Global (Collective) analysis of RNA-Seq experiment: multiple data sets & multiple genes

By
Bin Zhuo

A THESIS

submitted to

Oregon State University

University Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Statistics
(Honors Scholar)

Presented Month dd, yyyy
Commencement June 2016

AN ABSTRACT OF THE THESIS OF

Bin Zhuo for the degree of Honors Baccalaureate of Science in Statistics presented on Month dd, yyyy. Title:
Global (Collective) 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 (Collective) analysis of RNA-Seq experiment: multiple data sets & multiple genes

By
Bin Zhuo

A THESIS

submitted to

Oregon State University

University Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Statistics
(Honors Scholar)

Presented Month dd, yyyy
Commencement June 2016

Honors Baccalaureate of Science in Statistics project of Bin Zhuo presented on Month dd, yyyy

APPROVED:

Yanming Di, Mentor, representing Department of Statistics

Committee Member Name, Committee Member, representing Committee Member Department

---

Committee Member Name, Committee Member, representing Committee Member Department

---

Toni Doolen, Dean, Oregon State University Honors College

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

Bin Zhuo, Author

Contents

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 reflects 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/profiles of one or more genes from different sources. Factors for comparison can 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 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 profile or expression level 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 [12]

for a review). RNA-Seq, also known as *whole transcriptome shotgun sequencing* [57], 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 [21]: 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 tens of millions of reads (DESCRIBE IT); 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 profiles.

pros and cons about RNA-Seq

RNA-Seq technology offers several key advantages over other methods [79], 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 [31]. 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 processing them requires fast read mapping tools as well as a lot of computing resource [40, 44].

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 [39, 40], BWA [42, 43], Subread [70] and STAR [17]. In all situations, an index of either the reference genome or the reads are built at the beginning using hash tables or Burrows-Wheeler transform (BWT) [11]. 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 [30] or [45] 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 profiles. This can be done using HTSeq [2] or featureCounts [46], among other options. Finally, a read count matrix is obtained with each row representing a genomic feature unit and each column corresponding to a biological sample.

We assembled an in-house pipeline to process RNA-Seq data sets based on the R [62] 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 SRA (Sequence Read Archive) Toolkit [41] to convert SRA files to FASTQ files, and then Subread aligner [47] to map reads and featureCounts [46] to summarize counts (see Figure 1 for the work flow). We will use it to process multiple RNA-Seq data sets 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—whether a gene is expressed at different levels between two or more sample groups; and 3) gene set test—a type of downstream analysis in which a p-value is assigned to a set of genes as a unit.

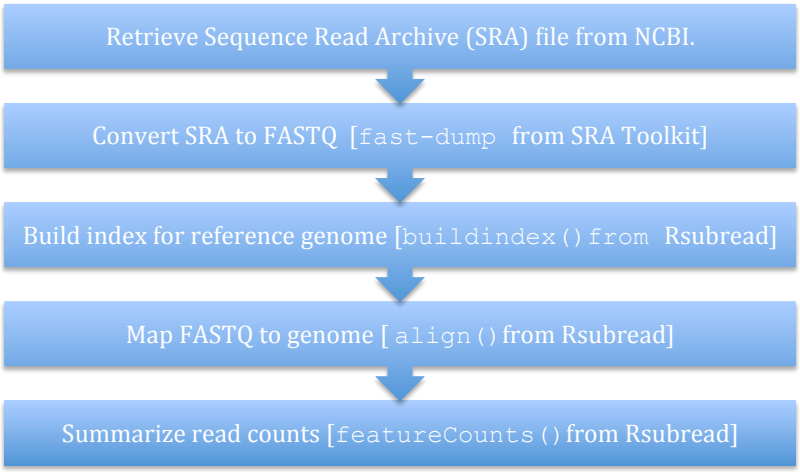

Figure 1: Work flow of data preprocessing: from raw reads sequencing data to read counts. Raw data in this workflow are retrieved from the NCBI. Data processing is based on two softwares SRA Toolkit [41] and Rsubread aligner [47]).

Normalization

Despite the optimistic claim that RNA-Seq does not need sophisticated normalization [79], many works have shown that normalization of count data is highly desirable before accessing differential expression to account for various sources of bias between samples [3, 16, 29, 66, 65, 68]. 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, such as the trimmed mean of M-values (TMM) [68], the DESeq normalization [3], and remove unwanted variation (RUV) [66].

DE analysis

Identification of DE genes is the key task in many biological studies. DE analysis uncovers the association between genes and responses/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 profile) or *down-regulated* (decreased expression profile) 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 [63, 69, 71]). In principle, most of those approaches are based on Poisson [51, 78] or Negative Binomial (NB) distribution [3, 15, 59, 67, 83] 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. A typical strategy is to perform gene set test—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. Gene set test enables researchers to examine an ensemble of genes simultaneously and thus improves interpretability of DE results. Depending on the definition of the null hypothesis, there are two types of gene set test: the *self-contained* test and the *competitive* test [24]. A self-contained test examines a set of genes by a fixed standard without reference to other genes in the genome (see, for example, [25, 26, 35, 77, 81]). A competitive test compares DE genes in the test set to those not in the test set [76, 80,

84]. The competitive gene set test is much more popular among genomic literatures [23, 24].

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 (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 [68] and DESeq [3] normalizations, assume that the majority of genes are not DE within the experiment under investigation. However, this assumption can be violated for some experiments, where over 50% of the genes' expression levels are affected by the treatments [50, 82]. 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 for count normalization. However, a number of works have shown that house-keeping genes are not necessarily stably expressed according to numerical stability measure (see, for example, [13, 36]). Another choice, the *spike-in genes*, is not reliable for normalization due to the same issue [66]. A popular approach has been to search from large sets of experiments for reference genes [13, 14, 22, 27, 72] which are evaluated by some numerical stability measure. Validation experiments (e.g. reverse transcription-PCR) show that reference genes identified by numerical methods generally outperform commonly used ones in terms of expression stability [13, 32]. 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 fo genes. Furthermore, making the reference set explicit will be beneficial during DE analysis, because often times biologists compare results from one experiment to ones that are publicly available.

1.1.3.2 Estimating correlations of test statistics

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 [34, 37, 56]). 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, for example, SigPathway [76], PAGE [38], MRSGE [55] or the 2×2 contingency-table-based tests [1, 33, 85]. However, it has been argued that such test procedures will result in inflated type I error [20, 24, 23, 80, 84], 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 [20, 23, 75]. 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 [75]. 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 experiment); second, it is computationally intensive; third, and most importantly, it implicitly alters the null hypothesis being tested and makes the null and alternative difficult to be characterized [24, 37, 80]. Another attempt has been to incorporate the inter-gene correlations into the formulation of gene set test procedure [80, 84]. CAMERA [80] estimates a *variance inflation factor* (VIF) from sample correlation (after the treatment effect removed), and then includes it in its gene set test statistic. The same VIF has also been used by QuSAGE [84] 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 profiles. 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 regression models are a natural choice for analyzing RNA-Seq data. In this section, we will first describe the formulation of generalized linear mixed models (GLMMs), and then review common methods for parameter estimation.

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 [53]. In the context of classical linear model, 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[1] component 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

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

[1] Unless specified otherwise, we assume there are n observations (i.e. $j = 1, \ldots, n$).

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 [64, 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. [48, 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) estimation 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. [48, Chapter 1].

1.2.1.3 Generalized linear models

We can take a different perspective of classical linear models by arranging Equation (1)–(3) into the following three parts [53, 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 the random component 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 result in the *generalized linear models* (GLMs), a framework that is especially suitable when the response can be no longer assumed to come from a normal distribution.

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 GLMM—Poisson log-linear mixed-effect model using data from RNA-sequencing experiments. Suppose we have RNA-Seq expression profiles (in the form of counts) randomly selected from three experiments, with two treatments nested in each experiment and two replicates for each treatment. We are not interested in the specific levels of treatment, and 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 observation unit from kth treatment of lth experiment. The link function $\eta_{jkl} = \log(\mu_{jkl})$ relates mean μ_{jkl} to linear predictors by Equation (12),

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

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

$\xi)$, $\boldsymbol{u} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\epsilon})$ 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 [58, 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 the random term $\boldsymbol{u}$ to 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 [49]. Gauss-Hermite quadrature is used for integrals of the form $\int_{-\infty}^{\infty} f(x)e^{-x^2}dx$, which 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 [49]. This procedure is called *Adaptive Gauss-Hermite* (AGH) quadrature [60].

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

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

[copied from SAS help] Let m be the number of quadrature points (i.e., the order of the Hermite polynomial) in each dimension for each random effect term. Let also Q be 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 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 computation for 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 [8].

1.2.2.2 Estimation based on linearization

maybe a brief introduction

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 (21) about estimates $\boldsymbol{\eta}$. Two approaches proposed by Breslow and Clayton [9]—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, given $\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{u}}$, of 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 linear mixed model (see Section 1.2.1.2), except that an estimate of $(\boldsymbol{\beta}, \boldsymbol{u})$ is needed for calculating 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 was done for 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 is performed at $\boldsymbol{\eta}_0 = \boldsymbol{X}\boldsymbol{\beta}_0$,

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

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

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

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

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

Next we 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 parameters $\boldsymbol{\beta}$ and variance component $\boldsymbol{\Sigma}$ is the same as that in PQL. 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 $\boldsymbol{u}$. MQL is also referred to as *population-averaged* estimate approach.

Breslow and Lin [10] and Pinheiro and Chao [61] showed 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

$$\log \pi(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{v}) = \log f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}) + \log \phi(\boldsymbol{u}|\boldsymbol{\Sigma}) \tag{34}$$

The optimal solution in Equation (34) can be obtained by *Expectation-Maximization* (EM) algorithm that can be readily implemented as follows:

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

$$E_{\boldsymbol{\beta}^{(k)}}[\log f(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{v})|\boldsymbol{y}] = Q_1(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)}),\ E_{\boldsymbol{\Sigma}^{(k)}}[\log \phi(\boldsymbol{\Sigma}|\boldsymbol{v})|\boldsymbol{y}] = Q_2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^{(k)}) \tag{35}$$

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. Unfortunately, the expectations in Equation (35) cannot be computed in closed form for GLMMs. However, they may be approximated by *Markov chain Monte Carlo* (MCMC). In light of this, McCulloch [54] 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

$$A_k(\boldsymbol{u}', \boldsymbol{u}) = \min\left\{1,\ \frac{f(\boldsymbol{u}'|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma})g(\boldsymbol{u})}{f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma})g(\boldsymbol{u}')}\right\} \tag{36}$$

If we choose $g(\boldsymbol{u}) = \phi(\boldsymbol{u}|\boldsymbol{\Sigma})$, the ratio term in Equation (36) can be simplified to

$$\begin{aligned}
&\frac{f(\boldsymbol{u}'|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\Sigma})g(\boldsymbol{u})}{f(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\Sigma})g(\boldsymbol{u}')} \\
&= \left[\frac{f(\boldsymbol{u}',\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\Sigma})}{f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\Sigma})}\phi(\boldsymbol{u}|\boldsymbol{\Sigma})\right] / \left[\frac{f(\boldsymbol{u},\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\Sigma})}{f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\Sigma})}\phi(\boldsymbol{u}'|\boldsymbol{\Sigma})\right] \\
&= \frac{f(\boldsymbol{y}|\boldsymbol{u}',\boldsymbol{\beta},\boldsymbol{\Sigma})\phi(\boldsymbol{u}'|\boldsymbol{\Sigma})\phi(\boldsymbol{u}|\boldsymbol{\Sigma})}{f(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{\beta},\boldsymbol{\Sigma})\phi(\boldsymbol{u}|\boldsymbol{\Sigma})\phi(\boldsymbol{u}'|\boldsymbol{\Sigma})} \\
&= \frac{f(\boldsymbol{y}|\boldsymbol{u}',\boldsymbol{\beta},\boldsymbol{\Sigma})}{f(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{\beta},\boldsymbol{\Sigma})}
\end{aligned} \tag{37}$$

The MCEM procedure combines the EM steps and Metropolis algorithm in estimating the fixed parameters and variance components 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. But 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) [54] and MCMC [28].

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 $\boldsymbol{\mu} = (\mu_{jkl})'$,

$$f(\boldsymbol{Y}|\boldsymbol{\mu}) = \prod_{j,k,l} f(y_{jkl}|\mu_{jkl}) = \prod_{j,k,l} \frac{[\mu_{jkl}]^{y_{jkl}} \exp(-\mu_{jkl})}{y_{jkl}!} \tag{38}$$

A re-expression of (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

$$f(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \prod_{j,k,l} \mu_{jkl}^{-1} \cdot \frac{1}{\sqrt{(2\pi)^{12}|\boldsymbol{\Sigma}|}} \exp[-\frac{1}{2}(\log \boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\log \boldsymbol{\mu} - \boldsymbol{\mu}_0)] \tag{39}$$

Since $Y_{jkl} \sim \text{Poisson}(\mu_{jkl})$, by combining Equation (38) and (39), we obtain the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{\mu}$,

$$f(\boldsymbol{Y}, \boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^{12}|\boldsymbol{\Sigma}|}} \exp[-\mathbf{1}^T \boldsymbol{\mu} - -\frac{1}{2}(\log \boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\log \boldsymbol{\mu} - \boldsymbol{\mu}_0)] \prod_{jkl} \frac{[\mu_{jkl}]^{y_{jkl}-1}}{y_{jkl}!}$$

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

$$L(\xi, \sigma_1^2, \sigma_2^2, \sigma_0^2|\boldsymbol{Y}) = f(\boldsymbol{Y}|\boldsymbol{\xi}, \boldsymbol{\Sigma}) = \int_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{\epsilon}} f(\boldsymbol{Y}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\epsilon}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) d\boldsymbol{a} d\boldsymbol{b} d\boldsymbol{\epsilon} \tag{40}$$

The integral in Equation (40) can be approximated by adaptive Gaussian-Hermite (AGH) quadrature or MCMC. For AGH quadrature, we first approximate the likelihood by Equation (20) and then estimate $\boldsymbol{\theta} = (\xi, \sigma_0^2, \sigma_1^2, \sigma_2^2)'$ maximizing the resulting

likelihood. R package `lme4` [5] has an inbuilt function `glmer()` for this procedure. The MCMC has been implemented in several packages, for example, `Rstan` [73] or `MCMCPack` [52].

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. [6] 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 [6], is the expected proportion of false positives among all significant calls (null rejected). FDR has been studied extensively ([7, 18, 19, 74] and more) over the past two decades. FDR is equivalent to FWER [6] 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 respectively. 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

3. Chapter 2

chapter 2

4. Chapter 3

Chapter 3.

5. Conclusion

References

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

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

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

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

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

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

[7] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

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

[9] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

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

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

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

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

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

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

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

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

[18] B. Efron. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99(465), 2004.

[19] B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2010.

[20] B. Efron and R. Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, pages 107–129, 2007.

[21] F. Finotello and B. Di Camillo. Measuring differential gene expression with rna-seq: challenges and strategies for data analysis. *Briefings in functional genomics*, 14(2):130–142, 2015.

[22] M. Frericks and C. Esser. 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, 2008.

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

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

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

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

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

[28] J. D. Hadfield et al. Mcmc methods for multi-response generalized linear mixed models: the mcmcglmm r package. *Journal of Statistical Software*, 33(2):1–22, 2010.

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

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

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

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

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

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

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

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

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

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

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

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

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

[42] H. Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

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

[44] H. Li and R. Durbin. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

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

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

[47] Y. Liao, G. K. Smyth, and W. Shi. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research*, 41(10):e108–e108, 2013.

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

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

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

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

[52] A. D. Martin, K. M. Quinn, and J. H. Park. Mcmcpack: Markov chain monte carlo in r. 2011.

[53] P. McCullagh and J. A. Nelder. Generalized linear models. 1989.

[54] C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170, 1997.

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

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

[57] R. D. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. J. Pugh, H. McDonald, R. Varhol, S. J. Jones, and M. A. Marra. Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing. *Biotechniques*, 45(1):81, 2008.

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

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

[60] J. C. Pinheiro and D. M. Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, 4(1):12–35, 1995.

[61] J. C. Pinheiro and E. C. Chao. Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15(1), 2006.

[62] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

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

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

[65] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit. Gc-content normalization for rna-seq data. *BMC bioinformatics*, 12(1):480, 2011.

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

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

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

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

[70] W. Shi and Y. Liao. Subread/rsubread users guide, 2013.

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

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

[73] Stan Development Team. *RStan: the R interface to Stan, Version 2.9.0*, 2016. URL `http://mc-stan.org`.

[74] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[75] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[76] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13544–13549, 2005.

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

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

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

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

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

[82] D. Wu, Y. Hu, S. Tong, B. R. Williams, G. K. Smyth, and M. P. Gantier. The

use of miRNA microarrays for the analysis of cancer samples with global miRNA decrease. *RNA*, 19(7):876–888, 2013.

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

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

[85] J. Ye, L. Fang, H. Zheng, Y. Zhang, J. Chen, Z. Zhang, J. Wang, S. Li, R. Li, L. Bolund, et al. Wego: a web tool for plotting go annotations. *Nucleic acids research*, 34(suppl 2):W293–W297, 2006.