genes* are used as reference genes. However, a number of works have shown that house-keeping genes are not necessarily stably expressed according to numerical stability measures (see, for example, [20, 50]). Another choice, the *spike-in genes*, is not reliable for normalization due to the same issue [87]. A popular approach has been to search from large sets of experiments for reference genes [20, 21, 33, 40, 96] whose expression stability are evaluated by some numerical stability measure. Validation experiments (e.g. reverse transcription-PCR) show that reference genes identified by numerical methods generally outperform house-keeping genes or spike-in genes in terms of expression stability [20, 46]. We will

follow the strategy of quantifying gene expression stability by numerical measures and identify stably expressed genes.

Identifying stably expressed genes not only helps count normalization, but also improves interpretability and comparability of RNA-Seq experiments in integrative analysis. Since genes are measured by relative frequencies, we argue that DE is a relative concept: when a normalization procedure is applied to a single data set, it effectively uses an implicit reference set of genes. Furthermore, making the reference set explicit will be beneficial during DE analysis, because often times biologists compare results from one experiment to those from others experiments whose data are publicly available.

1.1.3.2 Estimating correlations of test statistics

Differential expression analysis involves simultaneous hypothesis testing for an ensemble of genes. One common feature of such procedure is that a test statistic summarizing the magnitude of DE is calculated for each gene, and then all the test statistics are pooled together and treated as a sample from which to estimate sampling distribution of test statistics. However, since the expression levels are correlated, the test statistics calculated from the expression levels are also correlated [8, 27, 112]. Without replicating the experiment, there's no way to obtain the correlation structure of test statistics because only a single sample of test statistics is available.

To construct correlation structure for test statistics, the sample correlations (after any treatment effects removed) are often used. Efron [27] approximates the distribution of test statistics correlation by sample correlation, and uses this distribution to derive the *false discovery proportion* in the multiple hypothesis testing procedure. For gene set test procedures that account for correlation among test statistics, Barry et al. [8], Wu and Smyth [112] and Yaari et al. [114] also use sample correlation to summarize test statistics correlation. However, it has only shown by simulation that correlations among test statistics are almost the same as those among samples for

test statistics taking specific forms [8, 27]. We will explore, from a theoretical point of view, the relationship between test statistics correlations and sample correlations in Chapter 3.

1.1.3.3 Adjusting for correlations in competitive gene set test

Competitive gene set test compares DE genes in the set against those in its complementary set. A number of statistical methodologies have been developed for this purpose (literature reviews can be found in [47, 54, 76]). Broadly speaking, all of the competitive gene set tests fall into two categories based on whether they assume independence of expression profiles among genes. In earlier literatures, the inter-gene correlations were not taken care of in the enrichment analysis procedure, such as SigPathway [102], PAGE [55], MRGSE [75] or the 2×2 contingency-table-based tests [1, 48, 115]. However, it has been argued that such test procedures will result in inflated type I error [29, 34, 37, 112, 114], as genes within a gene set are often co-expressed and function together.

Several approaches have been proposed to address inter-gene correlation problems in competitive gene set test. One attempt is to evaluate the significance of the test set by permuting sample labels [29, 34, 100]. Sample permutation does not require an explicit understanding of the underlying correlation structure among genes, and is therefore supposed to protect the test against such correlations. One very famous example of this kind is the *gene set enrichment analysis* (GSEA) procedure [100]. Yet, sample permutation method has been criticized for several reasons: first, it cannot be applied to experiments having small number of biological replicates (e.g., three samples each for a two-group comparison, which is common in RNA-Seq experiments); second, it is computationally intensive since tens of thousands of DE tests are involved in each permutation; third, and most importantly, it implicitly alters the null hypothesis being tested and makes the null and alternative difficult to be characterized [37, 54, 112]. Another attempt has been to incorporate the inter-gene

correlations into the formulation of gene set test procedure [112, 114]. CAMERA [112] estimates a *variance inflation factor* (VIF) from sample correlation (after the treatment effects removed), and then includes it in its test statistic to assess the significance of the gene set. The same VIF has also been used by QuSAGE [114] to adjust for inter-gene correlations. However, accurate estimation of VIF relies on the assumption that correlation between any two gene-level statistics are almost the same as correlation between their corresponding expression levels. In Chapter 3, we will demonstrate that this assumption is easily violated when differentially expressed genes are present, and as a remedy, we will propose a new gene set test procedure in Chapter 4.

1.2. Statistical Methods

We have mentioned earlier that RNA-Seq data are essentially present in the form of count matrices. Therefore it might not be appropriate to impose normal distribution on gene expression profiles, especially when the sample size is small. Generalized linear models (GLMs) are a natural choice for analyzing RNA-Seq data. In addition, to account for random terms in biological experiments, GLMs are sometimes extended to generalized linear mixed models (GLMMs). In this section, we will first describe the formulation GLMMs, and then review commonly used methods for parameter estimation under this framework.

1.2.1. Generalized linear mixed models

GLMMs are a natural generalization of classical linear models. To illustrate this point, we will begin with classical linear models, and discuss how to generalize them to linear mixed models and then to GLMMs by relaxing different layers of assumptions.

1.2.1.1 Classical linear models

In a classical linear model, a vector $\boldsymbol{y}$ of n observations is assumed to be a realization of random variable $\boldsymbol{Y}$ whose components are identically distributed with mean $\boldsymbol{\mu}$. The systematic part of this model is a specification of the mean $\boldsymbol{\mu}$ over a few unknown parameters [72]. In the context of classical linear models, the mean is a function of p covariates $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$,

$$\boldsymbol{\mu} = \beta_0 + \sum_{i=1}^{p} \beta_i \boldsymbol{X}_i \tag{1}$$

where β's are unknown parameters and need to be estimated from data. For jth observation Y_j, we specify ϵ_j, a random term, to allow for measurement error. Assuming a linear relationship between response Y_j and predictors $(x_{1j}, \ldots, x_{pj})$, we present the linear model

$$Y_j = \beta_0 + \beta_1 x_{1j} + \ldots + \beta_p x_{pj} + \epsilon_j \tag{2}$$

It is often required that ϵ_i's meet *Gauss-Markov* assumption,

$$E(\epsilon_i) = 0,\ \mathrm{Var}[\epsilon_i] = \sigma^2 < \infty,\ \mathrm{Cov}[\epsilon_i, \epsilon_j] = 0, \forall i \neq j. \tag{3}$$

In practice, the error term is frequently, if not always, assumed to be normally distributed,

$$\boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}). \tag{4}$$

1.2.1.2 Linear mixed models

The Gauss-Markov assumption in equation (3) is vulnerable in practice, for example, nonconstant variance, or correlated data where $\mathrm{Cov}[\epsilon_i, \epsilon_j] \neq 0$. Equation (2) in either case, without loss of generality, can be expressed in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ E[\boldsymbol{\epsilon}] = \boldsymbol{0},\ \mathrm{Cov}[\boldsymbol{\epsilon}] = \boldsymbol{V} \tag{5}$$

where $\boldsymbol{V}$ is a known positive definite matrix. Let $\boldsymbol{Y}^* = \boldsymbol{V}^{-1/2}\boldsymbol{Y} = \boldsymbol{V}^{-1/2}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{V}^{-1/2}\boldsymbol{\epsilon}$. It follows that $\mathrm{Cov}(\boldsymbol{Y}^*) = \boldsymbol{I}$ and the techniques in classical linear models are readily applicable to estimate $\boldsymbol{\beta}$. However, this method relies on the assumption that $\boldsymbol{V}$ is known which is rarely, if ever, given. On the other hand, the structure of $\boldsymbol{V}$, which depends on experiment setup, can often be specified by a few unknown parameters.

Nonindependence can occur in the form of serial correlation or cluster correlation [86, chapter 17]. Serial correlation usually exists in experiments with repeated measurements—multiple measurements taken from a response variable on the same experimental unit. Several covariance structures are available for implementation (for more details, see Littell et al. [66, chapter 5]). Cluster correlation is present when measurements of a response variable are grouped in some way. In many situations, the covariance of cluster correlated data can be specified using an extension of standard linear model by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z_1}\boldsymbol{u_1} + \cdots + \boldsymbol{Z_q}\boldsymbol{u_q} + \boldsymbol{\epsilon} \tag{6}$$

equation (6) differs from equation (5) only in the $\boldsymbol{Z_i}\boldsymbol{u_i}$ terms, which is the key part of *linear mixed models*. The $\boldsymbol{Z}_i$ are known $n \times p_i$ full rank matrices, usually used to specify membership of predictors in various subgroups. The most important innovation in this model is that instead of estimating $\boldsymbol{u}_i$'s as fixed parameters, we assume them to be unknown random quantities, and $E[\boldsymbol{u}_i] = 0$, $\mathrm{Cov}[\boldsymbol{u}_i] = \sigma_i^2 \boldsymbol{I}_{p_i}$ for $i = 1, \ldots, q$. It is, in many cases, reasonable to require that $\boldsymbol{u}_i$ are mutually independent, and that $\boldsymbol{u}_i$ is independent of $\boldsymbol{\epsilon}$ for $i = 1, \ldots, q$. If we further impose normal distribution on the random terms and errors, then equation (6) can be casted in a Bayesian framework,

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q &\sim N_n(\boldsymbol{X}\boldsymbol{\beta} + \sum_{i=1}^{q} \boldsymbol{Z_i}\boldsymbol{u_i}, \sigma^2 \boldsymbol{I}_n), \\ \boldsymbol{u}_i &\sim N_{p_i}(0, \sigma_i^2 \boldsymbol{I}_{p_i}). \end{aligned} \tag{7}$$

The modeling issues are: (a) estimation of variance components σ_i^2 and σ^2; (b) estima-

tion of random effects u_i if needed. For the variance component estimation, there are primarily three approaches: (i) procedures based on expected mean squares from analysis of variance (ANOVA); (ii) maximum likelihood (ML); and (iii) restricted/residual maximum likelihood (REML). For more details, see Littell et al. [66, Chapter 1].

1.2.1.3 Generalized linear models

We can take a different perspective of classical linear models by arranging equations (1)–(3) into three parts, following the notations of McCullagh and Nelder [72, Chapter 2],

(i) the *random component* Y_j has constant variance σ^2 and $E[Y_j] = \mu_j$.

(ii) the *systematic component*—the linear predictor η_j is modeled by covariates $\boldsymbol{x}_j =: x_{1j}, \ldots, x_{pj}$,

$$\eta_j = \sum_{i=1}^{p} \beta_i x_{ij} = \boldsymbol{x_j}\boldsymbol{\beta}. \tag{8}$$

(iii) the *link function* relates the random components and the systematic components by

$$\eta_j = g(\mu_j). \tag{9}$$

The classical linear models fits within this framework if we assume that the random components Y_j's are independent and normally distributed, and that the link function is identity (i.e., $g(\mu_j) = \mu_j$).

We can extend part (i)—by allowing Y_j to come from an exponential family (e.g., Poisson, Gamma or Binomial distribution), and part (iii)— by requiring the link function to be monotonic differentiable (e.g., $g(\mu_j) = \log \mu_j$). These two extensions lead to the *generalized linear models* (GLMs), a framework that is especially suitable when a normal distribution is no longer appropriate to be assumed on the response.

1.2.1.4 Generalized linear mixed models

Generalized linear mixed models (GLMMs) is a further extension of GLMs that incorporates random components into part (ii), represented in a matrix notation

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{i=1}^{q} \boldsymbol{Z_i}\boldsymbol{u_i} \tag{10}$$

where $\boldsymbol{Z}_i$ and $\boldsymbol{u}_i$ are specified in equation (6).

To formally present GLMMs, we start with the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{u}$. It is typical to assume that vector $\boldsymbol{y}$ consists of conditionally independent elements, each coming from the exponential family (or similar to the exponential family),

$$\begin{aligned} y_j|\boldsymbol{u} &\sim \text{ indep. } f_{Y_j|\boldsymbol{u}}(y_j|\boldsymbol{u}), \\ f_{Y_j|u}(y_j;\theta,\phi|\boldsymbol{u}) &= \exp\left[\frac{y_j\theta_j - b(\theta_j)}{a_j(\phi)} + c(y_i,\phi)\right]. \end{aligned} \tag{11}$$

It can be verified that the conditional mean of y_j is related to θ_j in equation (11) by the identity $\mu_j = \partial b(\theta_j)/\partial\theta_j$. The transformation of the mean allows us to model the fixed and the random factors by a linear model

$$\begin{aligned} E[y_j|\boldsymbol{u}] &= \mu_j, \\ g(\mu_j) &= \eta_j = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{u}. \end{aligned} \tag{12}$$

Finally, we assign a distribution to the random effects

$$\boldsymbol{U} \sim \phi_{\boldsymbol{U}}(\boldsymbol{u}), \tag{13}$$

which completes the specification of GLMMs. It is often, if not always, assumed that $\boldsymbol{u}$ come from a normal distribution.

1.2.1.5 An example—Poisson log-linear mixed-effect model

We will illustrate one specific type of GLMMs—the Poisson log-linear mixed-effect model in the context of RNA-seq experiments. Suppose we have RNA-Seq expression profiles (in the form of counts) randomly selected from three experiments conducted in three different labs. For each experiment, there are two treatments and two biological replicates for each treatment. We are not interested in the specific levels of treatment, but focus more on the overall variation of treatments. In this sense, the treatment effects are also considered as random. For a single gene, let $Y_{jkl} \sim \text{Poisson}(\mu_{jkl})$ be the read count for jth biological sample from kth treatment of lth experiment. The link function $\eta_{jkl} = \log(\mu_{jkl})$ relates the mean μ_{jkl} to the linear predictors by equation (12)

$$\log(\mu_{jkl}) = \log(N_{jkl}R_{jkl}) + \xi + a_j + b_{k(j)} + \epsilon_{jkl}, \tag{14}$$

where $N_{jkl}R_{jkl}$ is the normalized library size (total number of read counts mapped to the genome), $j = 1, 2, 3$, $k = 1, 2$ and $l = 1, 2$; the random terms $a_j \sim N(0, \sigma_1^2), b_{k(j)} \sim N(0, \sigma_2^2)$ and $\epsilon_{jkl} \sim N(0, \sigma_0^2)$ represent the experimental, treatment and sample effects respectively, and are mutually independent. If the observations are sorted by experiment and then by treatment nested in experiment, we can present the model in the form of equation (10), with $\boldsymbol{\beta} = (\log[N_{111}R_{111}] + \xi, \ldots, \log[N_{223}R_{223}] + \xi)^T$, $\boldsymbol{u} =$

$(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\epsilon})$ and

$$q = 2, \boldsymbol{X} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \boldsymbol{Z}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \boldsymbol{Z}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{Z}_3 = \boldsymbol{I}_{12}.$$

Then it follows that

$$\boldsymbol{\Sigma} = \sigma_1^2 \boldsymbol{Z_1} \boldsymbol{Z_1'} + \sigma_2^2 \boldsymbol{Z_2} \boldsymbol{Z_2'} + \sigma_0^2 \boldsymbol{I}_{12} = \begin{bmatrix} \boldsymbol{\Sigma}_d & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Sigma}_d & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{\Sigma}_d \end{bmatrix},$$

where $\boldsymbol{O}$ is a 4×4 matrix of 0 and

$$\boldsymbol{\Sigma}_d = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 + \sigma_0^2 & \sigma_1^2 + \sigma_2^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 + \sigma_2^2 & \sigma_1^2 + \sigma_2^2 + \sigma_0^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma_2^2 + \sigma_3^2 & \sigma_1^2 + \sigma_2^2 \\ \sigma_1^2 & \sigma_2^2 & \sigma_1^2 + \sigma_2^2 & \sigma_1^2 + \sigma_2^2 + \sigma_0^2 \end{bmatrix}$$

The challenge due to the complexity of GLMM is the estimation of parameters. In the next section, we will summarize current available methods for estimating parameters and variance components.

1.2.2. Estimation of generalized linear mixed models

There are three general approaches for estimating parameters under GLMM settings [77, Chapter 7]: (i) using numerical method to approximate the integrals for the likelihood functions and obtaining the estimating equations; (ii) linearization of the conditional mean and then iteratively applying linear mixed model techniques to the approximated model; (iii) Bayesian approach.

In the following discussion, we assume conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{u}$ is $f_Y(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u})$, the link function is $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, and $\boldsymbol{\eta}$ relates the covariates by equation (12). We also assume that the random term $\boldsymbol{u}$ have some distribution $\boldsymbol{U} \sim \phi_{\boldsymbol{U}}(\boldsymbol{u}|\boldsymbol{\Sigma})$.

1.2.2.1 Likelihood function approach

It is straightforward to write down the likelihood function of $\boldsymbol{Y}$ by first obtaining the joint likelihood of $(\boldsymbol{Y}, \boldsymbol{u})$ and then integrating out the random term $\boldsymbol{u}$,

$$L(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u})\phi(\boldsymbol{u}|\boldsymbol{\Sigma})d\boldsymbol{u}. \tag{15}$$

A major challenge in estimating GLMMs is the integration in equation (15) over the n-dimensional distribution of $\boldsymbol{u}$. Numerical approximation are usually used in evaluating the integral. In this part we will discuss the *Gauss-Hermite* (GH) quadrature which is recognized as a higher order Laplace approximation [67]. Gauss-Hermite quadrature is used for integrals of the form $\int_{-\infty}^{\infty} f(x)e^{-x^2}dx$ that can be approximated by a weighted sum of $f(x)$:

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx \approx \sum_{i=1}^{m} w_i f(x_i) \tag{16}$$

In equation (16), x_i's are the zeros of mth order Hermite polynomial

$$H_m(x) = (-1)^m \exp(\frac{x^2}{2})\frac{d^m}{dx^m}\exp(-\frac{x^2}{2})$$

and w_i are the corresponding weights. For a Hermite polynomial of degree m, x_i and w_i can be calculated as

$$x_i = i\text{th zero of } H_m(x), \quad w_i = \frac{2^{m-1}m!\sqrt{\pi}}{m^2[H_{m-1}(x_i)]^2}. \tag{17}$$

Equation (16) gives the exact numerical value for all polynomials up to degree of $2m-1$. An improved version of the regular Gauss-Hermite quadrature is to center and scale the quadrature points by the empirical Bayes estimate of the random effects and the Hessian matrix from the Bayes estimate suboptimization [67]. This procedure is called *Adaptive Gauss-Hermite* (AGH) quadrature [79].

The AGH quadrature starts with maximizing the integrand $h(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) := f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u})\phi(\boldsymbol{u}|\boldsymbol{\Sigma})$ in equation (15) with respect to the random term $\boldsymbol{u}$. The resulting estimate $\hat{\boldsymbol{u}}^{(n)}$ at iteration n is the joint posterior modes for the random effects. Because $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are unknown, they are replaced by the current estimates $\hat{\boldsymbol{\beta}}^{(n)}$ and $\hat{\boldsymbol{\Sigma}}^{(n)}$. The Hessian matrix $\hat{\boldsymbol{H}}^{(n)}$ can be obtained by evaluating the second order partial derivatives of $\log(h(\boldsymbol{u}|\boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\Sigma}}^{(n)}))$ at $\hat{\boldsymbol{u}}^{(n)}$. Consequently, $\hat{\boldsymbol{\Omega}}^{(n)} = -\hat{\boldsymbol{H}}^{(n)}$ is the estimated covariance matrix for the random effects posterior modes. It follows from equation (15) that for the ith cluster

$$L(\boldsymbol{Y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{u})\phi(\boldsymbol{u}|\boldsymbol{\Sigma})d\boldsymbol{u} = \int \frac{f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{u})\phi(\boldsymbol{u}|\boldsymbol{\Sigma})}{\phi(\boldsymbol{u}|\hat{\boldsymbol{u}}^{(n)}, \hat{\boldsymbol{\Omega}}^{(n)})}\phi(\boldsymbol{u}|\hat{\boldsymbol{u}}^{(n)}, \hat{\boldsymbol{\Omega}}^{(n)})d\boldsymbol{u} \tag{18}$$

Let m be the number of quadrature points (i.e., the order of the Hermite polynomial) in each dimension for each random effect term, and Q the number of random effects. If $\boldsymbol{x} = (x_1, \ldots, x_m)$ are the nodes for standard Gauss-Hermite quadrature, and $\boldsymbol{x}_j^* = (x_{j_1}, \ldots, x_{j_Q})$ is a point on the Q dimensional quadrature grid, then the centered and scaled nodes are

$$\boldsymbol{a}_j^* = \hat{\boldsymbol{u}}^{(n)} + \sqrt{2}[\hat{\boldsymbol{\Omega}}^{(n)}]^{1/2}\boldsymbol{x}_j^* \tag{19}$$

The centered and scaled nodes, along with the Gauss-Hermite quadrature weights

$\boldsymbol{w} = (w_1, \ldots, w_m)$ are used to construct the Q dimensional integral of equation (18), approximated by

$$\begin{aligned} L(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}) &\approx \sum_{j_1=1}^{m} \cdots \sum_{j_Q=1}^{m} \frac{f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{a}_j^*)\phi(\boldsymbol{a}_j^*|\boldsymbol{\Sigma})}{\phi(\boldsymbol{a}_j^*|\hat{\boldsymbol{u}}^{(n)}, \hat{\boldsymbol{\Omega}}^{(n)})} w_{j_1} \cdots w_{j_Q} \\ &= (2)^{Q/2} |\hat{\boldsymbol{\Omega}}^{(n)}|^{1/2} \sum_{j_1=1}^{m} \cdots \sum_{j_Q=1}^{m} \left[f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{a}_j^*)\phi(\boldsymbol{a}_j^*|\boldsymbol{\Sigma}) \prod_{k=1}^{Q} w_{jk} \exp(x_{jk}^2) \right]. \end{aligned} \tag{20}$$

Thus the multidimensional unbounded integral is approximated by a finite summations. Now that the likelihood has the form of equation (20), a number of numerical methods (e.g. Newton-Raphson or Fisher's scoring) can be used to estimate $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$.

It should be noted, however, as the number of dimension Q increases, the computational burden for approximating equation (20) grows exponentially since the total number of nodes is m^Q. Therefore it is difficult to implement AGH procedure with more than three random effects [12].

1.2.2.2 Estimation based on linearization

Under GLMM framework, we have some conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{u}$. Without loss of generality, we assume

$$\begin{aligned} E[\boldsymbol{Y}|\boldsymbol{u}] &= \boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}), \\ \mathrm{Var}[\boldsymbol{Y}|\boldsymbol{u}] &= \boldsymbol{S}, \end{aligned} \tag{21}$$

where $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$. The linearization is done by Taylor expansion of equation (21) about estimates $\boldsymbol{\eta}$. Two approaches proposed by Breslow and Clayton [14]—the *penalized quasi-likelihood* (PQL) and the *marginal quasi-likelihood* (MQL)—may be used for this purpose.

Penalized Quasi-likelihood The PQL procedure uses a first order Taylor expansion of $\boldsymbol{\beta}$ and $\boldsymbol{u}$, at $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}$, respectively

$$g^{-1}(\boldsymbol{\eta}) \approx g^{-1}(\hat{\boldsymbol{\eta}}) + \tilde{\boldsymbol{\Omega}}_{PQL}(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}), \tag{22}$$

where $\tilde{\boldsymbol{\Omega}}_{PQL}$ is an $n \times n$ diagonal matrix whose (i,i) entry is $\partial g^{-1}(\boldsymbol{\eta}_i)/\partial \boldsymbol{\eta}_i$ evaluated at $\tilde{\boldsymbol{\eta}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{u}}$. Multiplying both sides by $\tilde{\boldsymbol{\Omega}}_{PQL}^{-1}$, equation (22) can be rearranged as

$$\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} \approx \tilde{\boldsymbol{\Omega}}_{PQL}^{-1}[g^{-1}(\boldsymbol{\eta}) - g^{-1}(\tilde{\boldsymbol{\eta}})] + \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{u}}. \tag{23}$$

Note that the right hand side of equation (23) is just the expected value, conditioning on $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}$, of the pseudo-response

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{\Omega}}_{PQL}^{-1}[\boldsymbol{Y} - g^{-1}(\tilde{\boldsymbol{\eta}})] + \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{u}}, \tag{24}$$

whose variance-covariance matrix given $\boldsymbol{u}$ is

$$\mathrm{Var}[\tilde{\boldsymbol{Y}}|\boldsymbol{u}] = \tilde{\boldsymbol{\Omega}}_{PQL}^{-1}\mathrm{Var}[\boldsymbol{Y}|\boldsymbol{u}]\tilde{\boldsymbol{\Omega}}_{PQL}^{-1} = \tilde{\boldsymbol{\Omega}}_{PQL}^{-1}\boldsymbol{S}\tilde{\boldsymbol{\Omega}}_{PQL}^{-1}. \tag{25}$$

Then we can consider the model

$$\tilde{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \tag{26}$$

which is a linear mixed model with pseudo response $\tilde{\boldsymbol{Y}}$ with covariance matrix

$$\boldsymbol{W} = \mathrm{Var}[\tilde{\boldsymbol{Y}}|\boldsymbol{u}] = \boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}' + \tilde{\boldsymbol{\Omega}}_{PQL}^{-1}\boldsymbol{S}\tilde{\boldsymbol{\Omega}}_{PQL}^{-1}. \tag{27}$$

Model (26) has exactly the same form as the linear mixed models (see Section 1.2.1.2), except that an estimate of $(\boldsymbol{\beta}, \boldsymbol{u})$ is needed for calculating the pseudo-response $\tilde{\boldsymbol{Y}}$ in equation (24). An iterative procedure can be used to estimate the parameters in

model (26) by substituting raw data $\boldsymbol{y}$ for $\tilde{\boldsymbol{y}}$ and identity matrix $\boldsymbol{I}$ for $\boldsymbol{S}$ as starting values. Techniques for fitting LMM such as REML can be readily applied to estimate variance components $\boldsymbol{\Sigma}$, upon which $\hat{\boldsymbol{W}}$ is calculated. The estimate for $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\hat{\boldsymbol{W}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}^{-1}\boldsymbol{X}\tilde{\boldsymbol{y}}, \tag{28}$$

and the estimate for random effect is

$$\hat{\boldsymbol{u}} = \hat{\boldsymbol{\Sigma}}\boldsymbol{Z}\hat{\boldsymbol{W}}^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{29}$$

Then the pseudo-response is updated and the procedure is repeated until convergence is reached for fixed effects and variance components. Note that equation (29) estimates a vector of random effect. For this reason, PQL is also referred to as *subject-specific* estimate procedure.

Marginal Quasi-likelihood One of the motivation for MQL is that usually one is more interested in estimating the marginal mean of the response than estimating the conditional mean as is done by equation (29) in PQL. Since $E[\boldsymbol{\eta}|\boldsymbol{u}] = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$, the unconditional mean is $E[\boldsymbol{\eta}] = E[E(\boldsymbol{\eta}|\boldsymbol{u})] = \boldsymbol{X}\boldsymbol{\beta}$. A first-order Taylor expansion of $E[\boldsymbol{Y}|\boldsymbol{u}]$ about $\boldsymbol{X}\boldsymbol{\beta}$ is given by

$$E[\boldsymbol{Y}|\boldsymbol{u}] = g^{-1}(\boldsymbol{\eta}) \approx g^{-1}(\boldsymbol{X}\boldsymbol{\beta}) + \tilde{\boldsymbol{\Omega}}_{MQL}(\boldsymbol{\eta} - \boldsymbol{X}\boldsymbol{\beta}) \tag{30}$$

where $\tilde{\boldsymbol{\Omega}}_{MQL}$ is evaluated at $\boldsymbol{X}\boldsymbol{\beta}$ (recall that for PQL, $\tilde{\boldsymbol{\Omega}}_{PQL}$ is evaluated at $\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}$). The unconditional expected value of $\boldsymbol{Y}$ is approximately $g^{-1}(\boldsymbol{X}\boldsymbol{\beta})$ by equation (30). The variance of $\boldsymbol{Y}$ can then be derived from the relation $\mathrm{Var}(\boldsymbol{Y}) = E[\mathrm{Var}(\boldsymbol{Y}|\boldsymbol{u})] + \mathrm{Var}[E(\boldsymbol{Y}|\boldsymbol{u})]$, which yields

$$\mathrm{Var}[\boldsymbol{Y}] = \tilde{\boldsymbol{\Omega}}_{MQL}\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}'\tilde{\boldsymbol{\Omega}}'_{MQL} + \boldsymbol{S}_{\boldsymbol{\eta}_0}. \tag{31}$$

A linearization performed at $\boldsymbol{\eta}_0 = \boldsymbol{X}\boldsymbol{\beta}_0$ leads to

$$g^{-1}(\boldsymbol{\eta}) \approx g^{-1}(\boldsymbol{X}\boldsymbol{\beta_0}) + \tilde{\boldsymbol{\Omega}}_{MQL}(\boldsymbol{\eta} - \boldsymbol{X}\boldsymbol{\beta}_0), \tag{32}$$

and multiplying both sides by $\tilde{\boldsymbol{\Omega}}^{-1}_{MQL}$, equation (32) then can be arranged to

$$\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} \approx \tilde{\boldsymbol{\Omega}}^{-1}_{MQL}[g^{-1}(\boldsymbol{\eta}) - g^{-1}(\boldsymbol{\eta}_0)] + \boldsymbol{X}\boldsymbol{\beta}_0. \tag{33}$$

Defining the pseudo-response $\tilde{\boldsymbol{Y}}_{MQL}$ as

$$\tilde{\boldsymbol{Y}}_{MQL} = \tilde{\boldsymbol{\Omega}}^{-1}_{MQL}[\boldsymbol{Y} - g^{-1}(\boldsymbol{\eta}_0)] + \boldsymbol{X}\boldsymbol{\beta}_0, \tag{34}$$

we next consider the linear mixed model

$$\tilde{\boldsymbol{Y}}_{MQL} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon},$$

where $\mathrm{Var}(\boldsymbol{\epsilon})$ is given by equation (31). The estimating procedure for fixed effect parameter $\boldsymbol{\beta}$ and variance component $\boldsymbol{\Sigma}$ is the same as those in the PQL approach. Note that the pseudo-response is not a function of $\boldsymbol{u}$ any more, so updating this quantity does not require calculating the random effects $\boldsymbol{u}$. Accordingly, the MQL approach is also referred to as *population-averaged* estimate approach.

Breslow and Lin [15] and Pinheiro and Chao [80] point out that PQL approach may lead to asymptotically biased estimates and hence to inconsistency. It is not recommended to use simple PQL method in practice.

1.2.2.3 Bayes approach

As mentioned earlier, for models with higher dimensional integrals, it is not practical to evaluate the likelihood function by AGH procedure. For mixed models, a typical strategy is to treat the random effects to be missing data. Following this idea, the

problem of estimating variance components associated with random effects can be simplified. Denote the *complete data* as $\boldsymbol{v} = (\boldsymbol{y}, \boldsymbol{u})$, the log-likelihood of $\boldsymbol{v}$ can be expressed as

$$\log \pi(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{v}) = \log f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}) + \log \phi(\boldsymbol{u}|\boldsymbol{\Sigma}) \tag{35}$$

The optimal solution for parameters in equation (35) can be obtained by *expectation-maximization* (EM) algorithm. The EM algorithm consists of two steps, readily implemented as follows:

1. **E-Step**. At $(k+1)$th iteration given $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$, calculate

$$\begin{aligned} E_{\boldsymbol{\beta}^{(k)}}[\log f(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{v})|\boldsymbol{y}] &= Q_1(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)}), \\ E_{\boldsymbol{\Sigma}^{(k)}}[\log \phi(\boldsymbol{\Sigma}|\boldsymbol{v})|\boldsymbol{y}] &= Q_2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^{(k)}). \end{aligned} \tag{36}$$

2. **M-Step**. Maximize Q_1 and Q_2 to update $\boldsymbol{\beta}^{(k+1)}$ and $\boldsymbol{\Sigma}^{(k+1)}$.

The **E** and **M** steps are alternated until convergence is reached. Unfortunately, the expectations in equation (36) cannot be computed in closed form for GLMMs. However, they may be approximated by *Markov chain Monte Carlo* (MCMC). In light of this, McCulloch [73] developed a Monte Carlo EM (MCEM) algorithm.

The Metropolis-Hastings algorithm is used for drawing samples from difficult-to-calculate density functions. For Metropolis algorithm, a proposal distribution $g(\boldsymbol{u})$ is selected, from which an initial value of $\boldsymbol{u}$ is drawn. The new candidate value $\boldsymbol{u}' = (u_1, u_2, \ldots, u_{k-1}, u_k', u_{k+1}, \ldots, u_Q)$, which has all elements the same as previous values expect the kth, is accepted (as opposed to keeping the previous value) with probability

$$A_k(\boldsymbol{u}', \boldsymbol{u}) = \min\left\{1, \ \frac{f(\boldsymbol{u}'|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) g(\boldsymbol{u})}{f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) g(\boldsymbol{u}')}\right\}. \tag{37}$$

If we choose $g(\boldsymbol{u}) = \phi(\boldsymbol{u}|\boldsymbol{\Sigma})$, the ratio term in equation (37) can be simplified to

$$\begin{aligned}
&\frac{f(\boldsymbol{u}'|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\Sigma})g(\boldsymbol{u})}{f(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\Sigma})g(\boldsymbol{u}')} \\
&= \left[\frac{f(\boldsymbol{u}',\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\Sigma})}{f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\Sigma})}\phi(\boldsymbol{u}|\boldsymbol{\Sigma})\right] \Bigg/ \left[\frac{f(\boldsymbol{u},\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\Sigma})}{f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\Sigma})}\phi(\boldsymbol{u}'|\boldsymbol{\Sigma})\right] \\
&= \frac{f(\boldsymbol{y}|\boldsymbol{u}',\boldsymbol{\beta},\boldsymbol{\Sigma})\phi(\boldsymbol{u}'|\boldsymbol{\Sigma})\phi(\boldsymbol{u}|\boldsymbol{\Sigma})}{f(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{\beta},\boldsymbol{\Sigma})\phi(\boldsymbol{u}|\boldsymbol{\Sigma})\phi(\boldsymbol{u}'|\boldsymbol{\Sigma})} \\
&= \frac{f(\boldsymbol{y}|\boldsymbol{u}',\boldsymbol{\beta},\boldsymbol{\Sigma})}{f(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{\beta},\boldsymbol{\Sigma})}
\end{aligned} \tag{38}$$

The MCEM procedure combines the EM steps and Metropolis algorithm in estimating the fixed parameters and variance components, summarized as follows:

1. Choose the starting value of $\boldsymbol{\beta}^{(0)}, \boldsymbol{\Sigma}^{(0)}$. Set $b = 0$

2. Generate the sequence $\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}, \ldots, \boldsymbol{u}^{(B)}$ from the conditional distribution of $\boldsymbol{u}$ given y with Metropolis algorithm.

3. Maximize $\sum_{b=1}^{B} \log f(\boldsymbol{y}|\boldsymbol{u}^{(b)}, \boldsymbol{\beta})/B$ and $\sum_{b=1}^{B} \log \phi(\boldsymbol{u}^{(b)}|\boldsymbol{\Sigma})/B$ to obtain $\boldsymbol{\beta}^{(m+1)}$ and $\boldsymbol{\Sigma}^{(m+1)}$

4. Iterate between step 2 and 3 until convergence is reached.

This method can be easily extended to allow for multiple random effects. Yet the advantage comes at a price. A major drawback of MCEM is the computational intensity. First, the convergence of EM algorithm is usually very slow, especially at the neighborhood of maximum of marginal likelihood. Second, the chain in Metropolis algorithm has to run long enough for reliable estimation.

In the Bayes framework, there are other alternatives to estimate the parameters and variance components, for example, *Monte Carlo Newton-Raphson* (MCNR) [73] and MCMC [41].

1.2.2.4 Example of estimating parameters

We will demonstrate the estimating procedure with the Poisson log-linear mixed-effect model discussed in Section 1.2.1. The estimation procedure starts from the joint density function of $\boldsymbol{Y} = (Y_{jkl})'$ given $\boldsymbol{\mu} = (\mu_{jkl})'$,

$$f(\boldsymbol{y}|\boldsymbol{\mu}) = \prod_{j,k,l} f(y_{jkl}|\mu_{jkl}) = \prod_{j,k,l} \frac{[\mu_{jkl}]^{y_{jkl}} \exp(-\mu_{jkl})}{y_{jkl}!}. \tag{39}$$

A re-expression of equation (14) in matrix form gives

$$\log \boldsymbol{\mu} = \boldsymbol{X\beta} + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 b} + \boldsymbol{I}_{12}\boldsymbol{\epsilon}.$$

Therefore $\boldsymbol{\mu} \sim \log N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}_0 = \boldsymbol{\xi} + \log(\boldsymbol{NR})$ and $\boldsymbol{\Sigma} = \sigma_1^2 \boldsymbol{Z_1 Z_1'} + \sigma_2^2 \boldsymbol{Z_2 Z_2'} + \sigma_0^2 \boldsymbol{I}_{12}$. The density function of $\boldsymbol{\mu}$ is then

$$f(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \prod_{j,k,l} \mu_{jkl}^{-1} \cdot \frac{1}{\sqrt{(2\pi)^{12}|\boldsymbol{\Sigma}|}} \exp[-\frac{1}{2}(\log \boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\log \boldsymbol{\mu} - \boldsymbol{\mu}_0)]. \tag{40}$$

Since $Y_{jkl} \sim \text{Poisson}(\mu_{jkl})$, by combining equation (39) and (40), we obtain the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{\mu}$,

$$\begin{aligned} &f(\boldsymbol{y}, \boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) \\ &= \frac{1}{\sqrt{(2\pi)^{12}|\boldsymbol{\Sigma}|}} \exp[-\mathbf{1}^T \boldsymbol{\mu} - -\frac{1}{2}(\log \boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\log \boldsymbol{\mu} - \boldsymbol{\mu}_0)] \prod_{jkl} \frac{[\mu_{jkl}]^{y_{jkl}-1}}{y_{jkl}!}. \end{aligned} \tag{41}$$

Therefore we can obtain the likelihood function of or the marginal distribution of $\boldsymbol{Y}$ by integrating out the random components $\boldsymbol{u}$,

$$L(\xi, \sigma_1^2, \sigma_2^2, \sigma_0^2|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{\Sigma}) = \int_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{\epsilon}} f(\boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\epsilon}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) d\boldsymbol{a} d\boldsymbol{b} d\boldsymbol{\epsilon}. \tag{42}$$

The integral in equation (42) can be approximated by adaptive Gaussian-Hermite (AGH) quadrature or MCMC. For AGH quadrature, we first approximate the like-

lihood by equation (20) and then estimate $\boldsymbol{\theta} = (\xi, \sigma_0^2, \sigma_1^2, \sigma_2^2)'$ by maximizing the resulting likelihood. The R package `lme4` [9] has an inbuilt function `glmer()` for this purpose. The MCMC procedure has also been implemented in several packages, such as the `Rstan` [97] and the `MCMCPack` [71].

1.3. Multiple hypothesis testing

Multiple hypothesis testing procedures deal with type I error rates in a family of tests. The problems arise when we consider a set of statistical inference simultaneously. For each of the individual tests or confidence intervals, there is a type I error which can be controlled by the experimenter. If the family of tests contains one or more true null hypotheses, the probability of rejecting one or more of these true null increases.

While traditional multiple testing procedures focus on modest number of tests, a different set of techniques are needed for large-scale inference, in which tens or even hundreds of thousands of tests are performed simultaneously. For example, in genomics study, expression levels of 50,000 genes for each of 100 individuals can be measured using modern technologies such as microarray or RNA-Sequencing. In testing differential expression (DE), 50,000 tests need to be conducted against the null that there is no DE between treatment/control. This has brought new challenge to the field of multiple hypothesis testing. Benjamini and Hochberg [10] points out that the control of familywise error rate (FWER), i.e. the probability of making one or more false discovery in a set of tests, tends to have substantially less power.

False discovery rate (FDR), introduced by Benjamini and Hochberg [10], is the expected proportion of false positives among all significant calls (null rejected). FDR has been studied extensively ([11, 26**?** , 99] and more) over the past two decades. FDR is equivalent to FWER [10] when all hypotheses are true but smaller if there are some true discoveries to be made. We will focus our attention on FDR in this part.

Let m, m_0 and m_1 be the number of tests, true nulls and true alternatives respec-

tively. Let also F and T be the number of true nulls and true alternatives among S tests that are declared as significant. Table (**??**) shows the relation among them. The FDR is

	Called significance	Called not significant	Total
Null True	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
total	S	$m - S$	m

1.4. Disertation Objective

2. Identifying stably expressed genes from multiple RNA-Seq data sets

Abstract

We examined RNA-Seq data on 211 biological samples from 24 different experiments carried out by different labs and identified genes that are stably expressed across biological samples, treatment conditions, and experiments. We fit a Poisson log-linear mixed-effect model to the read counts for each gene and decomposed the total variance into between-sample, between-treatment and between-experiment variance components. Identifying stably expressed genes is useful for count normalization and differential expression analysis. The variance component analysis that we explore here is a first step towards understanding the sources and nature of the RNA-Seq count variation. When using a numerical measure to identify stably expressed genes, the outcome depends on multiple factors: the reference sample sets used, the technology used for measuring gene expression, and the specific numerical stability measure used. Since DE is measured by relative frequencies, we argue that DE is a relative concept. We advocate using an explicit reference gene set for count normalization to improve interpretability of DE results, and recommend using a common reference gene set when analyzing multiple RNA-Seq experiments to avoid potential inconsistent conclusions.

2.1. Introduction

RNA sequencing (RNA-Seq) has become the technology of choice for transcriptome profiling over the last few years. The exponential growth in RNA-Seq studies have produced a large amount of *Arabidopsis thaliana* (Arabidopsis) data under a variety of experimental/environmental conditions. It is only natural to begin exploring how the large amount of existing data sets can help the analysis of future data. In this paper, we discuss identifying stably expressed genes from multiple existing RNA-Seq data sets based on a numerical measure of stability. We envision that such identified stably expressed genes could be used as a reference set or prior information

for count normalization and differential expression (DE) analysis of future RNA-Seq data sets obtained from similar or comparable experiments. We also fit a random-effect model to the read counts for each gene and decompose the total variance to into between-sample, between-treatment and between-experiment variance components. The variance component analysis is a first step towards understanding the sources and nature of the RNA-Seq count variation. To illustrate our methods, we examined RNA-Seq data on 211 Arabidopsis samples from 24 different experiments carried out by different labs and identified genes that were stably expressed across biological samples, experimental or environmental conditions, and experiments (labs).

A reference set of stably-expressed genes will be useful for count normalization. A key task of RNA-Seq analysis is to detect DE genes under various experimental or environmental conditions. Count normalization is needed to adjust for differences in sequencing depths or library sizes (total numbers of mapped reads for each biological sample) due to chance variation in sample preparation. In DE analysis, gene expression levels are often estimated from relative read frequencies. For this reason, normalization is also needed to account for the fact that non-differentially expressing genes may exhibit an apparent reduction or increase in relative read frequencies due to the respective increased or decreased relative read frequencies of truly differentially expressing genes. Many existing normalization methods, such as the trimmed mean of M-values normalization method (TMM) [89] and Anders and Huber's normalization [3], assume that the majority of the genes within an experiment are not DE, and examine the sample distribution of the fold changes between samples. If the experiment condition can affect expression levels of more than half of the genes, many of the existing normalization methods may be unreliable [68, 110]. This difficulty could be alleviated if one could identify a set of stably expressed genes whose expression levels are known or expected to not vary much under different experimental conditions. Our idea is to identify such a reference set based on a large number of existing data sets.

Our basic intuition is that a numerical quantification of expression stability—

which typically measures certain aspects of RNA-Seq count variation—can be more reliably estimated by using more data sets. There is, however, a caveat to this idea: as pointed out by [30] and [46], universally stably expressed genes may not exist. [46] showed that a subset of stably expressed genes from a specific biological context may have more variability than other genes if examined across a broader range of samples and conditions. Many studies have shown that stably expressed genes are subject to change from one experiment to another due to different experimental protocols, different tissue types, or other varying conditions [85, 44]. The top 100 stably expressed genes in the Arabidopsis developmental series of Czechowski et al. [20] shared only 3 genes with the top 50 stably expressed genes identified from Arabidopsis seed samples by Dekkers et al. [21]. In this study, we try to balance generality and specificity by identifying different reference gene sets for different tissue types of Arabidopsis.

We can also consider that when a normalization method is applied to a single data set, it effectively specifies an implicit reference set of stably expressed genes (those genes that have the least variation after normalization). From this perspective, we can view commonly used normalization techniques as using an internally identified reference set of genes. In contrast, what we are proposing is that one could alternatively identify a reference set externally by looking at past data sets. The internally and externally identified reference gene sets will provide different contexts for the DE analysis: in other words, one can choose to answer different scientific questions by using different reference sets. In any case, we advocate making the reference set explicit during a DE analysis and using a common reference set when analyzing multiple datasets.

We want to clarify that having stable gene expression is not equivalent to maintaining a stable biological function. Often times, we may not understand the biological functions of genes with numerically stable expression measures. From an operational point of view, however, numerical stability is more tractable. In pre-genomic era, the so-called "*house-keeping genes*" were often considered to be candidate reference genes

for normalization [17, 5]. House-keeping genes are typically constitutive genes that maintain basic cellular function, and therefore are expected to express at relatively constant levels in non-pathological situations. However, many studies have shown that house-keeping genes are not necessarily stably expressed according to numerical measures (a review can be found in [50] and reference therein). For example, in the microarray analysis of Arabidopsis, [20] showed that traditional house-keeping genes such as ACT2, TUB6, EF-1α are not stably expressed, and thus not good reference genes for normalization. Spike-in genes have also been considered as reference genes for normalization, but [87] showed that spike-in genes are not necessarily stably expressed according numerical measures either.

In this paper, we identify stably expressed genes from RNA-Seq data sets based on a numerical measure—the sum of three variance components estimated from a mixed-effect model. For microarray data, there have been many efforts to numerically find stably expressed genes by quantifying the variation of measured expression levels across a large number of microarray data sets. For example, [5] used a linear mixed model to estimate the between-group and within-group variances from expression profiles of microarray experiments, and then quantified expression stability by combining the two variance components using a Bayesian formulation. [20] measured the expression stability of each gene using the coefficient of variation (CV). Genes with lower CVs are considered more stably expressed. By investigating 721 arrays under 323 conditions throughout development, [20] suggested stably expressed (reference) genes under different experimental conditions for Arabidopsis. [96], Dekkers et al. [21], Gur-Dedeoglu et al. [40], and Frericks and Esser [33] screened a large number of microarray data sets to identify stably expressed genes in human blood, Arabidopsis seed, breast tumor tissues, and mice respectively. Validation experiments [20, 21, 50, 96] showed that these genes are more stably expressed than traditional house-keeping genes.

Our vision is that identifying stably expressed genes is the first step towards inte-

grative analysis of multiple RNA-Seq experiments. It will help to answer fundamental questions related to comparability, reproducibility and replicability of RNA-Seq experiments.

2.2. Methods

In Section 2.2.1, we describe the steps for collecting and processing RNA-Seq data sets from Arabidopsis experiments. In Section 2.2.2, we discuss count normalization methods and how to apply them to a subset of stably expressed genes. In Section 2.2.3, we introduce the generalized linear mixed model (GLMM, McCulloch and Neuhaus 74) for estimating three variance components from RNA-Seq data: the *between-sample*, *between-treatment* and *between-experiment* variances. We define the *total variance* measure for expression stability as the sum of estimated variance components. In Section 2.2.4, we review the CV and M-value measures for gene expression stability.

2.2.1. RNA-Seq data collection and processing

2.2.1.1 Overview of the RNA-Seq data sets

We examined RNA-Seq data from 49 Arabidopsis experiments stored on the NCBI GEO repository (see more details below). After screening, we retained data from 211 biological samples in 24 experiments. To illustrate our methods for finding stably expressed genes, we divided the experiments into three groups: *the seedling group* contains 60 Arabidopsis seedling samples from 9 experiments; *the leaf group* contains 60 Arabidopsis leaf samples from 5 experiments; the *multi-tissue group* contains 91 samples from 10 experiments on multiple tissue types (shoot apical, root tip, primary root, inflorescences and siliques, hypocotyl, flower, carpels, aerial tissue, epidermis, seed). Table 1 summarizes the basic information about the three groups.

To find stably expressed genes in each group, we processed the raw sequencing data

Table 1: Summary statistics for the three groups of Arabidopsis samples.

Group	# experiments	# treatments	# samples	# genes
seedling	9	27	60	22207
leaf	5	28	60	20967
multi-tissue	10	39	91	23611

and summarized the results as count matrices of mapped RNA-Seq short reads (see details below). We removed genes with low mean numbers (less than 3) of mapped read counts for all experiments. Such genes tend to be more prone to sequencing noise, less interesting to biologists, and also cause convergence issues when fitting statistical models. Many other researchers (such as Anders et al. 4) recommend removing such genes before analyzing RNA-Seq data. The number of remaining genes in each group is also summarized in Table 1.

2.2.1.2 Details of the data processing steps

The *Gene Expression Omnibus* (GEO) repository at *National Center for Biotechnology Information* (NCBI, `http://www.ncbi.nlm.nih.gov/`) stores raw sequencing data from a large number of RNA-Seq experiments. For this study, we restrict our attention to Arabidopsis experiments satisfying the following conditions: 1. Ecotype = "Columbia" (we kept only the Columbia samples from experiments that compare Columbia samples to other ecotypes); 2. There are at least two treatments and 2 biological replicates for each treatment; 3. Library strategy= "RNA-Seq"; 4. Library source = "transcriptomic"; 5. Library selection= "cDNA"; 6. Library layout = "Single end"; 7. If there are repeated measurements over time, we choose samples from one time point.

We screened all the Arabidopsis experiments available from the NCBI GEO repository up to May 31, 2015 and downloaded raw RNA-Seq data (Sequence Read Archive files) from 49 experiments.

We assembled our own in-house pipeline to process all the raw RNA-Seq data:

align the raw RNA-Seq reads to the reference genome and summarize the read counts at the gene level. In the GEO repository, the mapped read counts are unavailable for some experiments and the available ones are from different processing pipelines. Our pipeline, implemented using the software R [83], is summarized as follows:

1. Convert the Sequence Read Archive (SRA) files to FASTQ files using the NCBI SRA Toolkit ([60], version 2.3.5-2).

2. Download the reference genome

   `Arabidopsis_thaliana.TAIR10.22.dna.toplevel.fa`

   from the *Ensembl plants FTP server* (`http://plants.ensembl.org/info/data/ftp/index.html`) and build index using `build()` function from Subread aligner (RSubread, version 1.16.2, Liao et al. 65) in the software R [83]. The index allows fast retrieval of the sets of positions in the reference genome where the short reads are more likely to align.

3. Align short reads in FASTQ files to the Arabidopsis reference genome using the `align()` function from Rsubread.

4. Summarize the read counts at the gene level using the `featureCounts()` function from the Subread aligner and store the read counts as data matrix. The annotation file

   `Arabidopsis_thaliana.TAIR10.22.gtf`

   is downloaded from Ensembl plants FTP server. The multi-mapping or multi-overlapping reads were not counted.

Subread aligner is a recently developed sequence mapping tool that adopts a seed-and-vote paradigm to map the RNA-Seq short reads to the genome. It breaks each short read into a series of overlapping segments called subreads and uses the subreads

to vote on the optimal genome location of the original read. The subreads are shorter and can be mapped to the genome much faster. Compared to other aligners such as Bowite 2 [57] or BWA [62], Subread aligner is both faster and more accurate [65]. We compared results from the above pipeline to results from a pipeline described in [4] over several RNA-Seq experiment data, and Rsubread was more than three times faster and successfully mapped more reads to the reference genome. For researchers familiar with R, it also has the advantage that it is completely implemented in R.

We divided the experiments into three groups as summarized in Table 1. As an additional data quality control measure, we keep an experiment only when it has mapping quality (number of successfully mapped reads divided by total number of reads) $\geq 50\%$ for all samples. Then within each group, we computed an initial set of normalization factors from all samples combined using the method described in Section 2.2.2. An experiment is retained only when the normalization factors of all samples in the experiment are between 0.50 and 1.50. If the initial estimated normalization factor is too different from 1 for a sample, it often indicates that the read counts distribution in the corresponding sample is markedly different from the distributions of the rest of the samples. Such samples demand additional attention before being incorporated in the studies that we intend to do.

2.2.2. Count normalization

As explained in the introduction, count normalization is needed when analyzing RNA-Seq data to 1) adjust for differences in sequencing depths or library sizes; 2) to adjust for the apparent changes in relative read frequencies of non-DE genes that occur as a consequence of changes in relative read frequencies of truly DE genes.

For the second type of adjustment, we follow Anders and Huber's method [3] for estimating normalization factors. Let y_{ij} denote the read count for ith gene of the jth sample (m genes and n samples in total). We first create a pseudo-reference sample where each gene's expression value is the geometric mean expression over all

real samples for that gene,

$$y_{i,0} = (\prod_{j=1}^{n} y_{i,j})^{1/n}, i = 1, \ldots, m. \tag{43}$$

Next we calculate the median fold-change in relative frequency between each sample j and the pseudo-reference sample,

$$R'_j = \text{median}\left(\frac{y_{1,j}/N_j}{y_{1,0}/N_0}, \ldots, \frac{y_{m,j}/N_j}{y_{m,0}/N_0}\right), \tag{44}$$

where N_j is the library size for sample j (the sum of RNA-Seq counts mapped to all genes retained in each sample). Finally, the *normalization factor* R_j for sample j is calculated as

$$R_j = \frac{R'_j}{(\prod_{j=1}^{n} R'_j)^{1/n}}. \tag{45}$$

Using the estimated normalization factors, the relative frequencies will be computed as y_{ij}/N_jR_j, which we will call the *normalized relative frequency* for gene i in sample j. The assumption made here is that the median fold change between normalized relative frequencies in two samples should be 1. In other words, this normalization method assumes that the majority of genes are not DE. The NBPSeq package [23] has an inbuilt function for this procedure and it will be used for count normalization in this paper. With the estimates from equation (45), we see that the median fold change in normalized relative frequencies between each sample and the pseudo-reference sample will be set to 1:

$$\text{median}\left(\frac{y_{1,j}/N_jR_j}{y_{1,0}/N_0R_0}, \ldots, \frac{y_{m,j}/N_jR_j}{y_{m,0}/N_0R_0}\right) = 1, \tag{46}$$

where $R_0 = (\prod_{j=1}^{n} R'_j)^{-1/n}$.

We can apply equation (44) to a subset of reference genes to estimate normalization factors. In doing so, effectively, the median fold change in equation (46) among

the reference genes will be set to 1 in each sample j. Other normalization methods may make different assumptions than Anders and Huber's, but some assumptions of a similar nature seem unavoidable. For example, the TMM method of Robinson et al. [89] is based on a similar principle: assuming the majority of the genes are not DE. The TMM method can be applied to a subset of genes selected based on an initial screening of mean expression level and fold changes. In TMM method, one can also specify certain quantile (instead of the median) of the fold changes to be 1.

In this paper, we will identify stably expressed genes from multiple data sets based on numerical measure and use them as reference for estimating normalization factors (from equations (44) and (45)). However, to identify the stably expressed genes, we first need a set of initially estimated normalization factors. To tackle this circular dependence, we use a one-step iteration method to estimate the normalization factors:

1. First, we use all the genes to calculate the initial normalization factors;

2. Then, we fit a GLMM to each gene and estimate the total variance measure, incorporating the initial normalization factors as an offset term (see Section 2.2.3);

3. Next, we select the top 1000 stably expressed genes based on the total variance measure estimated from step 2 above, and use them as reference genes to recalculate the normalization factors.

In practice, this one-step method seems to be adequate and further iterations will only slightly change the set of 1000 stably expressed genes. For example, for the multi-tissue group of experiments, if we were to run one more iteration of steps 2 and 3, there would be 946 overlapping genes between the top 1000 genes from the first iteration and those from the second iteration.

2.2.3. Poisson log-linear mixed-effects regression model and the total variance measure of expression stability

We fit a Poisson log-linear mixed-effects regression model to the RNA-Seq counts mapped to each gene and measure gene expression stability using a total variance measure.

Let Y_{ijkl} be the number of RNA-Seq reads mapped to gene i in sample j from treatment group k in experiment l. We will fit regression models to each gene separately and suppress subscript i from the model equations. For each gene, we fit a Poisson log-linear mixed-effects regression model

$$Y_{jkl} \sim \text{Poisson}(\mu_{jkl}), \tag{47}$$

$$\log(\mu_{jkl}) = \log(R_{jkl}N_{jkl}) + \xi + \alpha_l + \beta_{k(l)} + \epsilon_{jkl}, \tag{48}$$

which is a specific type of generalized linear mixed model (GLMM, McCulloch and Neuhaus [74]). In equation (48), N_{jkl} and R_{jkl} are the library size and normalization factor discussed in Section 2.2.2. We will call $R_{jkl}N_{jkl}$ the *normalized library size*. The parameter ξ is a fixed-effect term for the baseline log mean of the *relative counts* (counts divided by the normalized library sizes). The values α, β, and ϵ represent the experiment effect, the treatment effect (nested within each experiment), and the sample effect respectively. We view α, β and ϵ as random effects and assume that they are independent and follow normal distributions:

$$\alpha_l \sim N(0, \sigma^2_{\text{experiment}}), \quad \beta_{k(l)} \sim N(0, \sigma^2_{\text{treatment}}), \quad \epsilon_{jkl} \sim N(0, \sigma^2_{\text{sample}}), \tag{49}$$

where $\sigma^2_{\text{experiment}}$, $\sigma^2_{\text{treatment}}$ and σ^2_{sample} are called *variance-components*—they quantify the overall variances of the corresponding random effect terms.

The sample effect ϵ represents the extra-Poisson variation in read counts among

samples in the same treatment group and σ^2_{sample} plays a similar role as the *over-dispersion* parameter in a negative binomial model (Anders and Huber 3, Di et al. 22). The experiment effect, α, accounts for all sources of variation at the experiment level, including differences in lab personnel and conditions, day light hours, age of the plants, temperature, sequencing platform, and other unidentified sources. The contributions from these different experiment-level sources are often difficult to separate statistically. We treat the experiment effect α as a random effect because we view the collected experiments as a random sample from the pool of all Arabidopsis RNA-Seq experiments. We also treat the treatment effect β as a random effect. In a DE test, β is usually considered as a fixed-effect term. Here for evaluation of expression stability, we are not interested in the specific levels of the individual β's and focus more on the overall variation of β under a range of treatment types.

We define the stability measure as the estimated *total variance*,

$$\hat{\sigma}^2 = \hat{\sigma}^2_{\text{sample}} + \hat{\sigma}^2_{\text{treatment}} + \hat{\sigma}^2_{\text{experiment}}. \tag{50}$$

The parameters $(\xi, \sigma^2_{\text{experiment}}, \sigma^2_{\text{treatment}}, \sigma^2_{\text{sample}})$ are estimated using the `glmer()` function of the R package `lme4` ([9], version 1.1.7), which uses a Gaussian-Hermite quadrature to approximate the likelihood function. We rank all the genes according to their values of $\hat{sigma}^2$ in increasing order (smallest first), and consider highly ranked (e.g., top 1000) genes to be stably expressed.

Normal models (equation (49)) are commonly assumed for the random effects in the GLMM settings. The normality assumption is likely a simplification of reality, yet it is a good starting point and should be adequate for finding genes with low total variation—the stably expressed ones.

2.2.4. Other stability measures

The assessment of gene expression stability depends on the specific stability measure used. [20] and [21] used the coefficient of variation (CV) measure, computed as *standard deviation devided by mean*, to find stably expressed genes from microarray data.

The *M-value* in geNorm [105] is a well-cited measure. For a set of m_0 genes, the M-value measure works as follows: first, the relative variation of gene i_1 to gene i_2 is calculated as the standard deviation of their log fold changes across all the n samples;

$$V_{i_1,i_2} = st.dev\left\{\log\left(\frac{y_{1,i_1}}{y_{1,i_2}}\right), \ldots, \log\left(\frac{y_{n,i_1}}{y_{n,i_2}}\right)\right\}$$

next, the M-value for gene i_1 is obtained by taking the average of $m_0 - 1$ relative variations, one for each pair of gene i_1 and the remaining $m_0 - 1$ genes;

$$M_{i_1} = \frac{\sum_{k \neq i_1} V_{i_1,k}}{m_0 - 1},$$

In this way, each of the m_0 genes is assigned an M value to represent its corresponding expression stability.

In the Results section, we compare the M-value to the total variance measure on RNA-Seq data from the multi-tissue group experiments, and compare the stably expressed genes identified from these two measures to those identified from microarray data using the CV measure.

2.3. Results

In Section 2.3.1, we summarize the stably expressed genes identified from three different experiment groups and emphasize that stability is context-dependent. In Section 2.3.2, we show that traditional house-keeping genes are not necessarily stably expressed according to our numerical measure, and that microarray data and RNA-Seq

data may often give different sets of stably expressed genes. In Section 2.3.3, we further demonstrate that when using a numerical measure to quantify gene expression stability, the outcome will depend on the specific numeric measure used. These points should be intuitive, but they are not often emphasized in practice. In Section 2.3.4, we discuss results from our variance component analysis. In Section 2.3.5, we discuss how to use the identified stably expressed genes for count normalization.

2.3.1. Stably Expressed Genes

Using the total variance, $\hat{\sigma}^2$, from the GLMM (see equation (48) in Section 2.2.3) as a stability measure, we identified stably expressed genes from the three groups of experiments described in Section 2.2.1: the group of seedling experiments, the group of leaf experiments, and the group of experiments on different tissue types (see Table 1 for a summary). As we mentioned in the Introduction, absolutely stably expressed genes may not exist. Choosing different sample sets as reference allows us to identify stably expressed genes for different biological contexts.

In Tables 1-3 in the online supplementary materials, we summarize the top 1000 most stably expressed genes in each group. In Figure 2, we provide the histograms of the mean Count Per Million (CPM) for the 1000 most stably expressed genes identified in each group. For each gene, the CPM is computed as

$$\frac{\text{count} \times 10^6}{\text{normalized library size}} \tag{51}$$

in each sample and the mean is computed over all samples.

The lists of the top 1000 genes in the three groups share 104 genes in common (see supplementary material for details). These genes are stably expressed under a wide range of experimental conditions and in different tissue types, and thus may be worth further study. This list of 104 genes has significant overlap with the top 100 stably expressed genes identified by [20] from a developmental series of microarray samples:

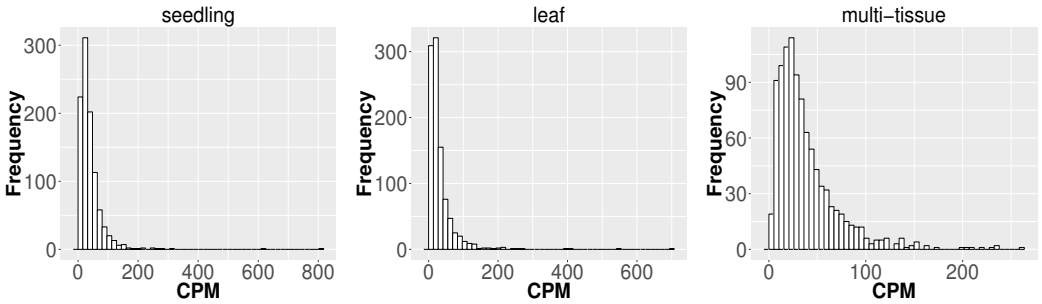

Figure 2: Histograms of the mean CPM (see equation (51)) for the top 1000 most stably expressed genes identified from the seedling (left), leaf (middle) and multi-tissue (right) groups using the total variance measure $\hat{\sigma}^2$. The mean CPM is computed over all samples within each respective group. Note that the x and y axis scales differ between the three plots.

9 out of these 104 genes (see Table 4 in the supplementary material for details),

AT1G13320, AT1G54080, AT2G20790, AT2G32170, AT3G10330,
AT4G24550, AT5G26760, AT5G46210, AT5G46630

appeared in the list of the top 100 stably expressed genes out of 14000 genes they examined (the probability is 4.8×10^{-9} for a list of 104 genes random selected from a set of 14000 genes to have an overlap of size 9 or more with a pre-selected list of 100 genes). In particular, one gene, AT1G13320, is in all but one of the ten lists of top 500 stably expressed genes identified by [20] for different experimental and experimental conditions (the only exception is the set of diurnal series), and is also identified by [44] as a stably expressed gene under all but one of the six experimental conditions they examined. This gene is ranked 159 (top 0.7%), 112 (top 0.5%), 513 (top 2.2%) according to our stability measure in the three groups we examined. This gene is a subunit of protein phosphatase type 2A complex and is involved in regulation of phosphorylation and regulation of protein phosphatase type 2A activity. It has been used as a reference gene for normalization in many papers (e.g., [13], [7]; these two papers cited [20] as reference).

2.3.2. Comparison to house-keeping genes and stably expressed genes identified from microarray data

[20] discussed the expression stability of house-keeping genes and showed that the house-keeping genes are not stably expressed according to their numerical measure. In particular, they compared the expression profiles of five traditional house-keeping genes (AT1G13440, AT3G18780, AT4G05320, AT5G12250, AT5G60390) and five genes (AT1G13320, AT5G59830, AT2G28390, AT4G33380 and AT4G34270) that they identified as stably expressed according to the CV measure from a developmental series of microarray experiments (see Figure 1 of that paper). In Figure 3, we compare the expression profiles of these 10 genes from [20] to the expression profiles of five genes (AT1G63110, AT1G79280, AT3G27530, AT4G02560, AT5G53540) that we randomly selected from the top 100 most stably expressed genes identified from the multi-tissue group RNA-Seq data according the total variance $\hat{\sigma}^2$. For each of the 15 genes, Figure 3 shows the expression levels measured in CPM over 91 samples in the eight experiments in the multi-tissue group, and Table 2 summarizes the variance components estimated from the GLMM in 2.2.3.

The five house-keeping genes show large total variation with all three variance-components relatively large as compared to the other 10 genes. This is consistent with Czechowski's observation that house-keeping genes are not necessarily stably expressed according to a numerical measure. Three of the five stably-expressed genes identified by Czechowski are among the top 1000 stably-expressed genes according to our stability measure, the total variance $\hat{\sigma}^2$. Czechowski et al. identified those five genes from microarray data and different experiments. It is not too surprising those genes might not be the most stable in RNA-Seq experiments: the two technologies differ in many aspects including coverage and sensitivity.

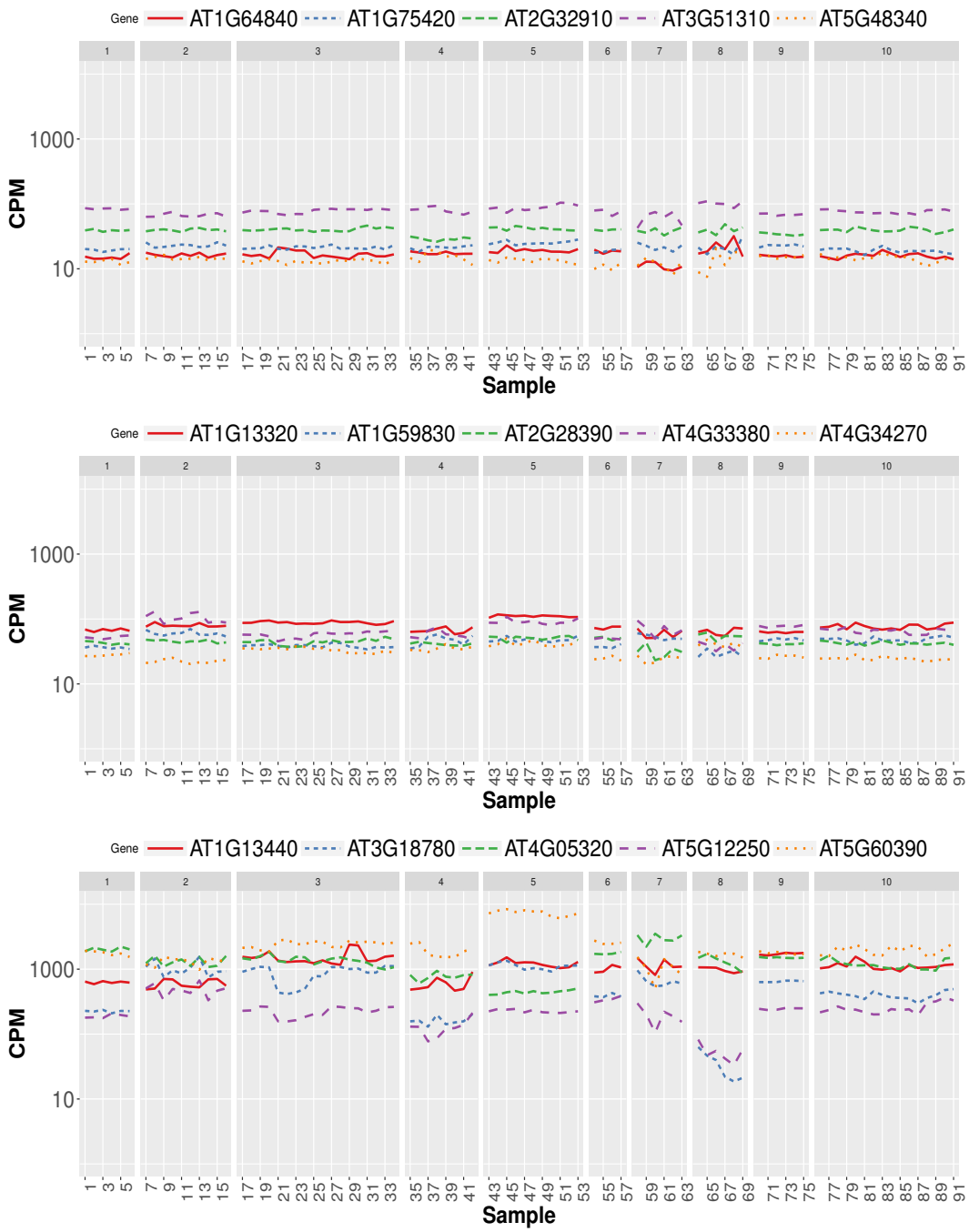

Figure 3: Expression profiles of 15 genes—as measured by CPM—across 91 samples in the multi-tissue group. The 15 genes include five stably expressed genes (randomly selected out of the top 100) identified by the total variance measure $\hat{\sigma}^2$ (GLMM), five stably expressed identified by Czechowski et al. [20] according to the CV measure from a developmental series of microarray experiments, and five traditional house-keeping genes (HKG) discussed in Czechowski et al. [20].

Table 2: Variance components estimated from the multi-tissue group for the 15 genes in Figure 3. Columns 3–5 are the estimated variance components. Column 6 specifies the ranking according to the total variance $\hat{\sigma}^2$ in the multi-tissue group.

Source	Gene	betweeen-sample	between-treatment	between-experiment	Rank
GLMM	AT1G75420	0.0012	0.0014	0.0050	5
	AT5G48340	0.0042	0.0019	0.0074	46
	AT2G32910	0.0007	0.0019	0.0113	53
	AT1G64840	0.0051	0.0008	0.0095	72
	AT3G51310	0.0028	0.0025	0.0100	73
Czechowski	AT2G28390	0.0034	0.0000	0.0111	62
	AT1G13320	0.0036	0.0003	0.0258	513
	AT4G34270	0.0063	0.0000	0.0365	1074
	AT1G59830	0.0044	0.0039	0.0370	1211
	AT4G33380	0.0103	0.0016	0.0747	3404
HKG	AT1G13440	0.0234	0.0058	0.1375	6562
	AT5G60390	0.0267	0.0068	0.2270	8867
	AT4G05320	0.0123	0.0094	0.2690	9409
	AT5G12250	0.0313	0.0128	0.3262	10589
	AT3G18780	0.0375	0.0211	1.0313	14951

2.3.3. Factors affecting stability ranking

The previous two subsections demonstrate that when using a numerical measure to quantify gene expression stability, the outcome is dependent on 1) the biological context reflected in the reference sample set used and 2) the technology used for measuring gene expression. It should also be intuitive, and we will further clarify in the second half of this subsection, that the stability ranking is also dependent on 3) the specific numerical measure used. In this section, we will first compare the lists of stably-expressed genes identified under different scenarios where one or more of the above three factors differ. We then further discuss the subtle roles played by the specific stability measure and the reference gene set by comparing the total variance $\hat{\sigma}^2$ measure from the GLMM (see equation (48)) to the M-value measure used in the geNorm method [105]. Last, we discuss the effect of an iterative elimination procedure used by geNorm.

We look at an additional five lists of stably expressed genes identified under different scenarios and examine how each of these five lists overlaps with the the top stably-expressed genes identified from the multi-tissue group of RNA-Seq experiments according to the total variance measure $\hat{\sigma}^2$ (see Section 2.2.3). The five lists are:

L_1: 100 top stably expressed genes from the multi-tissue group according to the M-value in geNorm (applied to (count + 1)) of [105] ;

L_2: 100 top stably expressed genes from the seedling group according to the total variance $\hat{\sigma}^2$ from the GLMM;

L_3: 100 top stably expressed genes from the leaf group according to the total variance $\hat{\sigma}^2$ from the GLMM;

L_4: 100 stably expressed genes identified from a developmental series of microarray experiments by [20] using the CV measure (see Section 2.2.4);

L_5: 50 stably expressed genes identified by [21] from microarray seed experiments using the CV measure.

For each list L_i above, we measure how it overlaps with the top stably expressed genes (the reference set) from the multi-tissue group using the *recall percentage*

$$\frac{\#\{L_i \cap \text{reference set}\}}{\#\{L_i\}} \times 100, \tag{52}$$

where $\#\{\}$ denotes the number of elements in the list. In Figure 4, we plot the recall percentage versus the number of top stably-expressed genes we selected as reference from the multi-tissue group.

We have the following observations:

1. The list L_1 is identified from the same set of RNA-Seq experiments as the reference sets, but using a different stability measure (M-value in geNorm). This list has significant overlap with the top stably-expressed genes identified using the total variance measure: 29 and 98 out of the 100 genes from the list L_1 are among the top 100 and 1000 most stably-expressed genes, respectively, from the multi-tissue group identified using the total variance measure.

2. The lists L_2 and L_3 are identified from different sets of RNA-Seq experiments (leaf and seedling experiments) using the same stability measure as used for the reference sets. The lists L_4 and L_5 are identified from microarray experiments (a developmental series and a seed group) and using the CV measure. The overlapping (recall) percentages are still statistically significant, but much less than in the case of L_1. This shows that differences in tissue type and in measuring technology both influence the expression stability ranking, and to comparable degrees. The lists L_3 and L_5 have the least overlapping percentages with the reference sets. These lists are identified from a leaf group and a seed group respectively. Our understanding is that the leaf group and the seed group are

more biologically homogeneous than the multi-tissue group and thus provide very different biological contexts for evaluating expression stability.

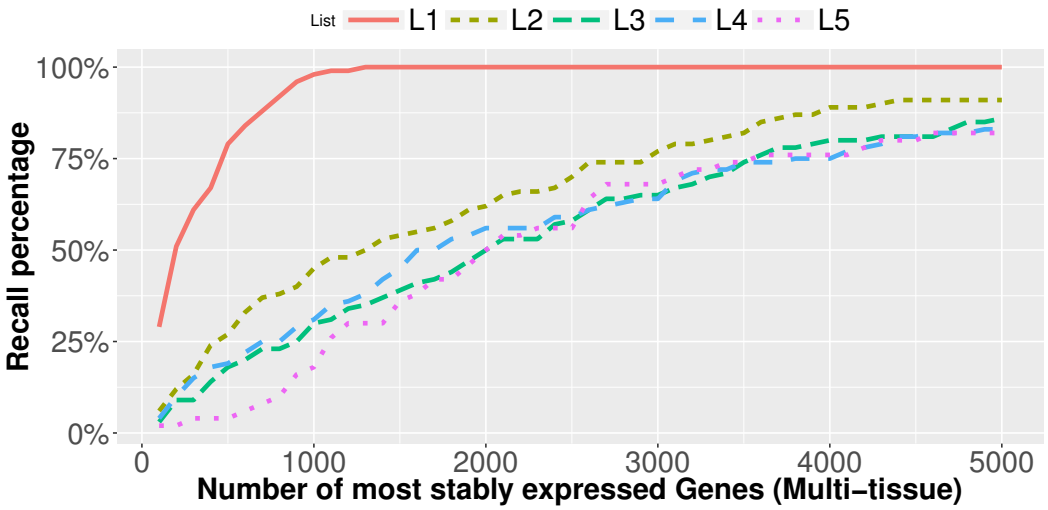

Figure 4: Comparison of top stably expressed genes identified under different scenarios. We choose the top 100 stably expressed genes as described in L_1–L_4, and the top 50 stably expressed genes in L_5 (see Section 2.3.3). and plot the recall percentages between these lists and the top most stably expressed genes identified from the multi-tissue group according to the total variance measure. The x-axis is the number of most stably expressed genes in multi-tissue group according to the total variance measure, and the y-axis shows the recall percentage (see equation (52)) for each of the five lists.

When applied to the same set of samples, the M-value and total variance measure $\hat{\sigma}^2$ give similar expression stability ranking: the rank correlation is 0.97 (see also, observation 1 above). We point out that the reason is because the M-value and normalization step needed for computing our total variance measure have similar fundamental assumptions. The basic principle behind the M-value is that the expression ratio of two stably-expressed genes should be identical in all samples. In formula, it means that the expression values of two stably-expressed genes i_1, i_2 in

any two samples j_1, j_2 should satisfy

$$\frac{y_{i_1,j_1}}{y_{i_2,j_1}} = \frac{y_{i_1,j_2}}{y_{i_2,j_2}}. \tag{53}$$

Our total variance measure $\hat{\sigma}^2$ is estimated from normalized data. The basic assumption in the normalization step is that majority of genes are not DE. In formula, it means for any stably-expressed gene i_1, its expression level as measured by the relative frequency should be stable across all samples,

$$\frac{y_{i_1,j_1}}{S_{j_1}} = \frac{y_{i_1,j_2}}{S_{j_2}}, \tag{54}$$

where S_{j_1} to S_{j_2} are the normalized library sizes (i.e., $R_j N_j$ in equation (48)). This implies for any two stably-expressed genes i_1 and i_2

$$\frac{y_{i_1,j_1}}{y_{i_1,j_2}} = \frac{y_{i_2,j_1}}{y_{i_2,j_2}} = \frac{S_{j_1}}{S_{j_2}}. \tag{55}$$

The first equation in (55) is equivalent to equation (53). (In practical application of both methods, the stability of any single gene is evaluated by comparing its expression to a set of reference genes. See the Method section 2.2.2 for more details.)

In practice, the geNorm program [105] is frequently used to rank a set of reference genes identified from other methods. An iterative elimination procedure is used along with the M-value to determine the final ranks of the expression stability: after each iteration, the gene receiving the largest M-value will be removed and a new set of M-values will be computed for the remaining genes, and the iteration will go on until there are only two genes left. We did not use such an iterative procedure in the comparisons above (i.e., we only computed one set of M-values for all genes).

This iterative elimination procedure creates an extra layer of complexity that is not well explored in literature. We use a toy example below to illustrate one subtle aspect of the iterative elimination procedure. In this example, we consider the expression

values of 7 genes in two samples shown in Table 3. When M-value is used to rank all 7 genes, the initial ranking of expression stability is given in column 4 of the table: gene 7 is the least stable and genes 4 and 5 are considered the most stable ones. Once genes 6 and 7 are eliminated, however, the recalculated M-values will rank genes 1–3 as more stable than genes 4 and 5 (see column 5 of Table 3). The root cause of this reversal of ranking is that when an iterative elimination procedure is used, effectively, the reference gene set is changing after each iteration: in the initial ranking, the expression patterns genes 4 and 5 are close to the "middle of the pack" and thus considered as the most stable, and the expression patterns of genes 1–3 and genes 6 and 7 are considered relatively more extreme; once genes 6 and 7 are removed, however, the "middle of the pack" is shifted towards the expression patterns of genes 1–3, and thus genes 1–3 become the most stably expressed. With this understanding, one could and should make a conscious decision on whether such a behavior as described above is desirable or not.

The point we want to emphasize is that gene stability is a relative concept and the stability ranking depends on which set of genes we use as reference. In an iterative elimination procedure, the reference gene set will change after each iteration. The procedure can thus give surprising results and the adoption of it in practice should not be automatic.

2.3.4. Sources of variation

For each gene, the GLMM (equation (48) of section 2.2.3) allows us to decompose total count variance into between-sample, between-treatment and between-experiment variance components. The estimated variance components tell us how much each component contributes to the overall count variation. Table 4 summarizes the percentages—averaged over all genes—of the total variance attributable to each of the three components for three groups of RNA-Seq samples (seedling, leaf and multi-tissue groups in Section 2.2.1). Figure 5 shows the histograms of the percentages. Figure 6 shows

Table 3: A toy example showing the effect of iterative elimination. Columns 2 and 3 represent expression levels for seven genes in two samples, column 4 is the stability ranking of genes by M-value without iterative elimination, and column 5 is the ranking after two geNorm iterations.

	Raw Counts		Rank	
Gene	sample 1	sample 2	rank 1	rank 2
Gene1	1	1	3	1
Gene2	1	1	3	1
Gene3	1	1	3	1
Gene4	1	2	1	4
Gene5	1	2	1	4
Gene6	1	3	6	
Gene7	1	4	7	
Library Size	7	14		

the stacked bar plot of variance components estimated from the multi-tissue group for 20 genes randomly selected from the top 1000 stably expressed genes and 20 genes randomly selected from 23611 genes. As expected, the between-experiment variance component, on average, explains the largest proportion of the total variation. The between-experiment variation is relatively smaller among the leaf samples, indicating that the leaf samples are more homogeneous. There is more variation in the relative percentages of total variance explained by the between-sample and between-treatment variance components. In principle, the between-treatment variation will be greater when there is a higher proportion of DE genes or when the samples are more homogeneous. In practice, the between-sample variance depends greatly on what samples are used as biological replicates.

2.3.5. Reference gene set for normalization

Once we have ranked the genes according to our numerical stability measure (i.e, the total variance measure, $\hat{\sigma}^2$), one application is to use an explicit set of most stably expressed genes as reference genes for count normalization. This new approach allows

Table 4: Percentages—averaged over all genes—of the total variance attributable to each of the three variance components (between-sample, between-treatment, between-experiment) for the three groups of RNA-Seq samples (the seedling, the leaf and the multi-tissue groups).

Source	Seedling	Leaf	Multi-tissue
between-sample	7.2%	16.0%	7.6%
between-treatment	20.1%	28.0%	5.1%
between-experiment	72.6%	56.0%	87.3%

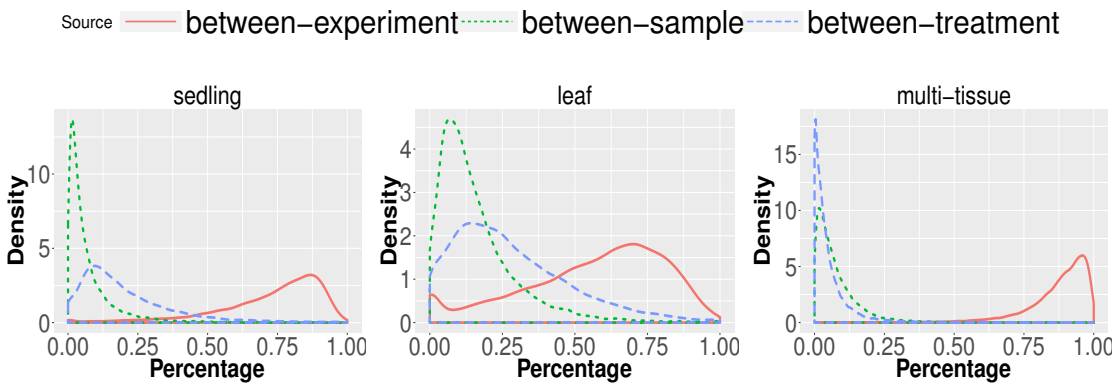

Figure 5: Distributions (over all genes) of the percentages of the total variance attributable to the between-sample, and between-treatment, or the between-experiment variance component, in the seedling, the leaf, and the multi-tissue groups.

investigators to prescribe a specific biological context for evaluating gene stability by choosing the most relevant reference samples and experiments when computing the stability measure. For example, the most stably expressed genes identified from the multi-tissue group and those identified from the seedling group will provide different biological contexts. In contrast, existing normalization approaches are often applied to the single data set under study, and thus provide a single, narrow context.

Even under a specific biological context, it is almost impossible to know whether the genes in any reference set are absolutely stably expressed, even though commonly used normalization methods often enforce some assumptions on the reference gene set: for example, when we use Anders and Huber's method to estimate the normalization factors based on a subset of reference genes, roughly speaking, the median fold change

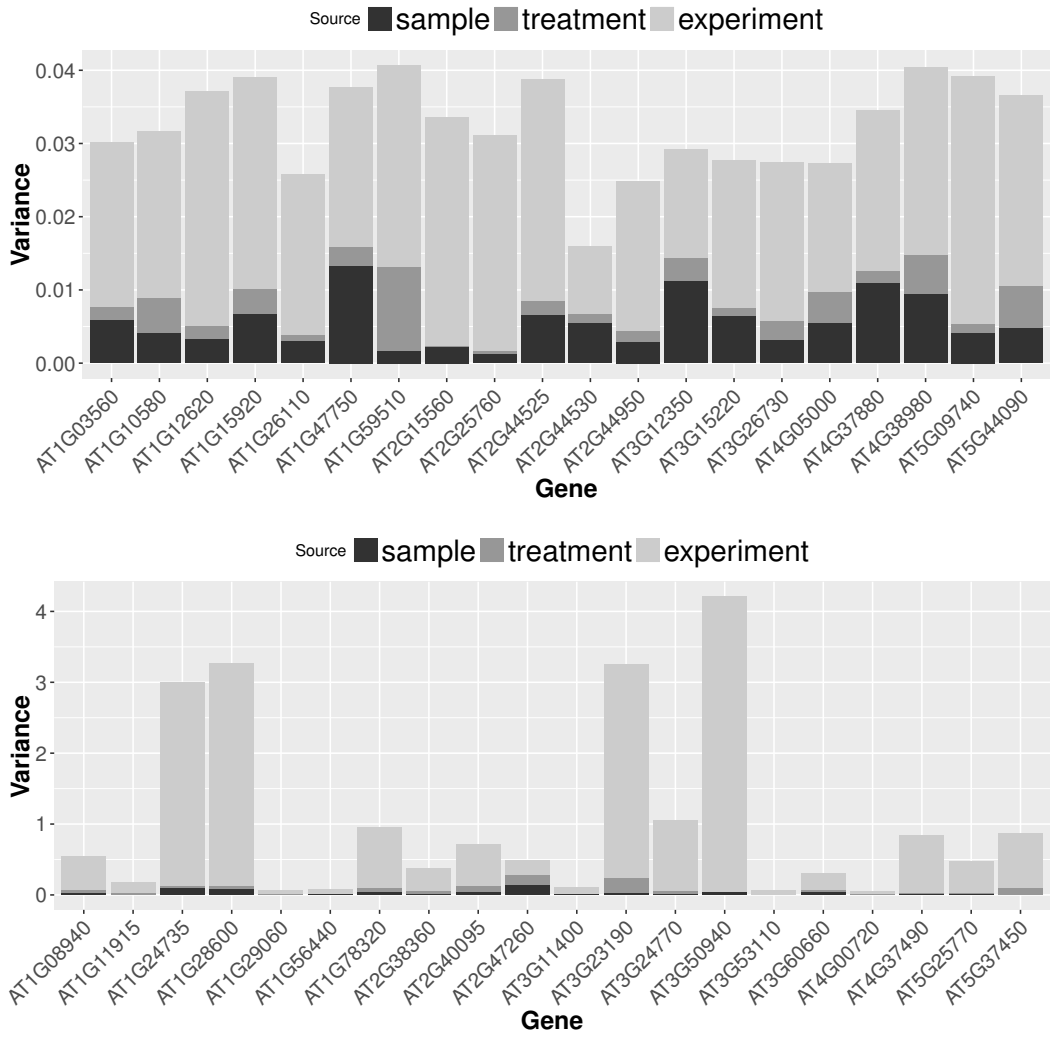

Figure 6: Stacked bar plots of the three variance components for selected genes in the multi-tissue group. Top: 20 genes randomly selected from top 1000 stably expressed genes; Bottom: 20 genes randomly selected from all the genes.

among the reference genes will be set to 1 (see Section 2.2.2 for more details). A subtle point we want to make is that since it is impossible to know how well such or similar assumptions on DE hold for a reference gene set, we can improve the interpretability of the DE test results by making the reference gene set explicit: we can slightly change our perspective and interpret all DE results as relative to the reference gene set. For example, a fold change of 2 inferred from the GLMM model can be interpreted as the

Table 5: A toy example for illustrating the importance of using a common explicit set of reference genes when comparing RNA-Seq data from multiple experiments. If a common reference gene set (e.g., genes 1–3) is used as reference for count normalization, we will notice that the DE behavior of gene 3 differs in the two experiments. If the two experiments are separately normalized using genes 1–3 as reference in experiment 1, but using genes 3–5 as reference in experiment 2, we may conclude that gene 3 is not DE in either group.

	Exp. 1		Exp. 2	
Gene	Control	Treatment	Control	Treatmetn
1	10	20	10	20
2	10	20	10	20
3	10	20	10	10
4	10	10	10	10
5	10	10	10	10

fold change of a gene is 2 times the true (but often unknowable) median fold change of the reference genes. When one estimates the normalization factors based on all genes, one is effectively specifying an implicit set of genes as a reference set. Our proposal is to make the reference set explicit and interpret DE results as relative to the reference gene set.

Interpreting the DE results as relative to an explicit reference set is especially beneficial when one wants to compare results from an experiment to ones that are publicly available. Furthermore, when the interest is in comparing different experiments, we recommend using a common reference set. For example, when two RNA-Seq data sets are separately normalized with different reference sets, a fold change of two observed in one experiment may not be directly comparable to a fold change of two observed in the other. This concern can be alleviated by using a common set of reference genes. We use a toy example to illustrate this point in Table 5 where we examine the mean counts for 5 genes in two two-group comparison experiments. If we use different reference gene sets for count normalization in the two experiments, for example, we use genes 1–3 as reference in experiment 1, but use genes 3–5 as reference in experiment

2, we may conclude that gene 3 is not DE in either experiment. If we use a common reference gene set—either genes 1–3 or genes 3–5—for normalization, however, we will be able to discover, in either case, that the DE behavior of gene 3 is different in the two experiments. Note that the DE conclusion in both experiments will depend on the reference genes used: if genes 1–3 are used as reference, gene 3 is not DE in experiment 1, but will be DE in experiment 2; if genes 3–5 are used as reference, gene 3 will be considered DE in experiment 1, but not DE in experiment 2. The point is, in either case, we will notice that the DE behavior of gene 3 is different between the two experiments. This information will be lost if one uses different reference sets to assess DE in the two experiments.

In practice, we recommend using the top 1000 most stably expressed genes for estimating normalization factors. The key is to avoid using too few (e.g., less than 10) or too many (e.g., using all genes) reference genes: intuitively, using too few, the estimates will be unstable; using too many, the results may be subject to influence from highly unstable genes. Our simple simulations suggest that using between 100 to 10000 genes seems to give stable results. In the first set of three examples, we use Anders and Huber's method (see equation (44)) to estimate normalization factors for samples in each of the seedling, leaf and multi-tissue groups of experiments (see Section 2.2.1). We use the top 10, 100, 1000, and 10000 stably expressed genes identified earlier (see Section 2.3.1 for details) as reference gene sets. Figure 7 shows the pairwise scatter plots and correlation coefficient between the normalization factors when different numbers of top stable genes are used as reference. A stronger correlation indicates the normalization factors estimated from the two settings are highly consistent. The plots and correlation coefficients suggest using between 100 and 1000 genes tend to give similar normalization factor estimates. We also used the top 10, 100, 1000, and 10000 stably expressed genes identified from the multi-tissue group as reference set for estimating normalization factors for a set of 48 root samples from a new experiment (GSE64410, [106]). The largest Pearson correlation 0.993 is between

the normalization factors estimated using the top 100 and top 1000 stably expressed genes as reference. Based on the above observations, using 1000 most stably expressed genes as reference seems to be a reasonable heuristic rule.

2.4. Conclusion and Discussion

In this paper, we advocate quantifying gene expression stability by applying a numerical stability measure to a large number of existing RNA-Seq data sets. Similar strategies have also been used by others to find stably expressed genes from microarray data. Since DE is measured by relative frequencies, we argue that DE is a relative concept and using an explicit reference gene set can improve interpretability of DE results, and furthermore, using a common reference gene set can avoid inconsistent conclusions when comparing multiple experiments (see Section 2.3.5).

It should be clear but worth emphasizing that when using a numerical measure to identify stably expressed genes, the outcome depends on multiple factors: the reference sample sets used, the technology used for measuring gene expression, and the specific numerical stability measure used. In this study, to illustrate our proposed methods, we identified three sets of stably expressed genes from three sets of Arabidopsis experiments. The major point is that stably expressed genes identified from different settings will provide different biological contexts for evaluating differential expression. In practice, researchers can choose the specific context. A practical challenge in applying such a philosophy is that no two experiments will have identical settings, and researchers have to decide what experiments can be considered comparable. This is a difficult question; however, we believe it has to be asked from now on: biologists perform comparative experiments with the intent that the conclusions from a single experiment will be generalizable beyond the context of a single lab. If we do not understand comparability between different experiments, such generalization is impossible. Defining and characterizing comparability is a challenging topic that we would like to investigate more in the future.

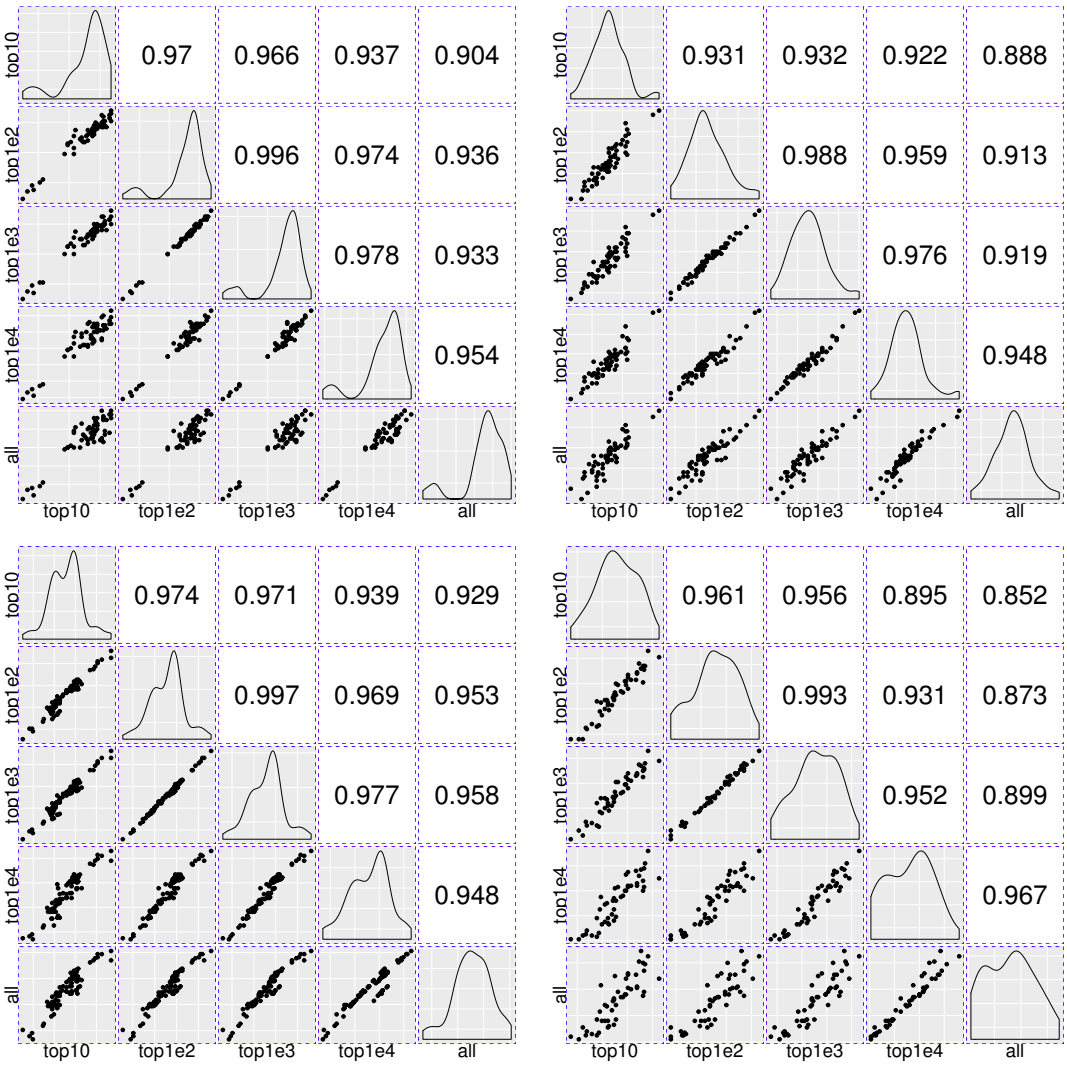

Figure 7: Matrices of scatter plots of normalization factors estimated using different reference gene sets. The upper-left, upper-right and lower-left plots show normalization factors estimated for the samples in the seedling, leaf, and multi-tissue groups correspondingly. In each case, the top 10, 100, 1000, and 10,000 stably expressed genes are used as reference to calculate the normalization factors. The lower right plot shows the normalization factors estimated for a new root experiment (GSE64410, with sample size 48) using the top 10–10,000 stably expressed genes identified from the multi-tissue group as reference. The normalization factors are estimated using the method described in Section 2.2.2.

To identify a set of stably expressed genes, our method still needs to estimate an initial set of normalization factors, which requires that we must make assumptions

about relative fold changes between samples. This kind of circular dependence seems unavoidable [105]. In this paper, we used a one-step iteration strategy to reduce the dependence on the initially estimated normalization factors. In future work, we intend to look at the genes through evolutionary genetics methods (e.g., 1001 genomes, [109]). For example, evolutionary genetics methods can help us test whether a gene is under negative, neutral, or positive selection and help us identify genes that are well conserved through the evolutionary history. We need to be mindful that a well conserved gene is not necessarily stably expressed, just like the house-keeping genes. However, it would be interesting to ask whether there is correlation between measures of expression stability and measures of conservativeness, and so on.

In the GLMM model we fit, the random effect terms such as the sample and treatment effects were modeled as normal random variables (Section 2.2.3). For the purpose of identifying stably expressed genes, this should be adequate, since we are mainly interested in the variances of these random effects (i.e., the variance components). In the future, it may also be of interest to model these random effects more accurately, for example, in order to build a prior distribution of the random effect terms for analyzing a new data set. A more careful examination of the individual data sets suggests that the between-sample variance varies greatly between experiments. Our observation suggests that different labs often have different understanding of what is deemed as "biological replicates".

The R codes for reproducing results in this paper are available at Github: `https://github.com/zhuob/StablyExpressedGenes`

Acknowledgment: Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM104977 (to YD, SCE, and JHC). We thank Duo Jiang and Wanli Zhang for helpful discussions. This article is part of doctor dissertation written by BZ under the supervision of YD.

Supplemental Material: The online version of this article offers supplementary material, available to authorized users.

3. Test statistics correlation may not converge to population correlation

Abstract

content...

3.1. Introduction

Inter-gene correlations are commonly observed in sequencing data generated from gene expression experiments [28, 81, 98, 49, 34]. The key task of expression analysis is to detect differentially expressed (DE) genes whose expression levels are associated with experimental or treatment variables under study. In such a task, a summary statistic is calculated for each gene to quantify the magnitude of DE. The test statistics are often of familiar form. For example, they may come from two-sample comparison or experimental design based regression models. However, since the expression levels are correlated, the test statistics calculated from the expression levels are also correlated [8, 27, 112]. This paper concerns the relationship between test statistics correlations and the corresponding expression level correlations.

The stochastic dependence of test statistics has brought methodological issues to statistical analysis accessing both individual genes and gene sets. The interest in examining individual genes is to find DE genes among tens of thousands of candidates. Multiple hypothesis testing procedures, such as *false discovery rate* (FDR) [10] and *q-value* [98], are needed to control type I error rate. In many cases, such techniques work only when test statistics are independent [10] or have positive regression dependency [11]. The goal of evaluating gene sets is to find molecular pathways or gene networks that are related to the experimental condition or factors of interest. Testing a gene set is usually done by pooling the test statistics of its member genes, and may or may not involve genes not in the test set [37]. In all situations, the correlation between test

statistics is a nuisance aspect, which, if not addressed appropriately, will undermine the applicability of the corresponding approaches [34, 112]. For example, Efron [27] showed in a simulation study that for a nominal FDR of 0.1, the actual FDR can easily vary by a factor of 10 when correlation between test statistics exists.

A number of attempts have been made to deal with issues of inter-gene correlation when testing either individual genes or gene sets. One approach is to derive certain summary statistic from correlation among test statistics and then use it in the hypothesis testing procedure. For testing individual genes, Efron [27] estimates some dispersion variate to summarize correlation among test statistics, and then calculates the *false discovery proportion* (FDP) conditioning on this dispersion variate. For testing gene sets, Wu and Smyth [112] estimate a *variance inflation factor* (VIF) associated with inter-gene correlation and incorporate it into their parametric/rank-based gene set test procedures. The same VIF is also used by Yaari et al. [114] to account for correlation in their distribution-based gene set test. Another approach is to permute the labels of biological samples, aiming to generate the null distribution of test statistic for each gene. This type of permutation preserves underlying correlation structure between genes, and thus protect the test against such correlations. The *gene set enrichment analysis* (GSEA) procedure [100] falls into this category. However, sample permutation method has an extra assumption, which states that the test statistics always follow the distribution they have under complete null that no gene is DE [28]. In other words, this assumption expects that the distribution of test statistics under the null is not affected by the presence of non-null cases. For this reason, we will not discuss sample permutation based methods in this paper.

Summarizing test statistics correlation requires that the correlations between test statistics are known or at least can be estimated from the data. Without replicating the experiment, however, there's no way to obtain the correlation between any pair of test statistics because only a single statistic is available for each gene. In the case of one-sided test (e.g., two sample t-test), one possible choice is to use

sample correlations (after gene treatment effects nullified) to represent correlations among test statistics [8, 27, 112, 114]. Efron [27] estimates the distribution of z-value (transformed from corresponding two sample t-test statistic) correlation by sample correlation. Barry et al. [8] show by Monte Carlo simulation of gene expression data that a nearly linear relationship holds between test statistic correlation and sample correlation for several types of test statistics they examine. This Monte Carlo simulation results are cited by Wu and Smyth [112] as a justification for estimating their VIF—a summary of correlation between test statistics—from sample correlation. In all of the works, it is shown by simulation only the equivalence (in terms of either distribution or numerical summarization) of sample correlation and test statistics correlation. To the best of our knowledge, such equivalence has not yet been justified or disproved theoretically.

We investigate the effect of testing procedures on inter-gene correlation. First, we present a formula for calculating correlation between test statistics when they take specific form and meet some assumption of independence. Then we apply this formula to a special case where two-group comparison experiment is considered. We show that 1) the test statistics correlation ρ_T is equal to the population correlation ρ when the test statistics are a linear combination of the expression levels, and that 2) ρ_T is no more larger than ρ in absolute value when the test statistics are derived from two sample t test. We conduct simulations to illustrate our findings.

A relevant research was done by Qiu et al. [81], in which they studied the effect of different normalization procedures on the inter-gene correlation structure for microarray data. They randomly assigned 330 arrays into 15 pairs, each containing 22 arrays within each array 12558 genes. Then 15 t-statistics were calculated for each gene to mimic 15 two-sample comparisons under null hypothesis of no DE. They compared the histogram of t-statistics correlation for different normalization algorithms, and concluded that the normalization procedures are unable to completely remove the correlation between the test statistics.

3.2. General setup

Correlation is a statistical quantity used to assess a possible linear relationship between two random variables or two sets of data sets. The degree of correlation is measured by *correlation coefficient*, a scaler taking values on the interval $[-1, 1]$. Correlation coefficient of $+1$ (-1) indicates perfect positive (negative dependence), while correlation coefficient of 0 implies no linear relationship between two random variables. Larger correlation coefficient (in absolute value) corresponds to stronger linear correlation.

3.2.1. Overview of Pearson's correlation coefficient

There are a number of ways to look at the correlation coefficient, many of which are special cases of *Pearson's correlation coefficient* [59]. For example, the *Kendall tau rank correlation coefficient* is computed as Pearson's correlation coefficient between the ranked variables. Throughout this paper, we will discuss Pearson's correlation under bivariate settings.

We will restrict our interest, following the notation of Lee Rodgers and Nicewander [59], to two types of Pearson's correlation coefficient. The first type of correlation, which we refer to as *population correlation*, is the standardized covariance

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{E[(X - \mu_X)(Y_j - \mu_Y)]}{\sigma_X \sigma_Y}. \tag{56}$$

In equation (56), μ_X and μ_Y are the expected values of random variables X and Y, and $\sigma_X < \infty$ and $\sigma_Y < \infty$ are the population standard errors. The second type of correlation, which we refer to as *sample correlation*, is a function of raw scores and means

$$r = \frac{\sum_j (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_j (x_j - \bar{x})^2 \sum_i (y_j - \bar{y})^2}}, \tag{57}$$

where $(\bar{x}, \bar{y})$ is the vector of arithmetic mean of the observations. Fisher [32] proved

that sample correlation r is a consistent estimator for population correlation ρ.

Let (X_j, Y_j) be a bivariate random variable representing two features (genes) of sample $j = 1, \ldots, m$, and (x_j, y_j) the corresponding realization. We assume that the population mean of (X_j, Y_j) may differ across samples, but that the population covariance structure remains the same, that is,

$$E \begin{pmatrix} X_j \\ Y_j \end{pmatrix} = \begin{pmatrix} \mu_{X,j} \\ \mu_{Y,j} \end{pmatrix} \stackrel{\text{def}}{=} \boldsymbol{\mu}_j, \quad \text{for } j = 1, \ldots, m \tag{58}$$

and

$$\text{Cov} \begin{pmatrix} X_j \\ Y_j \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \stackrel{\text{def}}{=} \boldsymbol{\Sigma} \tag{59}$$

where ρ is the population correlation defined by equation (56). In addition, we assume independence across samples (note that independence implies 0 correlation, but not vise versa),

$$\text{Cov}(X_{j_1}, X_{j_2}) = \text{Cov}(Y_{j_1}, Y_{j_2}) = 0 \quad \text{for } j_1 \neq j_2 \tag{60}$$

In the context of gene expression study, the goal is to detect DE—whether the expression level of a gene is significantly correlated with treatment or experimental variables. Let $\boldsymbol{a} := (a_1, \ldots, a_m)^T$ be a vector for a contrast of interest, then DE detection for gene X can be statistically formulated as

$$H_0 : \boldsymbol{a}^T\boldsymbol{\mu}_X = 0 \ \textit{Versus}\ H_1 : \boldsymbol{a}^T\boldsymbol{\mu}_X \neq 0, \tag{61}$$

where $\boldsymbol{X} = (X_1, \ldots, X_m)^T$ and $\boldsymbol{\mu}_X = (\mu_{X,1}, \ldots, \mu_{X,m})^T$. DE detection for gene Y can be obtained by applying the same contrast to $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ (simply replacing the subscript X by Y in equation (61)). This hypothesis testing procedure usually results in a "t-test similar" test statistic, in which the numerator is a linear combination of $\boldsymbol{X}$ and the denominator is its standard error. Without a loss of generality, we express

the test statistics as follows

$$T_X = \frac{\boldsymbol{a}^T \boldsymbol{X}}{S_X}, \quad T_Y = \frac{\boldsymbol{a}^T \boldsymbol{Y}}{S_Y}, \tag{62}$$

where S_X and S_Y are the standard error for $\boldsymbol{a}^T \boldsymbol{X}$ and $\boldsymbol{a}^T \boldsymbol{Y}$ respectively. Our main goal is to explore the relationship between population correlation (equation (56)) for the test statistics

$$\rho_T = \lim_{m \to \infty} \rho_T(m) = \lim_{m \to \infty} \mathrm{Corr}(T_X, T_Y), \tag{63}$$

and that for their corresponding expression levels

$$\rho = \mathrm{Corr}(X, Y). \tag{64}$$

We will examine two typical test statistics having the form of equation (62).

3.3. Results

In this section we present the exact formula of test statistics correlation $\rho_T(m)$ by making some assumptions about T_X and T_Y, and show that the test statistics correlation ρ_T does not always equal the population correlation ρ. For the case of two-group comparison, we prove that 1) if T_X (or T_Y) is a linear transformation of $\boldsymbol{X}$ (or $\boldsymbol{Y}$), then $\rho_T = \rho$, and that 2) if T_X (or T_Y) is the two sample t-test statistic for $\boldsymbol{X}$ (or $\boldsymbol{Y}$), then $|\rho_T| \leq |\rho|$. For 2), we show that the relationship between ρ_T and ρ depends on whether the hypotheses tests (equation 61) are true null or not. We perform simulations for the case of test statistics derived from two-sample t-test to illustrate our findings.

3.3.1. Theory

Theorem 1 *Let $(X_j, Y_j), j = 1, \ldots, m$ be independent random vectors with mean and covariance structures specified in equation (58). If $(\boldsymbol{a}^T \boldsymbol{X}, \boldsymbol{a}^T \boldsymbol{Y})$ is independent*

of (S_X, S_Y)*, then the correlation of* T_X *and* T_Y *in equation (62) can be expressed as*

$$\rho_T(m) = \frac{\rho E(S_X^{-1}S_Y^{-1}) + \frac{\boldsymbol{a}^T\boldsymbol{\mu}_X \cdot \boldsymbol{a}^T\boldsymbol{\mu}_Y}{\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}} Cov(S_X^{-1}, S_Y^{-1})}{\sqrt{\left[E(S_X^{-2}) + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_X)^2}{\sigma_X^2\boldsymbol{a}^T\boldsymbol{a}} Var(S_X^{-1})\right]\left[E(S_Y^{-2}) + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_Y)^2}{\sigma_Y^2\boldsymbol{a}^T\boldsymbol{a}} Var(S_Y^{-1})\right]}} \tag{65}$$

Proof: Since samples are independent, we have

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y}) &= \boldsymbol{a}^T\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{a} = \rho\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}, \\
\mathrm{Var}(\boldsymbol{a}^T\boldsymbol{X}) &= \sigma_X^2\boldsymbol{a}^T\boldsymbol{a}, \\
E(\boldsymbol{a}^T\boldsymbol{X})^2 &= (\boldsymbol{a}^T\boldsymbol{\mu}_X)^2 + \sigma_X^2\boldsymbol{a}^T\boldsymbol{a}, \\
E[(\boldsymbol{a}^T\boldsymbol{X})(\boldsymbol{a}^T\boldsymbol{Y})] &= E(\boldsymbol{a}^T\boldsymbol{X})E(\boldsymbol{a}^T\boldsymbol{Y}) + \mathrm{Cov}(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y}) \\
&= (\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y) + \rho\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}
\end{aligned} \tag{66}$$

Note that since S_X is independent of S_X, we have

$$\begin{aligned}
\mathrm{Var}(T_X) &= E\left[\left(\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X}\right)^2\right] - \left[E\left(\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X}\right)\right]^2 \\
&= E[\boldsymbol{a}^T\boldsymbol{X}]^2E[S_X^{-2}] - \left[E(\boldsymbol{a}^T\boldsymbol{X})\right]^2\left[E(S_X^{-1})\right]^2 \\
&= \sigma_X^2\boldsymbol{a}^T\boldsymbol{a}E(S_X^{-2}) + (\boldsymbol{a}^T\boldsymbol{\mu}_X)^2\mathrm{Var}(S_X^{-1})
\end{aligned} \tag{67}$$

Similarly,

$$\mathrm{Var}(T_Y) = \sigma_Y^2\boldsymbol{a}^T\boldsymbol{a}E(S_Y^{-2}) + (\boldsymbol{a}^T\boldsymbol{\mu}_Y)^2\mathrm{Var}(S_Y^{-1}) \tag{68}$$

and

$$\begin{aligned}
&\mathrm{Cov}(T_X, T_Y) = E\left[\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X^{-1}} \cdot \frac{\boldsymbol{a}^T\boldsymbol{Y}}{S_Y^{-1}}\right] - E\left[\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X^{-1}}\right]E\left[\frac{\boldsymbol{a}^T\boldsymbol{Y}}{S_Y^{-1}}\right] \\
&= E[(\boldsymbol{a}^T\boldsymbol{X})(\boldsymbol{a}^T\boldsymbol{Y})] \cdot E[S_X^{-1}S_Y^{-1}] - (\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y)E[S_X^{-1}]E[S_Y^{-1}] \\
&= [(\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y) + \rho\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}]E[S_X^{-1}S_Y^{-1}] - (\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y)E[S_X^{-1}]E[S_Y^{-1}]
\end{aligned} \tag{69}$$

The result follows by plugging equations (66)-(69) into equation (56).

corollary 1 *For any non zero* $\boldsymbol{a}$, $\rho_T(m) = \rho$ *if* S_X *and* S_Y *are constant with respect to* $\boldsymbol{X}, \boldsymbol{Y}$.

Proof: When S_X and S_Y are constants, $\mathrm{Cov}(S_X^{-1}, S_Y^{-1})$, $\mathrm{Var}(S_X^{-1})$ and $\mathrm{Var}(S_Y^{-1})$ are all 0, and equation (65) reduces to

$$\rho_T(m) = \frac{\rho E(S_X^{-1} S_Y^{-1})}{\sqrt{E(S_X^{-2})E(S_Y^{-2})}} = \rho. \tag{70}$$

Corollary 1 states that test statistics correlation and expression level correlation are equal under linear transformation of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

However, if we assume that (S_X, S_Y) is a non-constant function of $(\boldsymbol{X}, \boldsymbol{Y})$, then the test statistics correlation in equation (65) can be expressed as

$$\rho_T(m) = \frac{\frac{E(S_X^{-1} S_Y^{-1})}{\sqrt{\mathrm{Var}(S_X^{-1})\mathrm{Var}(S_Y^{-1})}}\rho + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y)}{\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}}\rho_s}{\sqrt{\left[\frac{E(S_X^{-2})}{\mathrm{Var}(S_X^{-1})} + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_X)^2}{\sigma_X^2\boldsymbol{a}^T\boldsymbol{a}}\right]\left[\frac{E(S_Y^{-2})}{\mathrm{Var}(S_Y^{-1})} + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_Y)^2}{\sigma_Y^2\boldsymbol{a}^T\boldsymbol{a}}\right]}} \tag{71}$$

where

$$\rho_s = \frac{\mathrm{Cov}(S_X^{-1}, S_Y^{-1})}{\sqrt{\mathrm{Var}(S_X^{-1})\mathrm{Var}(S_Y^{-1})}}. \tag{72}$$

The correlation between test statistics $\rho_T(m)$ depends on the form of test statistics, and in general, may not converge to the population correlation ρ.

3.3.2. Application of Theorem 1 under normal distribution

Many gene expression experiments are done to compare expression levels under two-treatment conditions. For the rest of this section, we discuss the relationship between ρ_T and ρ under such setting. Let $n = n_1 + n_2$ be the total number of samples, where

n_1 of them are from group 1 and n_2 from group 2, and let

$$\boldsymbol{a} = (\underbrace{\frac{1}{n_1}, \ldots, \frac{1}{n_1}}_{n_1}, \underbrace{-\frac{1}{n_2}, \ldots, -\frac{1}{n_2}}_{n_2})^T \tag{73}$$

be the contrast of interest. The mean expression levels are specified as

$$\begin{aligned} \boldsymbol{\mu}_j &= (\mu_X, \mu_Y)^T, \quad j = 1, \ldots, n_1, \\ \boldsymbol{\mu}_j &= (\mu_X, \mu_Y)^T + (\Delta_X, \Delta_Y)^T, \quad j = n_1 + 1, \ldots, n_1 + n_2. \end{aligned} \tag{74}$$

If we set $S_X = 1$, then T_X corresponds to mean difference between groups 1 and 2; instead, if $S_X = \sigma_X \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ where σ_X is known, then T_X corresponds to the statistic for two sample z-test. Therefore, according to Corollary 1, $\rho_T = \rho$ if we use mean difference or z-value as test statistics.

The two sample t-statistic is also a commonly used statistic in differential expression analysis. In the case of two sample t-test with equal variance, with the contrast $\boldsymbol{a}$ defined in equation (73), the test statistic for X is

$$T_X = \frac{\bar{X}_1 - \bar{X}_2}{S_{p,X} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{75}$$

where $S_{p,X}$ is the pooled variance

$$S^2_{p,X} = \frac{(n_1 - 1) S^2_{X,1} + (n_2 - 1) S^2_{X,2}}{n_1 + n_2 - 2}. \tag{76}$$

Similarly, we obtain T_Y by replacing the subscript "X" in equations (75) and (76). Under normal distribution assumption, we have the following theorem for two sample t-test with equal variance:

Theorem 2 *Let $(X_i, Y_i), i = 1, \ldots, n$ follow a bivariate normal distribution with mean specified by equations (74) and covariance $\boldsymbol{\Sigma}$ (see equation (58)). If T_X and T_Y*

are statistics for equal-variance two-sample t-test, then

$$Corr(T_X, T_Y) = \frac{\frac{\Delta_X \Delta_Y}{\sigma_X \sigma_Y} C\rho_s + \rho B + \rho_s \rho (A - B)}{\sqrt{\left[\frac{\Delta_X^2}{\sigma_X^2} C + A\right]\left[\frac{\Delta_Y^2}{\sigma_X^2} C + A\right]}} \tag{77}$$

where

$$\begin{aligned} A &= \frac{n_1 + n_2 - 2}{n_1 + n_2 - 4}, \quad B = \frac{(\frac{n_1+n_2-2}{2})\Gamma^2(\frac{n_1+n_2-4}{2} + \frac{1}{2})}{\Gamma^2(\frac{n_1+n_2-2}{2})}, \\ \rho_s &= Corr(S_X^{-1}, S_Y^{-1}), \quad C = \frac{(n_1 + n_2)(A - B)}{(2 + n_1 n_2^{-1} + n_1 n_2^{-1})}. \end{aligned} \tag{78}$$

The proof of Theorem 2 is presented in Section 3.4. Next we present the limit of $\mathrm{Corr}(T_X, T_Y)$.

Theorem 3 *If there exists positive constants M_1 and M_2, such that $M_1 \leq n_1 n_2^{-1} \leq M_2$, then*

$$\rho_T = \lim_{n_1+n_2\to\infty} Corr(T_X, T_Y) = \frac{\rho(1 + \beta \frac{\Delta_X \Delta_Y}{\sigma_X \sigma_Y} \rho)}{\sqrt{\left[1 + \beta \frac{\Delta_X^2}{\sigma_X^2}\right]\left[1 + \beta \frac{\Delta_Y^2}{\sigma_Y^2}\right]}} \tag{79}$$

where $\beta = \lim_{n_1+n_2\to\infty} C = (4 + 2n_1^{-1}n_2 + 2n_1 n_2^{-1})^{-1}$.

Theorem 3 says that as long as n_1 and n_2 grow proportionally to infinity, the quantity ρ_T is a function of population correlation ρ, the signal-to-noise ratios $(\delta_X, \delta_Y) = (\Delta_X/\sigma_X, \Delta_Y/\sigma_Y)$ and the sample ratio n_1/n_2. We have the following observations:

1. If both tests are true null (i.e., $\boldsymbol{\Delta} = \mathbf{0}$), then $\rho_T = \rho$.

2. If only one test is true null, then ρ_T is proportional to and smaller in absolute value than ρ (i.e., $\rho_T = \gamma_0 \rho,\ 0 < \gamma_0 < 1$).

3. If both tests are true alternative (i.e., $\boldsymbol{\Delta} \neq \mathbf{0}$), then $\rho_T \neq \rho$ in general. Specifically,

   i) when $\Delta_X \Delta_Y > 0$ (i.e., both genes are DE towards the same direction), we have $\rho_T > \rho$ for $\rho < 0$ and $0 \leq \rho_T \leq \rho$ for $\rho \geq 0$.

ii) when $\Delta_X \Delta_Y < 0$ (i.e., genes are DE towards different directions), we have $\rho < \rho_T < 0$ for $\rho < 0$ and $\rho_T < \rho$ for $\rho > 0$.

Therefore in either case, we have $|\rho_T| \leq |\rho|$.

We note that $|\rho_T| \leq |\rho|$ when test statistics are derived from two sample t test with equal variance. In other words, T_X and T_Y are always "no more correlated" than X and Y are. It's also interesting to note that when both genes are DE, $\rho_T = 0$ at $\rho = -\frac{\sigma_X \sigma_Y}{\beta \Delta_X \Delta_Y}$ and $\frac{\sigma_X \sigma_Y}{\beta \Delta_X \Delta_Y} \in (-1, 1)$. Figure 8 shows the contour plots of ρ_T versus the signal-to-noise ratios δ_X $(= \Delta_X/\sigma_X)$ and δ_Y $(= \Delta_Y/\sigma_Y)$ for different ρ's. The largest value of ρ_T (in absolute value) is always at the center, where both δ_X and δ_Y are 0 (i.e., $\Delta_X = \Delta_Y = 0$).

In addition, if $n_1/n_2 \to 0$ or ∞, then $\beta = 0$ and we have $\rho_T = \rho$. That is, when sample size of one group is not proportional to that of the other, $\mathrm{Corr}(T_X, T_Y)$ will converge to ρ regardless of whether the tests are under the null or not.

3.3.3. Simulation

We perform simulations to evaluate the correlations between test statistics and those between expression levels under two sample t-test. We simulate the expression data from normal distributions. Specifically, we let (X, Y) be the expression levels of genes X and Y, and

$$\begin{aligned} \begin{pmatrix} X_j \\ Y_j \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right], j = 1, \ldots, n_1 \\ \begin{pmatrix} X_j \\ Y_j \end{pmatrix} &\sim N\left[\begin{pmatrix} \Delta_X \\ \Delta_Y \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right], j = n_1 + 1, \ldots, n_1 + n_2 \end{aligned} \tag{80}$$

For each given ρ, we consider these $n = n_1 + n_2$ pairs of (X, Y) as observations from one *simulated* experiment. Out of this experiment, we calculate $q = (T_X, T_Y)$ where T_X and T_Y are the test statistics for gene X and gene Y respectively using two-

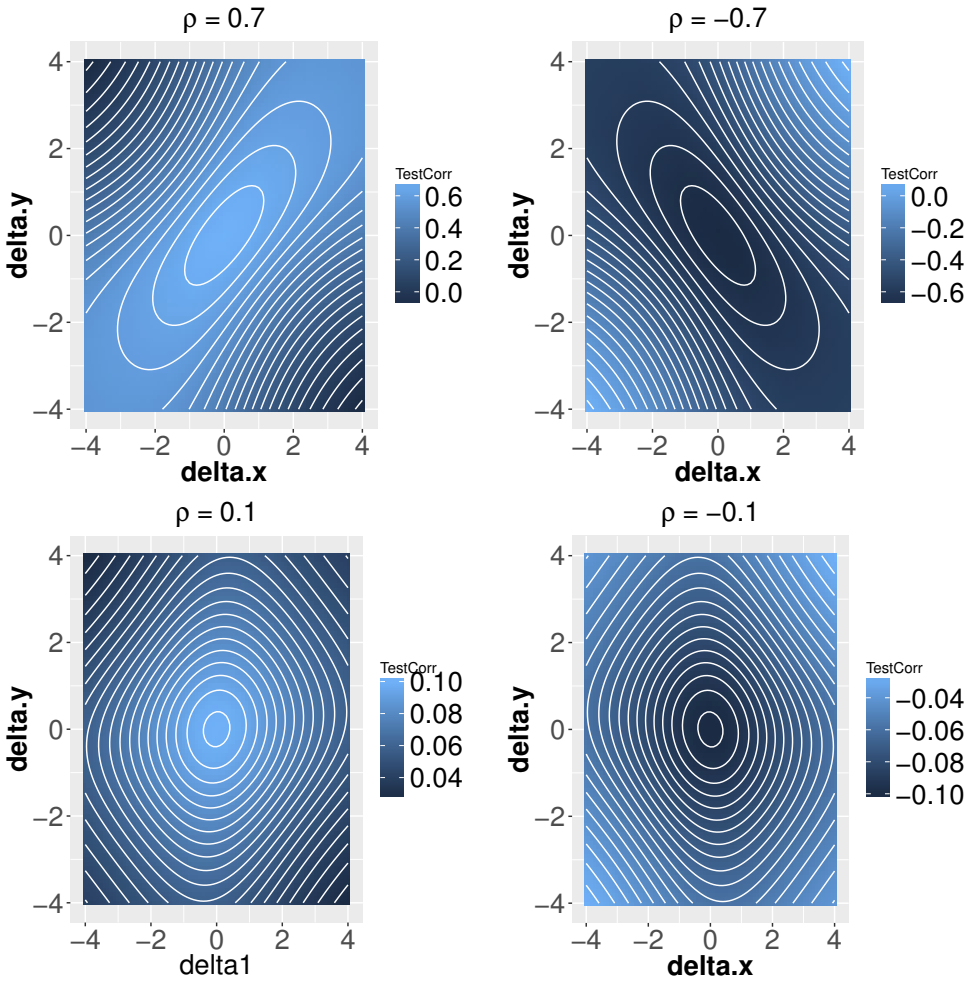

Figure 8: Contour plot of theoretical correlation between test statistics. For each fixed ρ and each pair of δ_X ($= \Delta_X/\sigma_X$) and δ_Y ($= \Delta_Y/\sigma_Y$), the theoretical correlation ρ_T is calculated according to equation (79).

sample t-test for equal variance procedure. We replicate the simulated experiment for $B = 1000$ times, resulting in a matrix $\boldsymbol{Q}_{1000\times 2}$. We take the correlation between the first and the second columns of $\boldsymbol{Q}$ as an estimate for test statistics correlation $r_{\text{statistics}}$. We increase ρ from -0.99 to 0.99 by fixed step size 0.01, and examine the relationship between $r_{\text{statistics}}$ and ρ under the following different cases:

a) $\delta_X = \delta_Y = 0$;

b) $\delta_X = 0, \delta_Y = 2$;

c) $\delta_X = 0.5, \delta_Y = 2$;

d) $\delta_X = 1, \delta_Y = 2$;

e) $\delta_X = 3, \delta_Y = 2$;

f) $\delta_X = -3, \delta_Y = 2$.

We conduct simulations for two different sample sizes: we set $n_1 = n_2 = 1000$ to assess asymptotic performance of $\rho_T(n)$ in equation (65), and set $n_1 = n_2 = 3$ to mimic small sample size scenarios which are typical in gene expression study.

In Figure 9, we plot $r_{\text{statistics}}$ against the underlying true population correlation ρ under both large and small sample size scenarios. In case a) where both tests are true null, $r_{\text{statistics}}$ is close to the true correlation ρ when sample size is large ($n_1 = n_2 = 1000$), but smaller (in absolute value) than ρ when sample size is small ($n_1 = n_2 = 3$). In cases b)—f) where there is at least one true alternative, the estimate $r_{\text{statistics}}$ can be very different from ρ. In case b) where only one gene is DE, the magnitude of $r_{\text{statistics}}$ is proportional to, and smaller in absolute value than ρ. It is more interesting to note that $r_{\text{statistics}}$ is not monotone with respect to ρ when both genes are DE. If genes are DE towards the same direction as in the case of e), $r_{\text{statistics}}$ first decreases until it reaches the minimum (a negative value), and then gradually increases to 1, as ρ grows from -1 to 1. When genes are DE towards opposite directions like the case f), however, the trend is reversed from that of e): $r_{\text{statistics}}$ increases from -1 to its maximum (a positive value), and then decreases. This set of simulation results is reflected in the test statistics correlation formula of equation (79). We also demonstrate the process of how ρ_T changes from being a linear function of ρ to a quadratic function, by fixing $\delta_Y = 2$ while increasing δ_X from 0 (case b) to 3 (case e).

We illustrate in Figure 9 the variation in $r_{\text{statistics}}$ with respect to change in sample size n. For each fixed ρ under cases a)—f), the absolute value of $r_{\text{statistics}}$ increases when we change the sample size n from 6 to 2000. The change in $r_{\text{statistics}}$ induced

by sample size could be substantial, especially when the population correlation is large (e.g., $\rho > 0.2$). This simulation shows that test statistics correlation ρ_T can be over-estimated by sample correlation, especially when sample size is small.

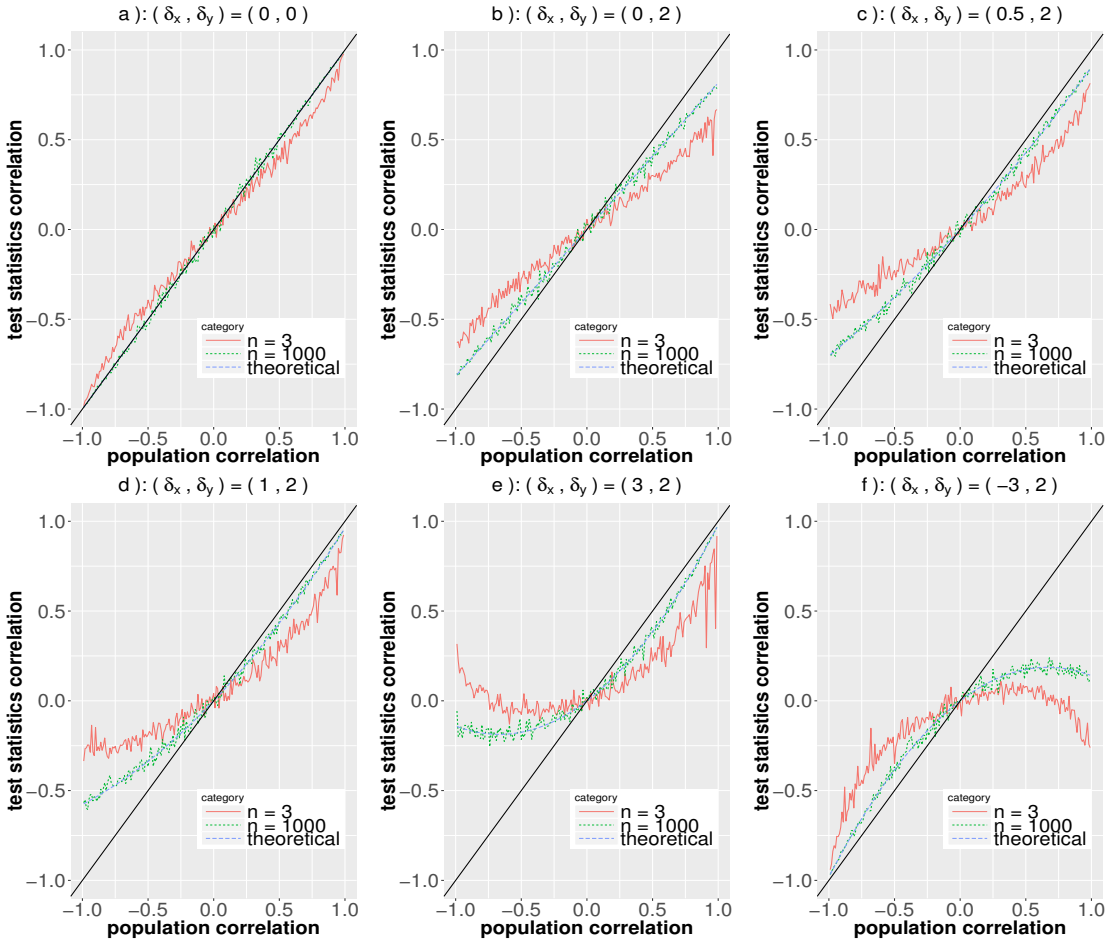

Figure 9: Plots of test statistics correlation against true population correlation. The test statistics are calculated using two sample t-test with equal variance, and the theoretical correlation is calculated by equation (79).

3.4. Method

Lemma 1 *The sample correlation coefficient r defined in equation (57) is a consistent estimator for the population correlation ρ,*

$$\sqrt{n}(r-\rho) \xrightarrow{D} N\left(0,(1-\rho^2)^2\right).$$

The proof of Lemma 1 can be found in Fisher [32].

To prove Theorem 2, it is useful to note that $\boldsymbol{U}=(\boldsymbol{a}^T\boldsymbol{X},\boldsymbol{a}^T\boldsymbol{Y})$ is independent of $\boldsymbol{S}=(S_X,S_Y)$, following from Lemmas 2 and 3.

Lemma 2 *Let $(X_j,Y_j), j=1\ldots,m$ be independent random variables satisfying equation (60), then $\boldsymbol{W}=(W_X,W_Y)=(\frac{(m-1)S_X^2}{\sigma_X^2},\frac{(n-1)S_Y^2}{\sigma_Y^2})$ follows a **bivariate chi square distribution** with density*

$$\begin{aligned} f(w_x,w_y) = {} & \frac{2^{-m}(w_xw_y)^{(n-3)/2}e^{-\frac{w_x+w_y}{2(1-\rho^2)}}}{\sqrt{\pi}\Gamma(\frac{m}{2})(1-\rho^2)^{(m-1)/2}}\times \\ & \sum_{k=0}^{\infty}[1+(-1)^k]\left(\frac{\rho\sqrt{w_xw_y}}{1-\rho^2}\right)^k\frac{\Gamma(\frac{k+1}{2})}{k!\Gamma(\frac{k+m}{2})} \end{aligned} \tag{81}$$

for $n>3$ and $-1<\rho<1$.

For proof of Lemma 2, interested readers are referred to Joarder [51]. It immediately follows from Lemma 2 that $\boldsymbol{W}_1=(\frac{(n_1-1)S_{X,1}^2}{\sigma_X^2},\frac{(n_1-1)S_{Y,1}^2}{\sigma_Y^2})$ follows bivariate chi-square distribution with degree of freedom n_1-1. Similarly, $\boldsymbol{W}_2=(\frac{(n_2-1)S_{X,2}^2}{\sigma_X^2},\frac{(n_2-1)S_{Y,2}^2}{\sigma_Y^2})$ follows a bivariate chi-square distribution with degree of freedom n_2-1. Note that $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are independent since the samples are independent.

Lemma 3 $\boldsymbol{U}=(U_X,U_Y)$ *is independent of* $\boldsymbol{S}=(S_X,S_Y)$.

Proof: By Lemma 2, the density function of $\boldsymbol{W}_1+\boldsymbol{W}_2$ only involves $\sigma_X^2,\sigma_Y^2,\rho$ and sample size n_1,n_2, therefore we can denote its density by some function $g(\sigma_X^2,\sigma_Y^2,\rho,n_1+$

n_2). Note that $\boldsymbol{S}^2 = \frac{(\sigma_X^2,\ \sigma_Y^2)}{n_1+n_2-2}(\boldsymbol{W}_1 + \boldsymbol{W}_2)^T$ is a linear transformation of $\boldsymbol{W}_1 + \boldsymbol{W}_2$, so its density also can be expressed in terms of $\sigma_1^2, \sigma_2^2, \rho, n_1, n_2$. Therefore $\boldsymbol{S} = (S_X, S_Y)$ is an ancillary statistic for $\boldsymbol{\Delta}$. On the other hand, it can be shown that $\boldsymbol{U} = (U_X, U_Y)$ is a complete sufficient statistic for $\boldsymbol{\Delta}$. It follows by Basu's theorem that $\boldsymbol{U}$ and $\boldsymbol{S}$ are independent.

Lemma 3 implies that $U_X U_Y$ is also independent of $S_X^{-1} S_Y^{-1}$, and therefore $E(\frac{U_X}{S_X} \cdot \frac{U_Y}{S_Y})$ can be expressed as $E(U_X U_Y)E(S_X^{-1} S_Y^{-1})$. We can apply Theorem 1 to calculate the correlation between T_X and T_Y under two sample t-test for equal variance.

Proof of theorem 2

First note that by Lemma 3 we have

$$\begin{aligned}\mathrm{Cov}(T_X, T_Y) &= E(T_X T_Y) - E(T_X)E(T_Y) \\ &= \frac{1}{c_0^2}\left[E(U_X U_Y)E(S_X^{-1} S_Y^{-1}) - E(\frac{U_X}{S_X})E(\frac{U_Y}{S_Y})\right]\end{aligned}$$

where $c_0 = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ and $\mathrm{Var}(T_X) = \mathrm{Var}(\frac{U_X}{c_0 S_X}) = \frac{1}{c_0^2}\mathrm{Var}(\frac{U_X}{S_X})$. Note that

$$\begin{aligned}\mathrm{Corr}(T_X, T_Y) &= \frac{\mathrm{Cov}(T_X, T_Y)}{\sqrt{\mathrm{Var}(T_X)\mathrm{Var}(T_Y)}} \\ &= \frac{E(U_X U_Y)E(S_X^{-1} S_Y^{-1}) - E(\frac{U_X}{S_X})E(\frac{U_Y}{S_Y})}{\sqrt{\mathrm{Var}(\frac{U_X}{S_X})\mathrm{Var}(\frac{U_Y}{S_Y})}}\end{aligned} \tag{82}$$

We need to calculate $E(U_X U_Y)$, $E(S_X^{-1} S_Y^{-1})$, $E(\frac{U_i}{S_i})$ and $\mathrm{Var}(\frac{U_i}{S_i})$ for $i = X, Y$.

1. Note that $U_i \sim N\left(\Delta_i, \sigma_i^2(\frac{1}{n_1} + \frac{1}{n_2})\right), i = X, Y$.

$$\begin{aligned}E(U_X U_X) &= \mathrm{Cov}(U_X, U_Y) + E(U_X)E(U_Y) \\ &= \rho\sigma_X\sigma_Y\left(\frac{1}{n_1} + \frac{1}{n_2}\right) + \Delta_X\Delta_Y\end{aligned} \tag{83}$$

2. Since $\frac{(n_1-1)S_X^2}{\sigma_X^2}$ and $\frac{(n_2-1)S_Y^2}{\sigma_Y^2}$ are independent and follow $\chi^2(n_1-1)$ and $\chi^2(n_2-1)$

respectively, , we have $W_X = \frac{(n_1+n_2-2)S_X^2}{\sigma_X^2} \sim \chi^2(n_1+n_2-2)$. It can be shown that

$$E(W_X^k) = \frac{2^k\Gamma(\frac{n_1+n_2-2}{2}+k)}{\Gamma(\frac{n_1+n_2-2}{2})}$$

Therefore

$$E\left(S_X^{-1}\right) = \frac{\sqrt{B}}{\sigma_X}, \quad \text{Var}\left(S_X^{-1}\right) = \frac{A-B}{\sigma_X^2} \tag{84}$$

Note that $\rho_s = \text{Corr}(S_X^{-1}, S_Y^{-1})$, we have

$$\begin{aligned} E(S_X^{-1}S_Y^{-1}) &= E(S_X^{-1})E(S_Y^{-1}) + \rho_s\sqrt{\text{Var}(S_X^{-1})\text{Var}(S_Y^{-1})} \\ &= \frac{B}{\sigma_X\sigma_Y} + \rho_s\frac{A-B}{\sigma_X\sigma_Y} \end{aligned} \tag{85}$$

3. $U_i \sim N\left(\Delta_i, \sigma_i^2(\frac{1}{n_1}+\frac{1}{n_2})\right)$ and $\frac{(n_1+n_2-2)S_i^2}{\sigma_i^2} \sim \chi^2(n_1+n_2-2)$ and by Lemma 3 U_i and $\frac{(n_1+n_2-2)S_i^2}{\sigma_i^2}$ are independent for $i = X, Y$, we have

$$\frac{\frac{U_i-\Delta_i}{\sigma_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\frac{(n_1+n_2-2)S_i^2}{\sigma_i^2}/(n_1+n_2-2)} = \frac{U_i-\Delta_i}{S_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2) \tag{86}$$

It follows from

$$E\left(\frac{U_i-\Delta_i}{S_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) = 0, \quad \text{Var}\left(\frac{U_i-\Delta_i}{S_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) = \frac{n_1+n_2-2}{n_1+n_2-4} \tag{87}$$

that

$$\begin{aligned} E\left(\frac{U_i}{S_i}\right) &= \frac{\Delta_i}{\sigma_i}\sqrt{B} \\ \text{Var}\left(\frac{U_i}{S_i}\right) &= A\left(\frac{1}{n_1}+\frac{1}{n_2}\right) + \frac{\Delta_i^2}{\sigma_i^2}(A-B) \end{aligned} \tag{88}$$

Finally, the test statistics correlation (77) is obtained by plugging equations (83–88) into equation (82).

Lemma 4 *If there exists a positive number M, such that $n_1n_2^{-1} \leq M$ and $n_1n_2^{-1} \leq$*

M, then the following results hold:

1. $\lim_{n_1+n_2\to\infty} A = 1.$
2. $\lim_{n_1+n_2\to\infty} B = 1.$
3. $\lim_{n_1+n_2\to\infty} C = \beta.$

where A, B and C are defined in equation (78), and $\beta = (4 + n_1 n_2^{-1} + n_1^{-1} n_2)^{-1}$.

Proof: Note that

$$B = \begin{cases} \frac{(k-1)\Gamma^2(k-\frac{3}{2})}{\Gamma^2(k-1)}, & \text{if } n_1+n_2 = 2k, k \geq 2 \\ \\ \frac{(k-\frac{1}{2})\Gamma^2(k-1)}{\Gamma^2(k-\frac{1}{2})}, & \text{if } n_1+n_2 = 2k+1, k \geq 2 \end{cases} \tag{89}$$

We will use second order Stirling's formula,

$$k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k (1 + \frac{1}{12k}) \tag{90}$$

Using Stirling's formula (90) and $\Gamma(k+\frac{1}{2}) = \frac{(2k)!}{4^k k!}\sqrt{\pi}$, it can be shown that

$$B \approx \begin{cases} \frac{(k-1)(k-2)(k-2+\frac{1}{24})^2}{(k-2+\frac{1}{12})^4}, & \text{if } n_1+n_2 = 2k, k \geq 2 \\ \\ \frac{(k-\frac{1}{2})(k-1+\frac{1}{12})^4}{(k-1+\frac{1}{24})^2(k-1)^3}, & \text{if } n_1+n_2 = 2k+1, k \geq 2 \end{cases} \tag{91}$$

It can also be shown following equation (91) that

$$A - B \approx \begin{cases} \frac{\frac{1}{4}(k-1)(k-2)^3+o((k-2)^4)}{(k-2)(k-2+\frac{1}{12})^4}, & \text{if } n_1+n_2 = 2k, k \geq 2 \\ \\ \frac{\frac{1}{4}(k-1)^3(k-\frac{1}{2})(k-3)+o((k-1)^4)}{(k-\frac{3}{2})(k-1+\frac{1}{24})^2(k-1)^3}, & \text{if } n_1+n_2 = 2k+1, k \geq 2 \end{cases} \tag{92}$$

And the results immediately follow.

Lemma 5 *Let $(X_j, Y_j), j = 1, \ldots, n$ be i.i.d. random variables under the two sample t-test for equal variance setting, with mean specified in equation (74) covariance structure in equation (58). Then we have* $\lim_{n\to\infty} \rho_s = \rho^2$.

Proof: Let's first look at samples $j = 1, \ldots, n_1$. Note that

$$S^2_{X,1} = \frac{1}{n_1}\sum_{j=1}^{n_1}(X_j - \bar{X}_1)^2 \tag{93}$$

is the *maximum likelihood estimator* (MLE) for σ^2_X. By invariance property of MLE, the pooled variance estimator

$$\begin{pmatrix} S^2_X \\ S^2_Y \end{pmatrix} = a_1 \begin{pmatrix} S^2_{X,1} \\ S^2_{Y,1} \end{pmatrix} + a_2 \begin{pmatrix} S^2_{X,2} \\ S^2_{Y,2} \end{pmatrix} \tag{94}$$

where

$$n = n_1 + n_2, \quad a_1 = \frac{n_1 - 1}{n-2}, \quad a_2 = \frac{n_2 - 1}{n-2}$$

is also MLE for $(\sigma^2_X, \sigma^2_Y)^T$ respectively. It can be shown that

$$\begin{gathered} E[S^2_X] = \sigma^2_X,\ E[S^2_Y] = \sigma^2_Y, \\ \mathrm{Var}[S^2_X] \to \frac{2\sigma^4_X}{n},\ \mathrm{Var}[S^2_Y] \to \frac{2\sigma^4_Y}{n},\ \mathrm{Cov}(S^2_X, S^2_Y) \to \frac{2\rho^2\sigma^2_X\sigma^2_Y}{n} \end{gathered} \tag{95}$$

We have

$$\sqrt{n}\left[\begin{pmatrix} S^2_{X,1} \\ S^2_{Y,1} \end{pmatrix} - \begin{pmatrix} \sigma^2_X \\ \sigma^2_Y \end{pmatrix}\right] \xrightarrow{d.} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, 2\begin{pmatrix} \sigma^4_X & \rho^2\sigma^2_X\sigma^2_Y \\ \rho^2\sigma^2_X\sigma^2_Y & \sigma^4_Y \end{pmatrix}\right] \tag{96}$$

If we let $g(x) = x^{-\frac{1}{2}}$, and apply δ-method to equation (96), we obtain

$$\sqrt{n}\left[\begin{pmatrix} S_X^{-1} \\ S_Y^{-1} \end{pmatrix} - \begin{pmatrix} \sigma_X^{-1} \\ \sigma_Y^{-1} \end{pmatrix}\right] \xrightarrow{d.} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} \sigma_X^{-2} & \rho^2\sigma_X^{-1}\sigma_Y^{-1} \\ \rho^2\sigma_X^{-1}\sigma_Y^{-1} & \sigma_Y^{-2} \end{pmatrix}\right] \tag{97}$$

It follows from equation (97) that $\text{Corr}(S_X^{-1}, S_Y^{-1}) \to \rho^2$.

3.5. Conclusion

This article discusses the relationship between population correlation ρ and the corresponding test statistics correlation ρ_T. We investigate ρ_T for test statistics of the form $(\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X}, \frac{\boldsymbol{a}^T\boldsymbol{Y}}{S_Y})$ (see equation (62)), where the denominator is the standard error of the numerator. Assuming independence between $(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y})$ and (S_X, S_Y), we derive the formula for test statistics correlation ρ_T, and show that ρ_T may not equal population correlation ρ.

In two group comparison setting, we conclude that $\rho_T = \rho$ when S_X (or S_Y) is constant with respect to $\boldsymbol{X}$ (or $\boldsymbol{Y}$). That is, $\rho_T = \rho$ under linear transformation of $\boldsymbol{X}$ and $\boldsymbol{Y}$, which is the case for two sample z-test. However, when S_X (or S_Y) is a function of $\boldsymbol{X}$ (or $\boldsymbol{Y}$), as is the case of two sample t-test, this equality may not hold. For two sample t-test, we prove that $\rho_T = \rho$ only if the null in equation (61) is true for all the tests considered, and that $|\rho_T| \leq |\rho|$ otherwise. In the case where one test is true null and the other true alternative, ρ_T is directly proportional to ρ, while when both tests are true alternatives, ρ_T is a quadratic function of ρ.

We note that cares need to be taken when estimating correlations between test statistics. In gene expression analysis, the two sample t-test [8, 27, 82] or moderated t-test [112] are used to calculate test statistics for DE detection, and the sample correlation (after treatment effects nullified) are used to adjust for correlation between those test statistics. Our study shows that, however, for DE genes, ρ_T may be overestimated when two genes are positively correlated, and underestimated when

they are negatively correlated. If there are true DE genes whose expression levels are correlated in either way, the VIF may not be accurately estimated in [112], resulting in biased test for their enrichment analysis (REF our paper?). Our results also indicate that the variance of ρ_T may also be overestimated in [27], which leads to larger variation in estimating their conditional FDP.

Theorem 1 and the subsequent results hold when the following two assumptions are met: 1) the test statistic has the of the form $\boldsymbol{a}^T\boldsymbol{X}/S_X$, and 2) $\boldsymbol{a}^T\boldsymbol{X}$ and S_X are independent. In practice, both assumptions are vulnerable. The test statistic may take different forms, depending on many factors such as the nature of the data (RNA-Seq or microarray), the experimental design structure, and the statistical hypothesis to be tested. The independence assumption between $\boldsymbol{a}^T\boldsymbol{X}$ and S_X are unlikely to hold unless the statistic is derived from two sample t-test for normally distributed random variables. Therefore, the application of Theorem 1 is somewhat limited. Yet one goal of this study is to raise awareness that the equality of ρ_T and ρ should not be taken for granted. In the future, we will explore the relationship between ρ_T and ρ for more general cases and for other types of statistics.

4. Accounting for correlations in competitive gene set test for improved interpretation of genome-scale data

Abstract

Competitive gene set test is a widely used tool for interpreting high-throughput biological data, such as gene expression and proteomics data. It aims at testing categories of genes for enriched association signals in a list of genes inferred from genome-wide data. Most conventional enrichment testing methods ignore or do not properly account for the widespread correlations among genes, which, as we show, can result in inflated type I error rates and power loss. We propose a new framework, MEQLEA, for gene set test based on a mixed effects quasi-likelihood model, where the data are not required to be Gaussian. Our method effectively adjusts for completely unknown, unstructured correlations among the genes. It uses a score test approach and allows for analytical assessment of p-values. Compared to existing methods such as GSEA and CAMERA, our method enjoys robust and substantially improved control over type 1 error and maintains good power in a variety of correlation structure and association settings. We also present two real data analysis to illustrate our approach.

4.1. Introduction

Gene set test is a statistical framework of studying the association between a test set—a *prior* set consisting of biologically related genes—and a set of genes that are significantly correlated with treatment or experimental design variables. A key task of gene expression analysis involves the detection of differentially expressed genes. Differential expression (DE) analysis evaluates each individual gene separately, and therefore it fails to provide insight into the relation between treatment variables and the prior gene set under study. Gene set test helps researchers better understand the underlying biological processes in terms of ensembles of genes.

Depending on the definition of the null hypothesis, there are two types of gene set test [37]: the *self-contained* test and the *competitive* test. A self-contained test examines a set of genes by a fixed standard without reference to other genes in the genome [39, 38, 104, 111, 49]. A competitive test compares DE genes in the

test set to those not in the test set [102, 112, 114]. Many methods, regardless of the type of test, perform a three-stage analysis [54]: on the first stage, a *gene-level statistic* is calculated for each gene in the whole genome to measure the association between the expression profiles and the experimental design variables; such gene-level statistic includes, among others, *signal-to-noise ratio* [100], *ordinary t-statistic* [102] or *moderated t-statistic* [93], *log fold change* [55] and *Z-score* [27]. On the second stage, a *set-level statistic* is obtained by utilizing the gene-level statistics from the first stage and their membership with respect to the test set (i.e., whether the gene belongs to the test set). Examples of the set-level statistic are *enrichment score* [100], *maxmean statistic* [29], and statistic derived from convoluted distribution of gene-level statistics [114], to name a few. On the third stage, a p-value is assigned to the test set by comparing the set-level statistic to its reference distribution. The competitive gene set test is much more popular among genomic literatures [37, 34].

Many competitive gene set tests rely on independence of gene-level statistics which further requires independence among gene expression levels. Those tests are parametric or rank-based procedures that assume the gene-level statistics to be independent and identically distributed, or gene permutation procedures that generate the same approximate null for the set-level statistics. For example, PAGE [55] conducts one-sample z-test by comparing the mean of gene-level statistics (i.e., the mean of log fold changes) in the test set to a normal distribution under the null. The 2×2 contingency-table-based tests examine the significance of the test set by dichotomizing the outcomes of DE analysis and cross-classifying the genes according to whether they are indicated as DE and whether they are in the test set (see [47] for a review and references therein). sigPathway [102] and "geneSetTest" in the limma package [93] evaluate the set-level p-values by permuting gene labels. However, tests assuming independence of genes may result in inflated false discovery rate [29, 37, 34, 112, 114], as genes within a gene set are often co-expressed and function together.

A handful of methods have been proposed to account for inter-gene correlation

in competitive gene set test. One attempt is to evaluate the set-level statistic by permuting the biological sample labels [100, 29]. Permuting sample labels does not require an explicit understanding of the underlying correlation structure among genes and thus protects the test against such correlation. Since permuting sample labels is computationally inefficient, Zhou et al. [116] proposed an analytic approximation to permutations for set-level score statistics, which preserves the essence of permutation gene set analysis with greatly reduced computational burden. However, an unavoidable problem arising from sample permutation approach is that it implicitly alters the null hypothesis being tested and it is therefore difficult to characterize the null and the alternative hypotheses [37, 54, 112]. Another attempt is to use set-level statistic that directly includes inter-gene correlation estimated from the data. For example, CAMERA [112] calculates a *variance inflation factor* (VIF) from sample correlation (after the treatment effect removed), and then incorporates it into their set-level statistics to account for inter-gene correlations. QuSAGE [114], which is a recent extension to CAMERA, also used the same VIF in their test procedure to adjust for inter-gene correlations. The VIF is a crucial factor and valid estimation of it relies on the assumption that correlation between any two gene-level statistics are almost the same as correlation between their corresponding expression levels. Barry et al. [8] showed by simulation that this assumption holds for several gene-level statistics (e.g., t-statistic, Wald-type statistic for regressing expression on censored time-to-event data through a Cox proportional hazards model). However, this assumption is likely to be problematic when a fraction of genes are truly DE, in which case the correlation among gene-level statistics (e.g., t-statistics) can be badly estimated by sample correlation (Zhuo and Di, unpublished work).

We propose a new framework for enrichment analysis that we will call Mixed Effects Quasi-Likelihood Enrichment Analysis (MEQLEA). Our idea is motivated by the discrepancy between correlations among expression levels and those among gene-level statistics caused by the presence of DE genes. To tackle such discrepancy, we

use differences in mean as gene-level statistics for a two group comparison experiment. We model the covariance of gene-level statistics by two variance components, one attributable to correlations among samples after treatment effect removed, and the other attributable to the DE effect associated with the treatment. The benefit of quasi-likelihood is that the data are not required to be Gaussian. Our method effectively adjusts for completely unknown, unstructured correlations among the genes. MEQLEA uses a score test approach and allows for analytical assessment of p-values. Compared to existing methods including GSEA and CAMERA, MEQLEA enjoys robust and improved control over type I error and maintains good power in a variety of correlation structure and association settings.

The rest of the paper is organized as follows: in Section 4.2 we describe the methodology of MEQLEA, as well as the simulation setup for evaluating type I error rate and power, and then we summarize some existing methods; in Section 4.3 we present results from comparison of MEQLEA to other existing methods by simulation study, and illustrate the application of our method by two real data sets; in Section 4.4 we conclude and also specifies the future work.

4.2. Methods

We consider a gene expression (e.g. RNA-Seq or microarray) experiment, in which we compare the expression levels of samples from two groups: a treatment group with n_1 samples referred to as "cases" and a control group with n_2 samples referred to as "controls" ($n_1, n_2 \geq 3$). Suppose the expression levels of a set of m genes are observed for each sample. An unknown subset of these genes are DE between cases and controls, with varying sign and magnitude of DE effects. The genes are also allowed to have (negatively or positively) correlated expression levels. In enrichment analysis, we are interested in a pre-defined set of genes, for example, from a known pathway or given by a functional annotation term from a database such as KEGG [52] or GO [6]. Our goal is to test whether this known gene set is enriched with differential expression

signals. Let $\boldsymbol{G}$ be an m-dimensional vector defining the gene set of interest, where $G_i = 1$ if and only if the i^{th} gene is in the set and $G_i = 0$ otherwise. Our analysis will condition on $\boldsymbol{G}$ and test if $\boldsymbol{G}$ is associated with enhanced DE effects. In the following sections, we will first construct a hierarchical model for the gene expression data incorporating possible correlations among the m genes, from which we will derive a quasi-likelihood model for the gene-level DE statistics jointly for all the genes. Based on this model, we will then present our enrichment test, and discuss its connections with CAMERA. Finally, we will describe our simulation studies used to evaluate our method. For the rest of this section, our presentation of the method is conditional on $\boldsymbol{G}$ unless otherwise indicated.

4.2.1. MEQLEA

4.2.1.1 A hierarchical model for gene expression data

We will start by presenting the hierarchical model for the observed gene expression data, which will incorporate the following features. Firstly, for a given sample, the expression levels of different genes are allowed to be correlated. We further assume that the correlation structure is the same across samples. Secondly, different genes may have different baseline expression levels, where "baseline" refers to the average among controls. Thirdly, for any given gene, its mean expression level in the treatment group can be either higher, lower or the same compared to the control group, depending on whether the gene is up-regulated, down-regulated, or not DE. For the genes that are differentially expressed, their DE effects are modeled additively and are allowed to have heterogeneous signs and magnitudes. Finally, given a gene, and its DE effect, the expression level is allowed to vary independently across samples, which captures measurement error and sample-level variability.

To present our model formally, we first introduce some notation. Let $n = n_1 + n_2$ be the total sample size. Let $\boldsymbol{X}$ be an n-dimensional known vector of 1's and 0's denoting the case-control membership of the samples, with $X_i = 1$ for a case and $X_i = 0$ for a

control. Let $\boldsymbol{Y}$ be an m by n matrix representing the expression data, in which each column is the expression profile for a sample and Y_{ij} $(1 \leq i \leq m, 1 \leq j \leq n)$ is the expression level of sample j at gene i. Let μ_i $(1 \leq i \leq m)$ be the baseline expression level for gene i. The quantities μ_i's are treated as nuisance parameters and as we will see later do not contribute to our analysis. Let $\boldsymbol{\Delta} = (\Delta_1, \cdots, \Delta_m)^T$ be a vector for the additive DE effects for the genes. Gene i is not DE if $\Delta_i = 0$, up-regulated if $\Delta_i > 0$ and down-regulated if $\Delta_i < 0$. We model $\boldsymbol{\Delta}$ as a random effect, for which we will detail our assumptions later. Given μ_i and Δ_i, the mean expression level for the control group and the treatment group are μ_i and $\mu_i + \Delta_i$, respectively. Given these means, the noise in the observed expression data for the j^{th} sample is denoted by the mean zero error vector $\epsilon_j = (\epsilon_{1j}, \cdots, \epsilon_{mj})^T$, $1 \leq j \leq n$. We assume $\boldsymbol{\epsilon} := (\epsilon_1, \cdots, \epsilon_m)$ to be independent of $\boldsymbol{\Delta}$ and to have mean zero. Without loss of generality, we also assume $\mathrm{Var}(\epsilon_{ij}) = 1$ for all genes and samples. For a real gene expression data set typically not satisfying this assumption, we can standardize the data by each gene to ensure that its empirical variance equals one before implementing our method (see Appendix for more detail). For the covariance structure of $\boldsymbol{\epsilon}$, we assume

$$\epsilon_{j_1} \text{ and } \epsilon_{j_2} \text{ are independent}, \quad j_1 \neq j_2, \tag{98}$$

$$\mathrm{Cov}(\epsilon_j | \boldsymbol{G}) = \boldsymbol{C}, \ \ 1 \leq i \leq n, \tag{99}$$

where $\boldsymbol{C}$ is an m by m inter-gene correlation matrix shared by all samples and is generally unknown.

Putting these elements together, we obtain the following model for the expression data $\boldsymbol{Y}$ given $\boldsymbol{X}$ and $\boldsymbol{G}$

$$Y_{ij} = \mu_i + X_j \cdot \Delta_i + \epsilon_{ij}, \tag{100}$$

for $1 \leq i \leq m, 1 \leq j \leq n$. The term $\boldsymbol{G}$ enters this model via Δ_i and possibly μ_i.

4.2.1.2 Assumptions on the DE effects Δ_i

Conditional on $\boldsymbol{G}$, we assume that the Δ_i's are mutually independent and come from either of the two distributions, $\mathscr{D}_1$ and $\mathscr{D}_2$, depending on whether $G_i = 0$ or 1. We denote the expected values of $\mathscr{D}_1$ and $\mathscr{D}_2$ by β_0 and $\beta_0 + \beta_1$, respectively, and their variances by σ_1^2 and σ_2^2, respectively. It follows that

$$E(\boldsymbol{\Delta}|\boldsymbol{G}) = \beta_0 + \beta_1 \boldsymbol{G}, \;\; \text{var}(\boldsymbol{\Delta}|\boldsymbol{G}) = \sigma_1^2 \boldsymbol{I}_1 + \sigma_2^2 \boldsymbol{I}_2, \tag{101}$$

where $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ are diagonal matrices of dimension m with 0's and 1's on their diagonals. The 1's in the diagonal of $\boldsymbol{I}_1$ correspond to the genes with $G_i = 1$ and those for $\boldsymbol{I}_2$ to the genes with $G_i = 0$.

Aside from the conditions in equation (101) on the first two moments, we do not impose any specific distributional assumptions such as normality on $\boldsymbol{\Delta}$. For example, the distribution of a given Δ_i can put positive mass on zero, which allows for the highly likely event that some of the genes are not DE. To further motivate our general framework for $\boldsymbol{\Delta}$, we present a simple model included by equation (101) as a special case. Suppose the m genes are independently sampled to be either DE or not. The probability for gene i to be DE is p_t if $G_i = 1$ or p_b if $G_i = 0$. For DE genes, their DE effects are sampled independently from a common distribution with mean μ_δ and variance σ_δ^2. Under these assumptions,

$$E(\Delta_i|\boldsymbol{G}) = p_i \mu_\delta, \;\; \text{Var}(\Delta_i|\boldsymbol{G}) = p_i \sigma_\delta^2 + p_i(1 - p_i)\mu_\delta^2, \tag{102}$$

where $p_i = p_t$ if $G_i = 1$ and $p_i = p_b$ if $G_i = 0$. It can be shown that this model is a special case of equation (101).

4.2.1.3 Model for gene-level statistics

For each gene i, we consider the gene-level statistic U_i given by

$$U_i = \frac{\sum_{j:X_j=1} Y_{ij}}{n_1} - \frac{\sum_{j:X_j=0} Y_{ij}}{n_2}, \tag{103}$$

which is sample mean difference in the expression levels between cases and controls. Given our assumption that ϵ_i has variance 1, U_i provides a DE metric for gene i. We will construct a quasi-likelihood model for $\boldsymbol{U} = (U_1, \cdots, U_m)^T$ by deriving the mean and covariance structures of $\boldsymbol{U}$ from the model for $\boldsymbol{Y}$ described in Sections 4.2.1.1 and 4.2.1.2. We first observe that combining equations (103) and (100) yields

$$U_i = \Delta_i + \eta_i, \text{where } \eta_i = \frac{1}{n_1} \sum_{j:X_j=1} \epsilon_{ij} - \frac{1}{n_2} \sum_{j:X_j=0} \epsilon_{ij}. \tag{104}$$

It can be shown based on equations (98), (99) and (101) that

$$E(\boldsymbol{U}|\boldsymbol{G}) = \beta_0 + \beta_1 \boldsymbol{G}, \tag{105}$$

$$\Sigma := \text{Var}(\boldsymbol{U}|\boldsymbol{G}) = \sigma_0^2 \boldsymbol{C} + \sigma_1^2 \boldsymbol{I}_1 + \sigma_2^2 \boldsymbol{I}_2, \tag{106}$$

where $\sigma_0^2 = 1/n_1 + 1/n_2$ is a known parameter. We note that in equation (106), the covariance structure of $\boldsymbol{U}$ has three components, a component with $\boldsymbol{C}$ which accounts for the contribution from sample-level noise $\boldsymbol{\epsilon}$, and two additional components from the DE effect $\boldsymbol{\Delta}$. It is noteworthy that both the $\boldsymbol{C}$ component and the $\boldsymbol{\Delta}$ components contribute to the variance of U_i's, whereas only the $\boldsymbol{C}$ component contributes to the correlation among U_i's.

4.2.1.4 The set-level test statistic

For a competitive gene set test, it is often unclear what the hypothesized null is and what is being tested [8, 112]. In our approach, to detect patterns of the DE signals

in the gene set of interest that stand out compared with genes not in the set, we test $H_0 : \mathscr{D}_0 = \mathscr{D}_1$ against $H_1 : \mathscr{D}_0 \neq \mathscr{D}_1$. For example, for the special scenario given by equation (102), this amounts to testing $p_b = p_t$ against $p_b \neq p_t$. To construct the test statistic, we focus on the part of the alternative space where $E(\mathscr{D}_0) \neq E(\mathscr{D}_1)$, or equivalently $\beta_1 \neq 0$. We first consider the less interesting case with uncorrelated genes, in which $\boldsymbol{C}$ equals $\boldsymbol{I}$, an m-dimensional identity matrix. Under the quasi-likelihood model for $\boldsymbol{U}$ given in Section 4.2.1.3, the quasi-score statistic for β_1 has the form $S \propto \boldsymbol{G}^T(\boldsymbol{U} - \hat{\beta}_0 \mathbf{1}_m)$, where $\hat{\beta}_0 = \overline{U}$ is an estimate for β_0 and $\mathbf{1}_m$ is a m-dimensional vector of 1's. To perform a quasi-score test, one would divide S^2 by its estimated variance under H_0 and the assumption that $\boldsymbol{C} = \boldsymbol{I}$. The resulting test statistic is

$$T_{\mathrm{u}} = \frac{S^2}{\widehat{\mathrm{Var}}_{0,\boldsymbol{C}=\boldsymbol{I}}(S|\boldsymbol{G})} = \frac{[\boldsymbol{G}^T(\boldsymbol{U} - \hat{\beta}_0 \mathbf{1}_m)]^2}{\boldsymbol{G}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{G}}, \tag{107}$$

where $\boldsymbol{H} = \dfrac{1}{m}\mathbf{1}_m\mathbf{1}_m^T$. The subscript "u" stands for "uncorrelated genes". For the case of interest when inter-gene correlation is present, $\boldsymbol{C}$ is a non-trivial correlation matrix. We will again form our test statistic based on S. However, for the denominator of the statistic, the null variance of S will be evaluated under the quasi-likelihood model with non-trivial $\boldsymbol{C}$. By equation (106), the variance of S is given by $\mathrm{Var}(S|\boldsymbol{G}) = \boldsymbol{G}^T(\boldsymbol{I} - \boldsymbol{H})\Sigma(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{G}$. Note that $H_0 : \mathscr{D}_0 = \mathscr{D}_1$ implies $\sigma_1^2 = \sigma_2^2$. Thus, under H_0, $\Sigma := \mathrm{Var}_0(\boldsymbol{U}|\boldsymbol{G}) = \sigma_0^2\boldsymbol{C} + \sigma_1^2\boldsymbol{I}$, where $\sigma_0 = 1/n_1 + 1/n_2$ is known and σ_1^2 is an unknown parameter. To estimate σ_1^2 under H_0, we observe that $\mathrm{Var}_0(U_i) = \sigma_0^2 + \sigma_1^2$ and use $\hat{\sigma}_1^2 = \sum_{i=1}^m (U_i - \overline{U})^2/(m-1) - \sigma_0^2$. Therefore, assuming $\boldsymbol{C}$ is known, we can obtain the MEQLEA test statistic given by

$$T = \frac{S^2}{\widehat{\mathrm{Var}}_0(S|\boldsymbol{G})} = \frac{[\boldsymbol{G}^T(\boldsymbol{U} - \hat{\beta}_0 \mathbf{1}_m)]^2}{\boldsymbol{G}^T(\boldsymbol{I} - \boldsymbol{H})\hat{\boldsymbol{\Sigma}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{G}}, \tag{108}$$

where $\hat{\boldsymbol{\Sigma}} = (1/n_1 + 1/n_2)\boldsymbol{C} + \hat{\sigma}_1^2\boldsymbol{I}$ is a null estimate of $\boldsymbol{\Sigma}$. Under suitable regularity conditions, significance of the test could then be assessed by comparing T to a χ_1^2

distribution.

In practice, the inter-gene covariance matrix $\boldsymbol{C}$ is usually unknown. Therefore we substitute $\boldsymbol{C}$ with $\hat{\boldsymbol{C}}$, the empirical covariance matrix of the expression data after controlling for possible DE effects by centering the expression levels of cases and controls separately around zero. Formally, $\hat{\boldsymbol{C}}$ is given by $\hat{C}_{ik} = \frac{1}{n}\sum_{j=1}^{n}(Y_{ij}-\alpha_{ij})(Y_{kj}-\alpha_{kj})$ where $\alpha_{ij} = \sum_{j':X_{j'}=X_j} Y_{ij'} / \sum_{j'=1}^{n} 1\{X_{j'} = X_j\}$ is the average expression level at gene i for all samples from the same group (either treatment or control) as sample j. In real data sets, the number of genes, m, is usually much greater than the sample size n, in which case $\boldsymbol{C}$ is a high-dimensional parameter that cannot be efficiently estimated by $\hat{\boldsymbol{C}}$. Interestingly, however, we find that the the test statistic T relies not on the accurate estimation of the entire $\boldsymbol{C}$, but only on three parameters involving $\boldsymbol{C}$, which can be much more realistically estimated by a moderate sample size. To demonstrate this, we re-arrange the order of the rows and columns of $\boldsymbol{C}$ to allow the partition $\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{12}^T & \boldsymbol{C}_{22} \end{bmatrix}$, where $\boldsymbol{C}_{11}$ is the correlation matrix for genes in the test set, $\boldsymbol{C}_{22}$ is that for gene in the background set (i.e., the complement of the test set), and $\boldsymbol{C}_{12}$ is the cross-correlation matrix between the two classes of genes. (To be continued....)

4.2.2. Simulation Methods

4.2.2.1 Simulation Setup

In this section, we will specify the parameter setup for type I error and power simulations. Let $\boldsymbol{Y}_j$ be a vector denoting the expression profile of sample j and $\mathrm{Cov}(Y_{i_1,j}, Y_{i_2,j}) = \rho_{i_1,i_2}$ for any two genes i_1 and i_2. We assume that genes have the same correlation if they are from the same category (whether the test set or the background set): $\mathrm{Cov}(Y_{i_1}, Y_{i_2}) = \rho_1$ if genes i_1 and i_2 are both from the test set (i.e., $G_{i_1} = G_{i_2} = 1$), $\mathrm{Cov}(Y_{i_1}, Y_{i_2}) = \rho_2$ if they are both from the background set (i.e., $G_{i_1} = G_{i_2} = 0$). For cross-category genes, $\mathrm{Cov}(Y_{i_1}, Y_{i_2}) = \rho_3$ if i_1 is from the test set

and i_2 is from the background set (i.e., $G_{i_1} = 1, G_{i_2} = 0$). We examine five different correlation structures, listed as follows:

(a): $\rho_1 = \rho_2 = \rho_3 = 0$; that is, the genes are independent of each other.

(b): $\rho_1 = \rho_2 = \rho_3 = 0.1$; that is, all genes are correlated, with an exchangeable correlation structure.

(c): $\rho_1 = 0.1$, $\rho_2 = \rho_3 = 0$; that is, only the genes in the test set are correlated. This corresponds to ... , and we envision what methods do well...

(d): $\rho_1 = 0.1$, $\rho_2 = 0.05$, $\rho_3 = 0$; that is, genes are correlated within the test set and within the background set, but any two genes, one from the test set and the other from the background set, are independent.

(e): $\rho_1 = 0.1$, $\rho_2 = 0.05$, $\rho_3 = -0.05$; that is, all genes are correlated, but the correlation between two genes depend on whether they belong to the test set or not.

The simulations run as follows: first, we consider an entire gene set containing $m = 500$ genes, of which $m_1 = 100$ genes are in the test set, and the remaining $m_2 = 400$ genes in the background set; second, we sample genes to be DE with probability p_t in the test set and with probability p_b in the background set, and for sampled DE genes, we simulate the DE effect Δ from a normal distribution $N(2, 1)$ (except in Table 7 we use $N(1, 0.5)$ to report calibrated power) and for non-DE genes we set $\Delta = 0$; third, we set the "true" mean expression values $\boldsymbol{\mu}_1 = \mathbf{0}_m$ and $\boldsymbol{\mu}_2 = \boldsymbol{\Delta}$, respectively, for the control and treatment groups; fourth, we simulate n_1 samples from MVN$(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ for the control group and n_2 samples from MVN$(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ for the treatment group, where the covariance $\boldsymbol{\Sigma} = [\text{Cov}(Y_{i_1}, Y_{i_2})]_{m \times m}$ may be one of the correlation structures in (a)-(e).

Further assumptions on p_t and p_b will complete our generating model used in the type I error and power simulations. (REF methods part about DE and no DE ....)

We have mentioned in the Introduction part that the test statistics correlations among genes may not equal their sample correlations when at least one gene is truly DE (see Section 4.1). Therefore, if there are DE genes in the entire gene set, approaches assuming almost equality of correlations among gene-level statistics and those among expression values may not perform well. (a heads-up on how the 2 groups are different: if non-GO-term genes

To illustrate this point, we perform two groups of simulations (A_1 and A_2) for each of (a)-(e) correlation structures. In both type I error and power simulations, we set the DE probability to be $0\%(S_0)$ in group A_1 and $10\%(S_0)$ in group A_2 for genes in the background set. That is, genes in the background set are allowed to be DE in group A_2 but not in group A_1. In the type I error simulation, we have $p_t = p_b$ under the null. In the power simulation, we considered four different scenarios for the alternative hypothesis of the presence of enrichment: for genes in the test set, we set DE probability to be $5\%(S_1), 10\%(S_2), 15\%(S_3)$ and $20\%(S_4)$ in group A_1, and $15\%(S_1), 20\%(S_2), 25\%(S_3)$ and $30\%(S_4)$ in group A_2. Table 6 summarizes the simulation setup for the two groups.

Table 6: DE probability configurations in type I error and power simulations. S_0 is for type I error simulation. S_1-S_4 represent the four scenarios considered in power simulations. p_b and p_t are the DE probability for genes in the background set and that in the test set, respectively.

Group	Background DE prob. (p_b)	DE prob. in test set (p_t)				
		S_0	S_1	S_2	S_3	S_4
A_1	0%	0%	5%	10%	15%	20%
A_2	10%	10%	15%	20%	25%	30%

4.2.2.2 Other methods considered

We will compare MEQLEA to six previously proposed gene set tests: GSEA [100], two versions of the CAMERA procedure —CAMERA-modt and CAMERA-rank [112], SigPathway [102], MRGSE [75], and QuSAGE [114]. Except SigPathway and MRGSE, all methods incorporate features intended for inter-gene correlation correction. GSEA

calculates an enrichment score for the test set by examining the ranking (according to some metric, for example, the signal-to-noise ratio) of its member genes, and determines the significance of the enrichment score by randomly permuting sample labels. CAMERA-modt uses moderated t-statistics [93] as gene-level statistics and estimates a VIF to account for inter-gene correlations in the set-level statistic, and CAMERA-rank is the rank version of the CAMERA-modt. MRGSE is a rank-based method assuming inter-gene independence, and is recommended by Tarca et al. [101] over a class of independence-assuming methods. SigPathway is a parametric version of MRGSE, and in this simulation we use the moderated t-statistics as the gene level statistics. QuSAGE generates from t-test a probability density function (PDF) for each gene, combines the individual PDFs using convolution, and quantifies enrichment of the test set by the convoluted PDF.

The software implementation is described as follows. The GSEA is modified from the original R-GSEA script (`http://software.broadinstitute.org/gsea/index.jsp`) to accommodate single gene set test. CAMERA and MRGSE are implemented in the limma package [94] in the Bioconductor project [35]. QuSAGE is available in the Bioconductor package of the same name, and SigPathway is implemented by ourselves.

In terms of type I error control and power, we expect some of the six tests to have different performances between group A_1 and A_2 simulations under one or more correlation structures.

4.3. Results

According to the simulation setup in Section 4.2.2.1, the test set is not enriched if DE probabilities are the same for genes in the test set and for those in the background set (i.e., $p_t = 0\%$ for group A_1 and $p_t = 10\%$ for group A_2), in which case we examine the type I error rate. As to power, we set DE probability according to each of the alternative scenarios S_1-S_4 (see Table 6) and calculate the proportion of data sets for

which a test would reject at a given level α. The results are based on $10,000$ simulated data sets.

4.3.1. Type I error simulations

We report the type I error simulation results for group A_1 and A_2 simulations. Figure 10 shows the uniform quantile-quantile (QQ) plots of p-values for the seven approaches (MEQLEA, SigPathway, MRGSE, CAMERA-modt, CAMERA-rank, GSEA and QuSAGE) under each of the five correlation structures (each row of plots, from top to bottom, corresponds accordingly to correlation structures (a)-(e)).

In group A_1 simulations (the left column of Figure 10), GSEA and MEQLEA hold the size of type I error rate correctly for all five correlation structures, with simulated p-values uniformly distributed on $[0, 1]$. The two versions of CAMERA control type I errors correctly for correlation structures (a) and (c), yet they are too conservative for the case of (b) and anti-conservative for correlation structures (d) and (e). SigPathway and MRGSE procedures have well-calibrated type I error for correlation structures (a) and (b), but are anti-conservative the case of (c), (d) and (e). QuSAGE has good type I error control for only (c), and is too conservative for (a), (d) and (e), and anti-conservative for (b).

In group A_2 simulations (the right column of Figure 10), MEQLEA continues to hold the size of type I error rate, whereas GSEA is skewed towards small p-values, under all five correlation structures. The two versions of CAMERA control type I error rate correctly for (a) where genes are simulated to be independent, but may be liberal in other situations. SigPathway and MRGSE have similar trends for p-values as they do, respectively, in group A_1 simulations. QuSAGE is conservative in (b) but anti-conservative in the remaining four correlation structures.

MEQLEA shows consistent accuracy for type I error control across all simulations, but the accuracy of the other six methods may be affected by two factors: the inter-gene correlation structures, and DE probability of each gene. MEQLEA controls the

size of type I error well because it uses difference in mean as gene-level statistic, and the correlations between such statistics are exactly the same as correlations between the samples (Zhuo and Di, unpublished work). GSEA evaluates the enrichment score of a test set by generating its null distribution from sample permutation. When there are no DE genes such as in the case of group A_1 simulations, GSEA performs extremely well since permuting sample labels won't change the underlying correlation structure. When DE genes exist, however, sample permutation will destroy the inter-gene correlation structure, which explains the complete failure of GSEA in controlling type I error for the case of group A_2 simulations. For CAMERA and QuSAGE, the VIF of the gene-level statistics (moderated t-test in [112]) may be over-estimated when a fraction of genes are DE (Zhuo and Di, unpublished work), and therefore the set-level test statistic is under-estimated. The performances of related methods—QuSAGE and two versions of CAMERA—are subject to the underlying correlation structures. Moreover, the performance of CAMERA is complicated by the fact that the set-level statistic takes into account only the inter-gene correlation in the test set without addressing that in the background set.

Different from the five methods mentioned above, SigPathway and MRGSE rely on independence between genes. It's not surprising that such gene permutation based methods control type I error correctly when genes are "equally-correlated": in (a) genes are simulated to be independent, and in (b) genes are simulated to have an exchangeable correlation structure. However, both SigPathway and MRGSE fail to hold type I error size for the remaining three correlation structures. These simulations show that even small inter-gene correlations (e.g., 0.05) will result in inflated type I error rate when the test does not account for inter-gene correlations.

4.3.2. Power simulation

We compare the power of MEQLEA to those of the other six methods under correlation structure (a) in which genes are simulated to be independent. Since some of

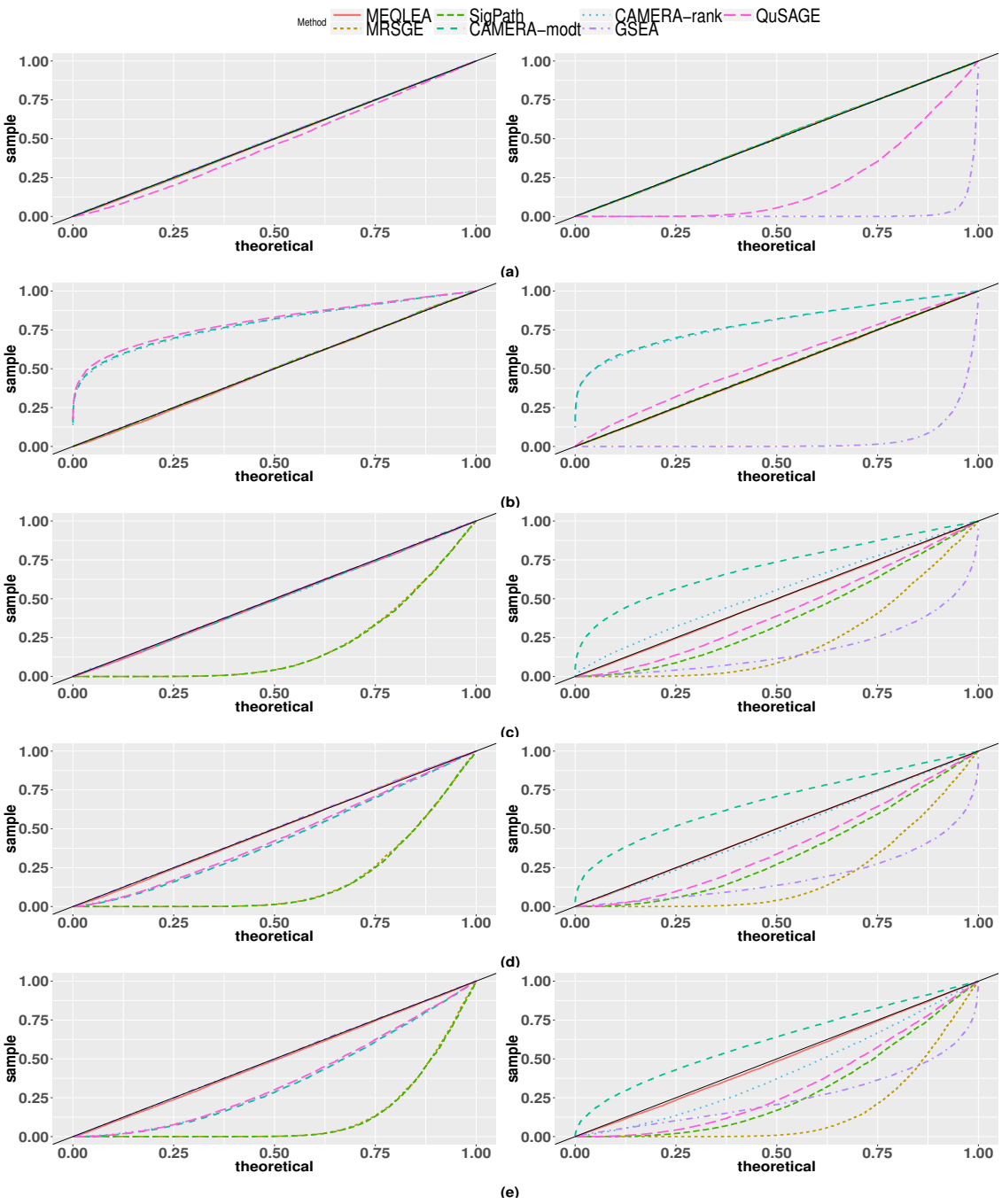

Figure 10: Uniform quantile-quantile plots for p-values by different methods. Each plot from top to bottom corresponds to correlation structures (a)-(e), respectively. The left column is for group A_1 simulation, and the right column for group A_2 simulation (see Table 6 for detail). Results are based on $10,000$ simulations.

these tests are not well calibrated at the sample size considered (see results in Section 4.3.1), we report calibrated power. For calibrated power, the critical value $c(\alpha)$ is chosen so that when the null hypothesis is true, exactly $100 \cdot \alpha\%$ of the resulting p-values are less than $c(\alpha)$; that is, $c(\alpha)$ is the α quantile of null distribution of p-values, where the null distribution is generated from simulation. Calibrated power allows a more fair comparison among tests, as tests that are too conservative under the null hypothesis will have greater power due to the tendency to produce small p-values, yet this apparent power does not truly distinguish between the null and the alternative.

Table 7 summarizes the calibrated power for the two groups of simulations (i.e., A_1 and A_2 in Table 6). For A_1 simulations, GSEA has the highest, and rank based methods (MRGSE and CAMERA-rank) have the lowest, calibrated power across all four alternative scenarios. CAMERA-modt, SigPathway and MEQLEA have no systematic difference in the calibrated power. In group A_2 simulations, GSEA shows virtually no power. MEQLEA, CAMERA-modt, and SigPathway have indistinguishable calibrated power and are among the best.

Table 7: Recalibrated power for different methods. The powers are summarized under four alternatives S_1-S_4 in each of the group A_1 and A_2 simulations (see Table 6 for detail). Results are based on $10,000$ simulations.

Group	Method	$c(\alpha)$	S_1	S_2	S_3	S_4
	MEQLEA	0.045	0.340	0.741	0.944	0.991
	MRGSE	0.051	0.111	0.284	0.533	0.766
	SigPathway	0.049	0.344	0.744	0.947	0.992
A_1	CAMERA-modt	0.051	0.336	0.737	0.943	0.990
	CAMERA-rank	0.053	0.108	0.280	0.519	0.758
	GSEA	0.051	0.517	0.894	0.989	0.999
	QuSAGE	0.028	0.385	0.784	0.959	0.995
	MEQLEA	0.050	0.180	0.478	0.777	0.939
	MRGSE	0.048	0.104	0.269	0.530	0.781
	SigPathway	0.049	0.175	0.473	0.773	0.936
A_2	CAMERA-modt	0.052	0.173	0.466	0.766	0.933
	CAMERA-rank	0.050	0.102	0.262	0.521	0.771
	GSEA	0.000	0.000	0.000	0.000	0.000
	QuSAGE	0.000	0.021	0.127	0.387	0.692

Figure 11 shows for MEQLEA, the variations in power according to different correlation structures across four alternative scenarios S_1-S_4. For each correlation structure and each alternative, we report the power (without recalibration) at a significance level of 0.05. The top is the power for group A_1, and the bottom for group A_2. The powers under correlation structures (a) and (b) are very similar, and are among the highest under each of the four alternatives. It's not surprising because they correspond to the simplest correlation structures: gene expression values in (a) are simulated to be independent and in (b) are simulated to have the same correlation 0.1. As the correlation structure becomes more complex, from (c) to (d) then to (e), the power decreases under every alternative scenario. The power under correlation structure (e) is the lowest for both A_1 and A_2 simulations.

4.3.3. Real Data

We apply MEQLEA to two example data sets, and compare the lists of enriched gene sets to those obtained by other three methods (GSEA, CAMERA-modt and MRGSE). Our results lend credence to previous studies in finding potential gene sets correlated with Huntington's disease and those correlated with chromosome Y and Y bands in lymphoblastoid cells.

4.3.3.1 Huntington's Disease Data

We examine the Huntington's Disease (HD) RNA-Sequencing (RNA-Seq) data [56] to identify enriched gene sets that are potentially responsible for HD. The mRNA expression profiles in human prefrontal cortex were obtained from 20 Huntington's Disease samples and 49 neurologically normal controls. Expression values are normalized and filtered as described in the methods section of Labadorf et al. [56]. The data, containing $28,087$ genes, is available as a series GSE64810 in the GEO database (`http://www.ncbi.nlm.nih.gov/geo/`). For each gene, we adjust for two covariates—age at death (DeathAge) and RNA Integrity Number (RIN), as also done

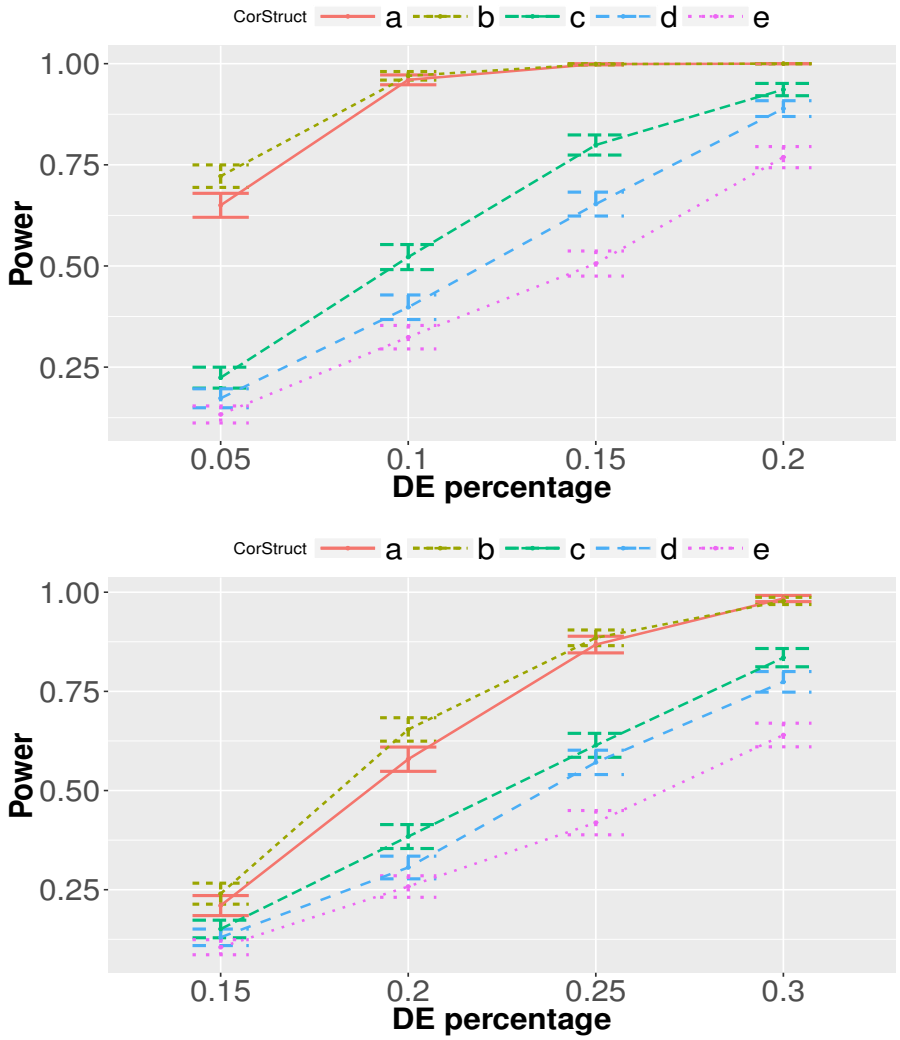

Figure 11: Power for MEQLEA under correlation structures (a)-(e) of Section 4.2.2.1. The top corresponds to group A_1 simulations, and the bottom to group A_2 simulations (see Table 6). The error bars are the 95% CIs based on $10,000$ simulations.

by Labadorf et al. [56]. We follow their strategy of treating the two covariates as categorical. Briefly, DeathAge is binned into intervals 0-45, 46-60, 61-75, 76-90 and 90+, and RIN is dichotomized as $>$ or ≤ 7. We regress the normalized expression levels on AgeDeath and RIN and use the resulting residuals as the *covariate-adjusted expression levels*.

We perform enrichment analysis on the covariate-adjusted data using the MsigDB

[100] C2 Canonical Pathways (February 5, 2016, data last accessed). The C2 Canonical Pathways have a collection of 1330 gene sets, with an average set size of 50 (the set sizes range from 3 to 1028, and the median is 29). Since the genes are named by HGNC symbols in C2 and by ensembl IDs in the HD expression data set, we convert the ensembl IDs in the expression data into HGNC symbols using *BioMart* (`http://uswest.ensembl.org/biomart/martview/`). We retain $26,941$ genes that have corresponding HGNC symbols.

We apply four test procedures (MEQLEA, GSEA, CAMERA-modt and MRGSE) to run enrichment analysis for the entire C2 Canonical Pathways, and compared the four tests in terms of resulting enriched gene sets. We use the Benjamini-Hochberg [10] procedure (BH) to control the false discovery rate (FDR) for multiple hypothesis testing (unless specified otherwise, all p-values in Section 4.3.3 are adjusted by BH procedure). The BH procedure is used when the test statistics under the null have non-negative correlations [11]. We note that since many pathways have overlapped genes, the BH procedure should be appropriate in our study.

In Figure 12 we plot $\log 10$ p-values of MEQLEA against those of GSEA, CAMERA-modt and MRGSE. The p-values of CAMERA-modt are overwhelmingly larger than their counterparts of GSEA or MEQLEA, yet smaller than those of MRGSE, even if p-values between MEQLEA and other three methods are highly correlated (Pearson's correlation of $\log 10$ p between MEQLEA and GSEA, CAMERA-modt and MRGSE are 0.90, 0.96, and 0.87 respectively). This trend of p-values is consistent with our earlier simulation (see results in simulation section 4.3.1) that CAMERA-modt could produce large p values. The p-values of MRGSE are in general smaller than the corresponding p-values of MEQLEA, leading to more significant calls.

Using MEQLEA, we find 89 significant signals out of the entire 1330 gene sets at FDR level of 0.05. GSEA finds 3 enriched gene sets—2 of them were also among those 89 gene sets (the one that is not significant according to MEQLEA had a p-value of 0.013 and FDR 0.100). MRGSE identified 387 gene sets which include all the 89

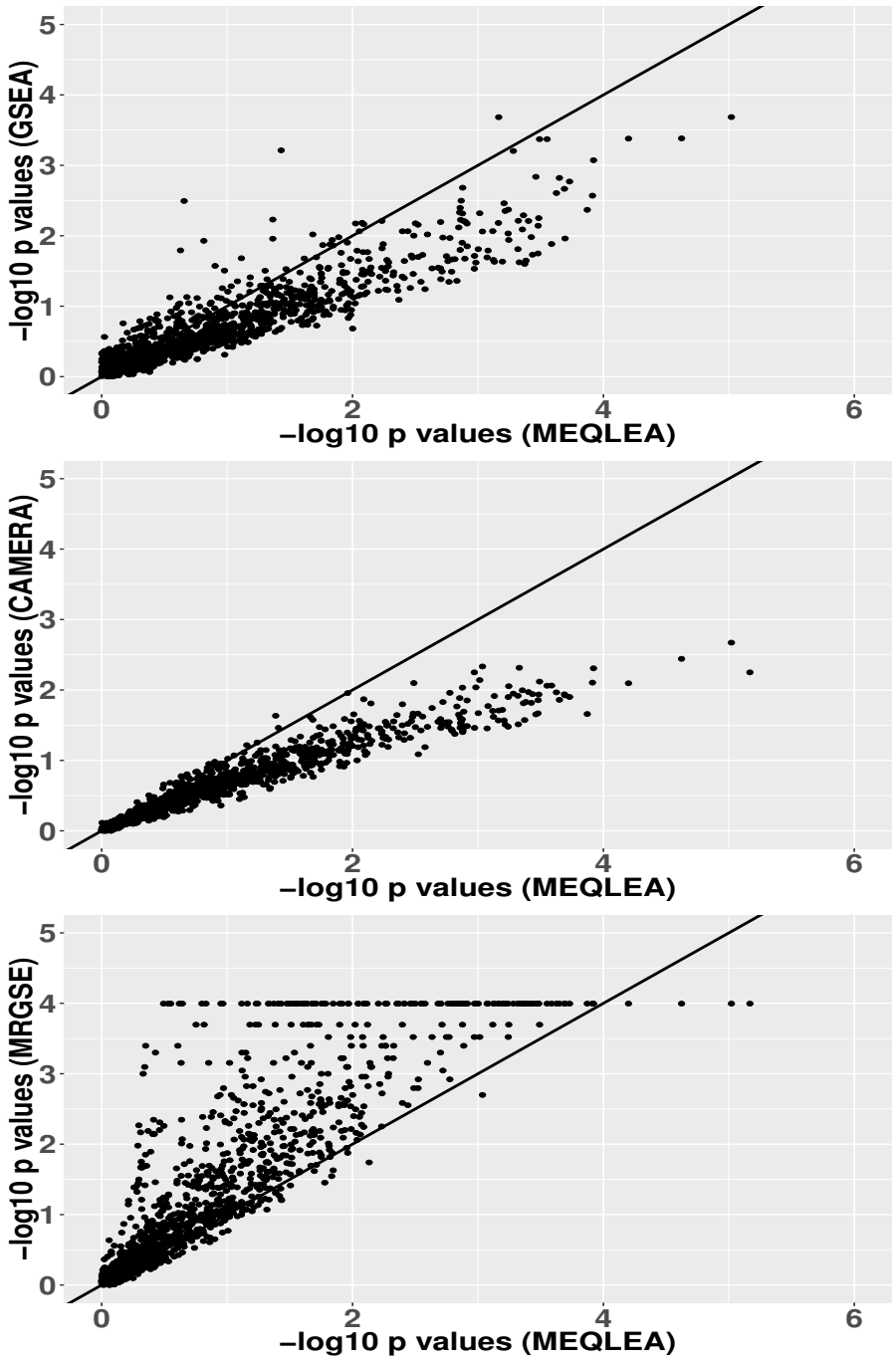

Figure 12: Pairwise comparisons of p-values for MEQLEA, GSEA, and CAMERA-modt. The p-values are reported from enrichment test of each gene set in the C2 Canonical Pathway gene sets.

sets MEQLEA identified, and CAMERA-modt identified none. Originally, Labadorf et al. [56] used the same HD data set to conduct enrichment analysis using topGo [1]. They note that the enriched gene sets they identified show a clear immune response and inflammation-related pattern, including "REACTOME INNATE IMMUNE SYSTEM", "PID IL4 2PATHWAY", and "PID NFKAPPAB CANONICAL PATHWAY". These three gene sets rank (by nominal p-values) 18,10 and 3 respectively in the 89 enriched gene sets.

In Table 8, we report the top 30 enriched gene sets (ordered by nominal p values) identified using MEQLEA. We also label the enriched gene sets from GSEA by "$*$" in the table. Many of our enriched gene sets have been shown to be closely related to HD pathogenesis. For example, the top enriched gene set by MEQLEA, "PID SMAD2 3NUCLEAR PATHWAY", is responsible for regulation of nuclear SMAD2/3 signaling. Katsuno et al. [53] showed that nuclear SMAD2/3 are related to polyglutamine disease, which includes HD. The third enriched gene set, "PID NFKAPPAB CANONICAL PATHWAY", is a canonical NF-kappaB pathway, and its dysregulation causes HD immune dysfunction [103]. Also, Marcora and Kennedy [69] found that reduced transport of NF-kappaB out of dendritic spines and its activity in neuronal nuclei may contribute to the etiology of HD. Another gene set, "REACTOME INNATE IMMUNE SYSTEM", contributes to HD pathogenesis [103, 56]. Chiang et al. [19] demonstrated that the systematic downregulation of PPARγ, related to "BIOCARTA PPARA PATHWAY", seems to play a critical role in the dysregulation of energy homeostasis observed in HD, and that PPARγ is a potential therapeutic target for this disease. For "PID P53 DOWNSTREAM PATHWAY", Ghose et al. [36] showed the likely involvement of NFkB (RelA), p53 and miRNAs in the regulation of cell death in HD pathogenesis.

Table 8: Enriched gene sets (ordered by nomial p-values) identified by MEQLEA for HD data. The $\hat{\rho}_1$, $\hat{\rho}_2$ and $\hat{\rho}_3$, respectively, are the average estimated sample correlation between genes in the test set, between genes in the background set, and between two genes–one from the test set and the other from the background set. The enriched gene sets are noted by "$*$" for GSEA. No gene set was identified as enriched by CAMERA-modt and all the 30 gene sets are also identified as enriched by MRGSE. For all methods, a gene set is called significant when its FDR using Benjamini-Hochberg (BH) correction is < 0.05.

Gene Set	Size	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	p-value	FDR	
PID SMAD2 3NUCLEAR PATHWAY	79	0.063	0.013	0.015	5.8E-06	5.7E-03	$*$
REACTOME YAP1 AND WWTR1 TAZ STIMULATED GENE EXPRESSION	23	0.121	0.013	0.014	8.5E-06	5.7E-03	
PID NFKAPPAB CANONICAL PATHWAY	22	0.127	0.013	0.019	2.3E-05	1.0E-02	
BIOCARTA NTHI PATHWAY	23	0.130	0.013	0.023	6.2E-05	2.1E-02	
BIOCARTA TID PATHWAY	18	0.101	0.013	0.012	1.2E-04	2.2E-02	
PID HIV NEF PATHWAY	35	0.065	0.013	0.013	1.2E-04	2.2E-02	
KEGG PATHWAYS IN CANCER	311	0.028	0.013	0.010	1.3E-04	2.2E-02	
PID MYC REPRESS PATHWAY	60	0.057	0.013	0.013	1.9E-04	2.2E-02	
BIOCARTA TOLL PATHWAY	36	0.083	0.013	0.018	2.0E-04	2.2E-02	
PID IL4 2PATHWAY	59	0.081	0.013	0.010	2.0E-04	2.2E-02	
KEGG TGF BETA SIGNALING PATHWAY	82	0.055	0.013	0.011	2.2E-04	2.2E-02	
BIOCARTA DEATH PATHWAY	33	0.067	0.013	0.013	2.4E-04	2.2E-02	
KEGG NOD LIKE RECEPTOR SIGNALING PATHWAY	55	0.045	0.013	0.008	2.6E-04	2.2E-02	
BIOCARTA CTCF PATHWAY	23	0.083	0.013	0.015	2.8E-04	2.2E-02	
ST TUMOR NECROSIS FACTOR PATHWAY	28	0.031	0.013	0.014	3.2E-04	2.2E-02	
BIOCARTA TNFR2 PATHWAY	17	0.151	0.013	0.022	3.3E-04	2.2E-02	
KEGG APOPTOSIS	82	0.036	0.013	0.008	3.3E-04	2.2E-02	
REACTOME INNATE IMMUNE SYSTEM	209	0.039	0.013	0.009	3.3E-04	2.2E-02	
PID HES HEY PATHWAY	47	0.071	0.013	0.019	3.4E-04	2.2E-02	
REACTOME DOWNSTREAM TCR SIGNALING	31	0.082	0.013	0.011	3.7E-04	2.2E-02	
PID TCPTP PATHWAY	42	0.076	0.013	0.010	3.7E-04	2.2E-02	
BIOCARTA 41BB PATHWAY	14	0.110	0.013	0.023	3.9E-04	2.2E-02	
PID FRA PATHWAY	34	0.154	0.013	0.008	4.1E-04	2.2E-02	
PID P53 DOWNSTREAM PATHWAY	131	0.045	0.013	0.012	4.2E-04	2.2E-02	
PID EPO PATHWAY	34	0.069	0.013	0.013	4.3E-04	2.2E-02	
BIOCARTA PPARA PATHWAY	53	0.031	0.013	0.008	4.4E-04	2.2E-02	
BIOCARTA EPONFKB PATHWAY	11	0.068	0.013	0.010	4.7E-04	2.2E-02	
BIOCARTA HIVNEF PATHWAY	58	0.063	0.013	0.019	4.8E-04	2.2E-02	
BIOCARTA CD40 PATHWAY	13	0.165	0.013	0.026	4.8E-04	2.2E-02	
BIOCARTA IL7 PATHWAY	17	0.100	0.013	0.016	5.2E-04	2.3E-02	

4.3.3.2 Male vs Female Lymphoblastoid Cells Data

We analyze the mRNA expression profiles from lymphoblastoid cell lines derived from 17 females and 15 males. Subramanian et al. [100] examined this data set with their GSEA method, testing the enrichment of the cytogenetic gene sets (C1). The C1 includes 24 sets, one for each of the 24 human chromosomes, and 295 sets corresponding to cytogenetic bands. For the comparison "male VS female", they expected to find gene sets on chromosome Y, not on chromosome X. We run enrichment analysis with the four tests (MEQLEA, GSEA,CAMERA-modtand MRGSE). In Table 9, we summarize all the gene sets that are called significant at FDR level 0.05. Unanimously, three gene sets—"chrY", "chrYq11" and "chrYp11"—are found to be enriched by all of the four methods. It is interesting to note that only MEQLEA is able to identify another Y band, "chrYp22", as enriched. In fact, these four gene sets are the only four pathways containing at least 3 genes in C1 and corresponding to chromosome Y or Y bands. MEQLEA does not produce small p-value (< 0.01) for the remaining three gene sets in Table 9, which is just as expected in that study.

Table 9: Enriched gene sets and their nominal p values for lymphoblastoid cells data. Reported are gene sets with FDR$<$ 0.05 for at least one of the MEQLEA, GSEA, CAMERA-modt and MRGSE methods using Benjamini-Hochberg(BH) procedure.

Gene set	Size	MEQLEA	GSEA	CAMERA-modt	MRGSE
chrY	40	0.0E+00	0.0E+00	1.0E-05	5.9E-07
chrYq11	16	0.0E+00	0.0E+00	7.2E-08	8.5E-06
chrYp11	18	2.1E-15	0.0E+00	2.8E-04	5.1E-04
chrYp22	8	3.6E-04	1.2E-02	1.0E-02	1.3E-02
chr6	614	5.6E-02	6.0E-01	6.1E-01	2.1E-04
chr1	1104	6.1E-02	5.5E-01	6.3E-01	5.3E-05
chr12	571	8.7E-02	2.6E-01	4.0E-01	5.1E-09

4.4. Conclusion and Discussion

MEQLEA is a mixed effects quasi-likelihood model for competitive gene set test. It effectively adjusts for completely unknown, unstructured correlations among the genes.

It uses a score test approach and allows for analytical assessment of p-values. Compared to existing approaches, MEQLEA controls type I error correctly and maintains good power under different correlation structures.

Under competitive gene set test framework, a number of methods have been proposed to account for correlation among genes. One approach is to evaluate the set-level statistic by permuting sample labels to generate the null distribution, as adopted by the widely used GSEA [100]. However, sample permutation method has been criticized for altering the null hypotheses being tested [37, 54]. Instead, CAMERA [112] proposed to correct for the correlation among genes by estimating a VIF directly from the data. Incorporating VIF into set level test statistic has also been used by Yaari et al. [114] for their QuSAGE procedure. The major problem with this VIF approach is that it tries to estimate correlations among gene-level test statistics directly from sample correlation. Zhuo and Di have argued (unpublished work) that the correlations among gene-level statistics are not necessarily equal to those among samples due to the presence of DE genes. The estimated VIF could be biased without taking into account such a discrepancy, which will undermine the performance of CAMERA and QuSAGE. In contrast, MEQLEA avoids the discrepancy by using the differences in mean as gene-level statistics for a two group comparison experiment. It models the covariance of gene-level statistics by two variance components, one attributable to correlations among samples after treatment effect removed, and the other attributable to the DE effect associate with the treatment. We note that for MEQLEA, the estimation of covariance among gene-level statistics need not be exact: MEQLEA uses a score test that involves linear combinations of the entries of the covariance matrix. The denominator in the score test statistic (see equation (108)) can usually be accurately approximated given the high dimensionality of the covariance matrix. MEQLEA is based on quasi-likelihood, therefore it does not require normal assumption of expression data, and could be applied to both microarray and RNA-Seq experiments.

We compare the performance of MEQLEA to those of other existing approaches through both simulation study and real data analysis. In the simulation study, we examine the calibration of MEQLEA and other six method (SigPathway, MRGSE, CAMERA-modt, CAMERA-rank, GSEA and QuSAGE) in terms of type I error control and power. We demonstrate that MEQLEA holds correct type I error size under all correlation structures considered, whereas all other methods may fail in one or more situations. MEQLEA is also among the best for power performance under independent correlation structure. In the real data analysis, we run enrichment analysis using four methods—MEQLEA, CAMERA-modt, GSEA and MRGSE on two data sets. The p-values of MEQLEA are smaller than those produced by GSEA and CAMERA (which are intended to adjust for inter-gene correlation), but larger than those of MRGSE (which assumes independence between genes). MEQLEA is able to identify a moderate size of enriched gene sets, some of which are insightful in the corresponding studies yet are not found by other three methods.

Currently, MEQLEA only supports enrichment test for two-group comparisons. In many gene expression experiments, however, researchers might use more complex design to study different factors of interest, in which case a linear model would be more appropriate. Our future work will focus on generalizing MEQLEA to allow for more complicated design structures.

The R codes for reproducing results in this paper are available at `https://github.com/zhuob/EnrichmentAnalysis`.

Acknowledgments

We thank Yanming Di, Sarah Emerson and Wanli Zhang for helpful discussion in preparing this manuscript. We thank Dr. Adam Labadorf for providing information about the HD gene expression data. This article is part of doctor dissertation written by BZ under the supervision of YD.

Conflict of interest statement.

None declared.

Appendix

Standardization Standardization for each gene: first, we obtain the residuals by subtracting off the means within each treatment group;

$$r_{ijk} = y_{ijk} - \sum_{j=1}^{n_k} y_{ijk}/n_k; \tag{109}$$

then we calculate the pooled standard deviation from the residuals,

$$s_i = std(r_{ijk}); \tag{110}$$

next we get the standardized expression by dividing the original expression levels by the standard deviation,

$$y^*_{ijk} = y_{ijk}/s_i \tag{111}$$

We perform the standardization procedure to every gene in the data set.

Calculating covariance matrix for test statistics First $E(\Delta_i) = E(Z_i\delta_i) = E(Z_i)E(\delta_i) = p_i\mu_\delta$. Next note that

$$\begin{aligned}
\text{Var}(\Delta_i) &= E[(Z_i\delta_i)^2] - [E(Z_i\delta_i)]^2 \\
&= \text{Var}(Z_i)[E(\delta_i)]^2 + \left[(EZ_i)^2 + \text{Var}(Z_i)\right]\text{Var}(\delta_i) \\
&= p_i\sigma^2_\delta + p_i(1-p_i)\mu^2_\delta
\end{aligned}$$

Let $T_i = \bar{Y}_{i,2} - \bar{Y}_{i,1}$ be the difference in mean expression levels between the treatment group and the control group. We have

$$E(T_i) = E(\bar{Y}_{i,2}) - E(\bar{Y}_{i,1}) = E(\Delta_i) = E(Z_i\delta_i) = p_i\mu_\delta$$

The covariance between two genes i_1 and i_2 is given by (I HAVE CONCERNS HERE, IS IT VALID TO ASSUME THAT DE EFFECTS ARE INDEPENDENT BE-

TWEEN GENES? WE SEE CO-EXPRESSION!! OR WE'VE ALREADY TAKEN THAT INTO ACCOUNT BY CORRELATION BETWEEN GENES"),

$$\begin{aligned}\mathrm{Cov}(T_{i_1}, T_{i_2}) &= E\left[\mathrm{Cov}(T_{i_1}, T_{i_2}|\Delta_{i_1}, \Delta_{i_2})\right] \\ &\quad + \mathrm{Cov}\left[E(T_{i_1}|\Delta_{i_1}), E(T_{i_2}|\Delta_{i_2})\right] \\ &= E\left(\frac{1}{n_1}\rho_{i_1,i_2} + \frac{1}{n_2}\rho_{i_1,i_2}\right) + \mathrm{Cov}(\Delta_{i_1}, \Delta_{i_2}) \\ &= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\rho_{i_1,i_2}\end{aligned} \tag{112}$$

For gene i, the variance $\mathrm{Var}(T_i) = \mathrm{Var}(\bar{Y}_{i,1}) + \mathrm{Var}(\bar{Y}_{i,2})$, with

$$\mathrm{Var}(\bar{Y}_{i,1}) = \frac{1}{n_1}$$

$$\begin{aligned}\mathrm{Var}(\bar{Y}_{i,2}) &= \frac{1}{n_2^2}\left[\sum_{j=1}^{n_2}\mathrm{Var}(Y_{ij2}) + 2\sum_{1\le j_1<j_2\le n_2}\mathrm{Cov}(Y_{ij_12}, Y_{ij_22})\right] \\ &= \frac{1}{n_2}\mathrm{Var}(Y_{ij2}) + \frac{n_2-1}{n_2}\mathrm{Cov}(Y_{ij_12}, Y_{ij_22}) \\ &= \frac{1}{n_2}\left[E\left(\mathrm{Var}(Y_{ij2}|\Delta_i)\right) + \mathrm{Var}\left(E(Y_{ij2}|\Delta_i)\right)\right] \\ &\quad + \frac{n_2-1}{n_2}E\left(\mathrm{Cov}(Y_{ij_12}, Y_{ij_22}|\Delta_i)\right) \\ &\quad + \frac{n_2-1}{n_2}\mathrm{Cov}\left(E(Y_{ij_12}|\Delta_i), E(Y_{ij_22}|\Delta_i)\right) \\ &= \frac{1}{n_2} + \mathrm{Var}(\Delta_i)\end{aligned} \tag{113}$$

Therefore $\mathrm{Var}(T_i) = \frac{1}{n_1} + \frac{1}{n_2} + \mathrm{Var}(\Delta_i)$, and it follows that

$$\mathrm{Cov}(\boldsymbol{T}) = \boldsymbol{D} + \sigma_2^2\boldsymbol{C} \tag{114}$$

where $\boldsymbol{D}$ is a diagonal matrix with $\mathrm{Var}(\Delta_i) = p_i\sigma_\delta^2 + p_i(1-p_i)\mu_\delta^2$ as its ith diagonal element, and $\sigma_2^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$.

case	MEQLEA	MRGSE	SigPathway	CAMERA-modt	CAMERA-rank	GSEA	QuSAGE
a0PCT	0.056	0.049	0.051	0.049	0.047	0.049	0.078
a10PCT	0.050	0.052	0.051	0.048	0.050	0.946	0.491
b0PCT	0.059	0.050	0.051	0.000	0.000	0.048	0.000
b10PCT	0.052	0.051	0.051	0.000	0.000	0.837	0.027
c0PCT	0.056	0.513	0.517	0.051	0.044	0.051	0.052
c10PCT	0.054	0.442	0.188	0.000	0.021	0.290	0.131
d0PCT	0.059	0.586	0.594	0.114	0.104	0.051	0.106
d10PCT	0.052	0.522	0.235	0.001	0.049	0.220	0.175
e0PCT	0.058	0.674	0.679	0.213	0.197	0.053	0.203
e10PCT	0.054	0.614	0.334	0.004	0.116	0.113	0.267

5. Conclusion

6. Appendix

Chapter 3.

References

[1] Alexa, A. and Rahnenfuhrer, J. (2010). topgo: enrichment analysis for gene ontology. *R package version*, 2(0).

[2] Anders, S. (2010). HTSeq: Analysing high-throughput sequencing data with Python. *URL http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html*.

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

[4] Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.*, 8(9):1765–1786.

[5] Andersen, C. L., Jensen, J. L., and Ørntoft, T. F. (2004). Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res.*, 64(15):5245–5250.

[6] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene Ontology: tool for the unification of biology. *Nat. Genet.*, 25(1):25–29.

[7] Baron, K. N., Schroeder, D. F., and Stasolla, C. (2012). Transcriptional response of abscisic acid (ABA) metabolism and transport to cold and heat stress applied at the reproductive stage of development in Arabidopsis thaliana. *Plant Sci.*, 188:48–59.

[8] Barry, W. T., Nobel, A. B., and Wright, F. A. (2008). A statistical framework for testing functional categories in microarray data. *Ann. Appl. Statist.*, pages 286–315.

[9] Bates, D., Maechler, M., and Bolker, B. (2012). lme4: Linear mixed-effects models using s4 classes.

[10] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy. Stat. Soc. B Met.*, pages 289–300.

[11] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, pages 1165–1188.

[12] Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.*, 24(3):127–135.

[13] Bournier, M., Tissot, N., Mari, S., Boucherez, J., Lacombe, E., Briat, J.-F., and Gaymard, F. (2013). Arabidopsis ferritin 1 (AtFer1) gene regulation by the phosphate starvation response 1 (AtPHR1) transcription factor reveals a direct molecular link between iron and phosphate homeostasis. *J. Biol Chem.*, 288(31):22670–22680.

[14] Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, 88(421):9–25.

[15] Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82(1):81–91.

[16] Burrows, M. and Wheeler, D. (1994). A block-sorting lossless data compression algorithm. In *DIGITAL SRC RESEARCH REPORT*. Citeseer.

[17] Bustin, S. (2002). Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. *J. Mol. Endocrinol.*, 29(1):23–39.

[18] Casassola, A., Brammer, S. P., Chaves, M. S., Ant, J., Grando, M. F., et al. (2013). Gene expression: A review on methods for the study of defense-related gene differential expression in plants. *Am. J. Plant. Sci.*, 2013.

[19] Chiang, M.-C., Chen, C.-M., Lee, M.-R., Chen, H.-W., Chen, H.-M., Wu, Y.-S., Hung, C.-H., Kang, J.-J., Chang, C.-P., Chang, C., et al. (2010). Modulation of energy deficiency in Huntington's disease via activation of the peroxisome proliferator-activated receptor gamma. *Hum. Mol. Genet.*, page ddq322.

[20] Czechowski, T., Stitt, M., Altmann, T., Udvardi, M. K., and Scheible, W.-R. (2005). Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Plant Physiol.*, 139(1):5–17.

[21] Dekkers, B. J., Willems, L., Bassel, G. W., van Bolderen-Veldkamp, R. M., Ligterink, W., Hilhorst, H. W., and Bentsink, L. (2012). Identification of reference genes for RT–qPCR expression analysis in Arabidopsis and tomato seeds. *Plant Cell Physiol.*, 53(1):28–37.

[22] Di, Y., Schafer, D. W., Cumbie, J. S., and Chang, J. H. (2011). The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat. Appl. Genet. Mol. Biol.*, 10(1):1–28.

[23] Di, Y., Schafer, D. W., and Di, M. Y. (2014). Package 'NBPSeq'. *Mol. Biol.*, 10:1.

[24] Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.*, 14(6):671–683.

[25] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

[26] Efron, B. (2004). Large-scale simultaneous hypothesis testing. *J. Amer. Statist. Assoc.*, 99(465).

[27] Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Am. Stat. Assoc.*, 102(477).

[28] Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

[29] Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *Ann. Appl. Statist.*, pages 107–129.

[30] Fernandes, J. M., Mommens, M., Hagen, Ø., Babiak, I., and Solberg, C. (2008). Selection of suitable reference genes for real-time PCR studies of Atlantic halibut development. *Comp. Biochem. Phys. B*, 150(1):23–32.

[31] Finotello, F. and Di Camillo, B. (2015). Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief. Funct. Genomics.*, 14(2):130–142.

[32] Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, pages 507–521.

[33] Frericks, M. and Esser, C. (2008). A toolbox of novel murine house-keeping genes identified by meta-analysis of large scale gene expression profiles. *BBA-Gene Regul. Mech.*, 1779(12):830–837.

[34] Gatti, D. M., Barry, W. T., Nobel, A. B., Rusyn, I., and Wright, F. A. (2010). Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics*, 11(1):574.

[35] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open

software development for computational biology and bioinformatics. *Genome Biol.*, 5(10):R80.

[36] Ghose, J., Sinha, M., Das, E., Jana, N. R., and Bhattacharyya, N. P. (2011). Regulation of miR-146a by RelA/NFkB and p53 in ST Hdh Q111/Hdh Q111 Cells, a Cell Model of Huntington's Disease. *PLoS One*, 6(8):e23837.

[37] Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987.

[38] Goeman, J. J., Oosting, J., Cleton-Jansen, A.-M., Anninga, J. K., and Van Houwelingen, H. C. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21(9):1950–1957.

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

[40] Gur-Dedeoglu, B., Konu, O., Bozkurt, B., Ergul, G., Seckin, S., and Yulug, I. G. (2009). Identification of endogenous reference genes for qRT-PCR analysis in normal matched breast tumor tissues. *Oncol. Res.*, 17(8):353–365.

[41] Hadfield, J. D. et al. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.*, 33(2):1–22.

[42] Hansen, K. D., Irizarry, R. A., and Zhijin, W. (2012). Removing technical variability in RNA-Seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216.

[43] Hatem, A., Bozdağ, D., Toland, A. E., and Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, 14(1):1.

[44] Hong, S. M., Bahn, S. C., Lyu, A., Jung, H. S., and Ahn, J. H. (2010). Identification and testing of superior reference genes for a starting pool of transcript normalization in Arabidopsis. *Plant Cell Physiol.*, 51(10):1694–1706.

[45] Howald, C., Tanzer, A., Chrast, J., Kokocinski, F., Derrien, T., Walters, N., Gonzalez, J. M., Frankish, A., Aken, B. L., Hourlier, T., et al. (2012). Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res.*, 22(9):1698–1710.

[46] Hruz, T., Wyss, M., Docquier, M., Pfaffl, M. W., Masanetz, S., Borghi, L., Verbrugghe, P., Kalaydjieva, L., Bleuler, S., Laule, O., et al. (2011). Refgenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. *BMC Genomics*, 12(1):156.

[47] Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37(1):1–13.

[48] Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H. C., et al. (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.*, 35(suppl 2):W169–W175.

[49] Huang, Y.-T. and Lin, X. (2013). Gene set analysis using variance component tests. *BMC Bioinformatics*, 14(1):210.

[50] Huggett, J., Dheda, K., Bustin, S., and Zumla, A. (2005). Real-time RT-PCR normalisation; strategies and considerations. *Genes Immun.*, 6(4):279–284.

[51] Joarder, A. H. (2009). Moments of the product and ratio of two correlated chi-square variables. *Stat. Pap.*, 50(3):581–592.

[52] Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30.

[53] Katsuno, M., Adachi, H., Minamiyama, M., Waza, M., Doi, H., Kondo, N., Mizoguchi, H., Nitta, A., Yamada, K., Banno, H., et al. (2010). Disrupted transforming growth factor-β signaling in spinal and bulbar muscular atrophy. *J. Neurosci.*, 30(16):5702–5712.

[54] Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS. Comput. Biol.*, 8(2):e1002375.

[55] Kim, S.-Y. and Volsky, D. J. (2005). PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6(1):144.

[56] Labadorf, A., Hoss, A. G., Lagomarsino, V., Latourelle, J. C., Hadzi, T. C., Bregu, J., MacDonald, M. E., Gusella, J. F., Chen, J.-F., Akbarian, S., et al. (2015). RNA sequence analysis of human huntington disease brain reveals an extensive increase in inflammatory and developmental gene expression. *PloS One*, 10(12):e0143563.

[57] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359.

[58] Langmead, B., Trapnell, C., Pop, M., Salzberg, S. L., et al. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol.*, 10(3):R25.

[59] Lee Rodgers, J. and Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *Am. Stat.*, 42(1):59–66.

[60] Leinonen, R., Sugawara, H., and Shumway, M. (2010). The sequence read archive. *Nuc. Acids Res.*, page gkq1019.

[61] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

[62] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[63] Li, H. and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.*, 11(5):473–483.

[64] Liao, Y., Smyth, G. K., and Shi, W. (2013a). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, page btt656.

[65] Liao, Y., Smyth, G. K., and Shi, W. (2013b). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nuc. Acids Res.*, 41(10):e108–e108.

[66] Littell, R. C., Stroup, W. W., Milliken, G. A., Wolfinger, R. D., and Schabenberger, O. (2006). *SAS for mixed models*. SAS institute.

[67] Liu, Q. and Pierce, D. A. (1994). A note on Gauss—Hermite quadrature. *Biometrika*, 81(3):624–629.

[68] Lovén, J., Orlando, D. A., Sigova, A. A., Lin, C. Y., Rahl, P. B., Burge, C. B., Levens, D. L., Lee, T. I., and Young, R. A. (2012). Revisiting global gene expression analysis. *Cell*, 151(3):476–482.

[69] Marcora, E. and Kennedy, M. B. (2010). The Huntington's disease mutation impairs Huntingtin's role in the transport of NF-κB from the synapse to the nucleus. *Hum. Mol. Genet.*, 19(22):4373–4384.

[70] Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18(9):1509–1517.

[71] Martin, A. D., Quinn, K. M., and Park, J. H. (2011). Mcmcpack: Markov chain monte carlo in R.

[72] McCullagh, P. and Nelder, J. A. (1989). Generalized linear models.

[73] McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.*, 92(437):162–170.

[74] McCulloch, C. E. and Neuhaus, J. M. (2001). *Generalized linear mixed models*. Wiley Online Library.

[75] Michaud, J., Simpson, K. M., Escher, R., Buchet-Poyau, K., Beissbarth, T., Carmichael, C., Ritchie, M. E., Schütz, F., Cannon, P., Liu, M., et al. (2008). Integrative analysis of RUNX1 downstream pathways and target genes. *BMC Genomics*, 9(1):363.

[76] Mishra, P., Törönen, P., Leino, Y., and Holm, L. (2014). Gene set analysis: limitations in popular existing methods and proposed improvements. *Bioinformatics*, 30(19):2747–2756.

[77] Myers, R. H., Montgomery, D. C., Vining, G. G., and Robinson, T. J. (2012). *Generalized linear models: with applications in engineering and the sciences*, volume 791. John Wiley & Sons.

[78] Oberg, A. L., Bot, B. M., Grill, D. E., Poland, G. A., and Therneau, T. M. (2012). Technical and biological variance structure in mRNA-Seq data: life in the real world. *BMC Genomics*, 13(1):304.

[79] Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J. Comp. Graph. Stat.*, 4(1):12–35.

[80] Pinheiro, J. C. and Chao, E. C. (2006). Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *J. Comp. Graph. Stat.*, 15(1).

[81] Qiu, X., Brooks, A. I., Klebanov, L., and Yakovlev, A. (2005a). The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics*, 6(1):120.

[82] Qiu, X., Klebanov, L., and Yakovlev, A. (2005b). Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Stat. Appl. Genet. Mol. Biol.*, 4(1).

[83] R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[84] Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, 14(9):R95.

[85] Reid, K. E., Olsson, N., Schlosser, J., Peng, F., and Lund, S. T. (2006). An optimized grapevine RNA isolation procedure and statistical determination of reference genes for real-time RT-PCR during berry development. *BMC Plant Biol.*, 6(1):27.

[86] Rencher, A. C. and Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons.

[87] Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-Seq data using factor analysis of control genes or samples. *Nat. Biotech.*, 32(9):896–902.

[88] Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, 12(1):480.

[89] Robinson, M. D., Oshlack, A., et al. (2010). A scaling normalization method for differential expression analysis of RNA-Seq data. *Genome Biol.*, 11(3):R25.

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

[91] Seyednasrollah, F., Laiho, A., and Elo, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.*, 16(1):59–70.

[92] Shi, W. and Liao, Y. (2013). Subread/Rsubread users guide.

[93] Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3, Article3.

[94] Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer.

[95] Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1):1.

[96] Stamova, B. S., Apperson, M., Walker, W. L., Tian, Y., Xu, H., Adamczy, P., Zhan, X., Liu, D.-Z., Ander, B. P., Liao, I. H., et al. (2009). Identification and validation of suitable endogenous reference genes for gene expression studies in human peripheral blood. *BMC Med. Genom.*, 2(1):49.

[97] Stan Development Team (2016). *RStan: the R interface to Stan, Version 2.9.0.*

[98] Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.*, pages 2013–2035.

[99] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, 100(16):9440–9445.

[100] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A*, 102(43):15545–15550.

[101] Tarca, A. L., Bhatti, G., and Romero, R. (2013). A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS one*, 8(11):e79217.

[102] Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci. U.S.A*, 102(38):13544–13549.

[103] Träger, U., Andre, R., Lahiri, N., Magnusson-Lind, A., Weiss, A., Grueninger, S., McKinnon, C., Sirinathsinghji, E., Kahlon, S., Pfister, E. L., et al. (2014). HTT-lowering reverses Huntington's disease immune dysfunction caused by NFκB pathway dysregulation. *Brain*, 137(3):819–833.

[104] Tsai, C.-A. and Chen, J. J. (2009). Multivariate analysis of variance test for gene set analysis. *Bioinformatics*, 25(7):897–903.

[105] Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.*, 3(7):research0034.

[106] Vragović, K., Sela, A., Friedlander-Shani, L., Fridman, Y., Hacham, Y., Holland, N., Bartom, E., Mockler, T. C., and Savaldi-Goldstein, S. (2015). Translatome analyses capture of opposing tissue-specific brassinosteroid signals orchestrating root meristem differentiation. *Proc. Natl. Acad. Sci. USA*, 112(3):923–928.

[107] Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1):136–138.

[108] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63.

[109] Weigel, D. and Mott, R. (2009). The 1001 genomes project for arabidopsis thaliana. *Genome Biol.*, 10(5):107.

[110] Wu, D., Hu, Y., Tong, S., Williams, B. R., Smyth, G. K., and Gantier, M. P. (2013a). The use of miRNA microarrays for the analysis of cancer samples with global miRNA decrease. *RNA*, 19(7):876–888.

[111] Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M.-L., Visvader, J. E., and Smyth, G. K. (2010). ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182.

[112] Wu, D. and Smyth, G. K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.*, 40(17):e133–e133.

[113] Wu, H., Wang, C., and Wu, Z. (2013b). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, 14(2):232–243.

[114] Yaari, G., Bolen, C. R., Thakar, J., and Kleinstein, S. H. (2013). Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic Acids Res.*, page gkt660.

[115] Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., Wang, J., Li, S., Li, R., Bolund, L., et al. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.*, 34(suppl 2):W293–W297.

[116] Zhou, Y.-H., Barry, W. T., and Wright, F. A. (2013). Empirical pathway analysis, without permutation. *Biostatistics*, page kxt004.