ABSTRACT

Title of dissertation: SHAPE ANALYSIS OF HIGH-THROUGHPUT GENOMICS DATA

Kwame Baabu Okrah, Doctor of Philosophy, 2015

Dissertation directed by: Dr. Héctor Corrada Bravo
Department of Computer Science

RNA sequencing refers to the use of next-generation sequencing technologies to characterize the identity and abundance of target RNA species in a biological sample of interest. The recent improvement and reduction in the cost of next-generation sequencing technologies have been paralleled by the development of statistical methodologies to analyze the data they produce. Coupled with the reduction in cost is the increase in the complexity of experiments. Some of the old challenges still remain. For example the issue of normalization is important now more than ever. Some of the crude assumptions made in the early stages of RNA sequencing data analysis were necessary since the technology was new and untested, the number of replicates were small, and the experiments were relatively simple.

One of the many uses of RNA sequencing experiments is the identification of genes whose abundance levels are significantly different across various biological conditions of interest. Several methods have been developed to answer this question. Some of these newly developed methods are based on the assumption that the data observed or a transformation of the data are relatively symmetric with light tails,

usually summarized by assuming a Gaussian random component. It is indeed very difficult to assess this assumption for small sample sizes (e.g. sample sizes in the range of 4 to 30).

In this dissertation, we utilize L-moments statistics as the basis for normalization, exploratory data analysis, the assessment of distributional assumptions, and the hypothesis testing of high-throughput transcriptomic data. In particular, we introduce a new normalization method for high-throughput transcriptomic data that is a modification of quantile normalization. We use L-moments ratios for assessing the shape (skewness and kurtosis statistics) of high-throughput transcriptome data. Based on these statistics, we propose a test for assessing whether the shapes of the observed samples differ across biological conditions. We also illustrate the utility of this framework to characterize the robustness of distributional assumptions made by statistical methods for differential expression. We apply it to RNA-seq data and find that methods based on the simple t-test for differential expression analysis using L-moments statistics as weights are robust. Finally we provide an algorithm based on L-moments ratios for identifying genes with distributions that are markedly different from the majority in the data.

SHAPE ANALYSIS OF HIGH-THROUGHPUT GENOMICS DATA

by

Kwame Baabu Okrah

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Dr. Héctor Corrada Bravo, Chair/Advisor
Dr. Najib El-Sayed
Dr. Stephen Mount
Dr. Eric Slud
Dr. Sridhar Hannenhalli

Dedication

In loving memory of Joseph Lewis Okrah

Acknowledgments

I would like to thank my advisor Dr. Héctor Corrada Bravo and the other members of my dissertation committee: Dr. Stephen Mount, Dr. Najib El-Sayed, Dr. Sridhar Hannenhalli, and Dr. Eric Slud. I found Dr. Héctor Corrada Bravo's style of mentorship to be unique, challenging, and rewarding. His breadth of knowledge in diverse areas of science is impressive. I would gladly recommend him to any student. I would like to acknowledge Dr. Stephen Mount for his open door policy and for the many discussions we have had over basic biology and chemistry. I would like to acknowledge Dr. Najib El-Sayed and members of his lab for presenting me with challenging datasets, some of which are the motivation for some of the work in this dissertation. I want to commend Dr. Sridhar Hannenhalli for organizing the Center for Bioinformatics and Computational Biology (CBCB) research in progress talks. I have had the opportunity to present some of the work in this dissertation at these talks. I found the feedback from the audience to be very constructive and useful. I would like to acknowledge Dr. Eric Slud for raising the bar in the mathematical statistics courses that I took with him. Much of what I learned in that class was applied to this dissertation.

I would like to acknowledge Stephanie Hicks and Rafael Irizarry, both of whom I collaborated with on the work of Smooth Quantile Normalization (Chapter 5). I would like to thank my colleagues at CBCB. Especially the members of the Bravo Lab. I would like to thank Hisham Talukder for being a great friend and lab mate throughout my time at Maryland. I would like to thank Eddie Kim, Joyce Hsiao, and

Joseph Paulson for their advice and encouragement during tough times. I would like to thank Keith Hughitt for organizing the Bring Your Own Bioinformatics talks, an avenue where graduate students and faculty can openly interact and share ideas. I would like to thank Ashton Trey Belew for our insightful discussion on quantitative polymerase chain reaction. I would like to acknowledge Christopher Hill for our many discussions on science and mathematics education policy. I cannot forget to mention Mr. Senthil Muthiah for the many lively discussions we have had about some of the work in this thesis, astrology, marriage, family, and every aspect of life.

I would like to thank my mother Theresa Kyeremateng, my brothers Kwame Agyapong Okrah and Kwabena Kwarteng Okrah, and my dear sister Abena Pokua Okrah. My family provided me with support and love throughout my years in graduate school. I would like to especially thank my uncle Mr. Kwame Asafo-Adjei and his wife Victoria Asafo-Adjei. This dissertation is only possible because of their love and support.

I would like to acknowledge Alverda McCoy, Celeste Regalado, and Denise Cross. All of whom have made this dissertation possible. Finally, I would like to thank the University of Maryland, College Park and the tax payers of the state of Maryland for making the campus a conducive and exciting environment for higher learning.

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Introduction

The central dogma of molecular biology describes the metabolic pathway in which chains of amino acids (*polypeptides*), the precursors of proteins, are synthesized. First, certain sections of the genetic material DNA (*deoxyribonucleic acid*) in the nucleus of the eukaryotic cell are transcribed into mRNA (*messenger ribonucleic acid*). As shown in Figure 1.1 the mRNA is then transported into the cytosol and translated into polypeptides. The polypeptides are then translocated into certain areas of the cytosol (or outside the cell) where they may bond with other polypeptides to form proteins [1]. A *gene* is a section of DNA or RNA that encodes the information necessary for synthesizing a protein or a functional RNA molecule. Usually these are proteins and RNA. The entire set of DNA molecules found in an organism is known as its *genome* [1]. One of the main functions of the genome is to specify, regulate, and drive the creation of proteins that will provide structure to the cell and control the metabolism of the cell [1]. Although the genome of most cells found in the human body is the same, the genes that are transcribed and translated into proteins (*gene expression*) are not. The type, the time, the location, and the rate of gene expression are different in different type of cells. Or cells under different kinds of treatments; for example diseased cells versus diseased cells that have been treated with medicine. The entire set of RNA molecules found in a cell is called the *transcriptome*. A typical cell contains approximately 20 to 30 pg of RNA, which is about 1% of the cell's total mass; while mRNA forms only about 4% of the total RNA in a cell [2]. See Figure 1.2. Due to the relatively simple chemical structure of DNA and RNA as compared with proteins; they have been the focus of some of the biomolecular technologies, in particular techniques that can survey most or all the RNA species found in a biological sample have been developed [3, 4].

The development of genomic high-throughput assays that seek to detect and quantify the abundance of target molecules such as mRNA has influenced the manner in which basic biological and clinical research is conducted. Specifically, biologists are now able to gain further insight into biological systems from a holistic point of view [5–7]. The impact of genomic high-throughput technologies can also be felt in medicine and drug development [8]. As these technologies have matured; experiments have grown more complex. Data produced by genomic high-throughput technologies can be used to verify biological hypotheses, develop biomarkers [9, 10], or can be used in an exploratory manner. However before any analysis can be performed, the data has to be normalized to account for potential systematic technical biases [11–14]. After normalization, one of the most common statistical tests used in the analysis of high-throughput experiments data is to find genes that are differentially expressed across various biological conditions of interest [15–17].

In this chapter we begin by reviewing the basic structure of nucleic acids (DNA and RNA) necessary for the understanding of this dissertation. Next we discuss RNA sequencing (RNA-seq). We then summarize the current state of the art statistical methods for normalization and differential expression. We end the chapter by providing an outline of this dissertation.

1.2 Basics of DNA and RNA Molecular Structure

Nucleic acids, which include DNA and RNA, play an important role in all known forms of life [1]. Nucleic acids are large and complex molecules whose structure can be categorized into four levels: (1) nucleotides (2) single strand (3) double helix, and (4) three-dimensional protein-nucleic acid complex.

Nucleotides are the fundamental units of DNA and RNA. Every nucleotide consists of three components: one or more phosphate groups, a pentose sugar, and a nitrogenous base. The pentose sugar contains five carbon atoms which are labelled as $1'$ through to $5'$. There are two main types of sugar: *ribose*, which is found in RNA only, and *deoxyribose*, which is found in DNA only. The ribose sugar has a hydrogen (H) atom and a hydroxyl (HO) group covalently bonded to the $2'$-carbon, whereas the deoxyribose sugar has two hydrogen atoms bonded to the $2'$-carbon. The loss of an oxygen atom at the $2'$-carbon of deoxyribose makes DNA more chemically stable compared with RNA. This stability of DNA plays a key role in many high-throughput assays. There are five different kinds of nitrogenous base: adenine (A), guanine (G), and cytosine (C) which can be found in both DNA and RNA; thymine (T) which is found in DNA only; and uracil (U) which is found in RNA only. In both DNA and RNA the nitrogenous base is covalently bonded to the sugar at the

$1'$-carbon. Nucleotides can have one or more phosphate groups covalently bonded to the $5'$-carbon of the pentose sugar. The name given to a nucleotide is based on the type of base, the type of sugar, and the number of phosphate groups attached to the sugar. For example a nucleotide with an adenine base, ribose sugar, and one phosphate group is called *adenosine monophosphate* (AMP). If the nucleotide has a deoxyribose sugar instead of a ribose sugar then it is called *deoxyadenosine monophosphate* (dAMP). In this dissertation, as is commonly done, we will denote a nucleotide by the initial letter of its nitrogenous base. Details on the type of nucleotide will be given in this dissertation if the context requires it.

The phosphate group (attached at the $5'$-carbon of the sugar) of a nucleotide can covalently bond only at the $3'$-carbon of the sugar of another nucleotide. A sequence of such sugar-phosphate bonds creates a single linear nucleic acid strand. For example, suppose that we are given the following four nucleotides $5'$-A-$3'$, $5'$-G-$3'$, $5'$-C-$3'$, and $5'$-U-$3'$ (here we have emphasized the $5'$ and $3'$ carbons of the pentose sugar for illustration). A possible single strand is $5'$-G-$3' - 5'$-U-$3' - 5'$-A-$3' - 5'$-C-$3'$; which is usually denoted as $5'$-GUAC-$3'$ or simply GUAC. The key point to note here is that the strand has a natural direction ($5' \rightarrow 3'$).

Two DNA strands (and sometimes RNA strands) can interact with each other to form a double strand (*double helix*), twisted together around a common axis. This double helix is made stable by the hydrogen bonds made between the bases on the opposite strands (*base pairs* (bp)). A key feature of the base pairs is that they are specific. An adenine (A) pairs with a thymine (T) in DNA, and with a uracil (U) in RNA; a guanine (G) pairs with a cytosine (C) in both DNA and RNA. It is this base pair specificity that forms the basis of most high-throughput genomics technology, such as RNA-seq and microarrays [3, 4]. Three hydrogen bonds occur between G and C but only two between A and T. This makes a DNA double strand with a high proportion of G and C (GC content) to be more stable when compared to one with a low GC content. Besides the difference in sugars and T replaced with U; the structure of an RNA strand is similar to a DNA strand. During transcription DNA is used as a template to create a single-stranded RNA molecule. Complementary sequences on the same RNA strand (or between two separate RNA strands) can bond at certain locations to form a folded single strand with double stranded regions. RNA coupled with proteins can form a functional RNA product such as transfer RNA (tRNA) and ribosome RNA (rRNA). RNA strands that become mRNA are modified at the $5'$ and $3'$ ends. A cap (a specially altered nucleotide) is placed at the $5'$ end. The $3'$ end is modified by attaching a sequence AMPs, known as the poly(A) tail, to it.

To fit within the nucleus of a eukaryotic cell, chromosomal DNA, a very long double stranded DNA (typically millions of bp long) must be efficiently packaged into a three-dimensional conformation [1]. This is possible with the aid of DNA-binding proteins, such as histone proteins.

1.3 Data Generation Process

In this section we will briefly describe the process by which gene expression measurements are obtained using RNA-seq. We will end the section by describing the

External RNA Control Consortium spike-in mix (ERCC spike-in mix), a synthetic set of 92 mRNA molecules with known concentrations that is designed to approximate mRNA species found in eukaryotic organisms. The ERCC spike-in mix can be used as a negative or positive control in high-throughput experiments. Please see Figure 1.3.

1.3.1 RNA Sequencing

RNA sequencing (RNA-seq) refers to the use of next-generation sequencing technologies to characterize the transcriptome of a biological sample [4]. DNA sequencing is a biomolecular assay for determining the exact sequence of a given DNA molecule. A typical first generation high-throughput sequencing platform (also known as Sanger sequencing [18]) using 96 capillaries to sequence single stranded DNA molecules of length 600-1000 nucleotides can generate a dataset containing about 1×10^5 nucleotides. On the other hand a standard run of a second generation high-throughput experiment; sequencing single stranded DNA molecules of length 100 nucleotides can yield a dataset containing approximately 6×10^{11} nucleotides [19]; orders of magnitude higher than the first generation high-throughput genomic technologies. It is this improvement in throughput of second generation sequencing technologies that make them ideal for identifying and quantifying each RNA molecule in a biological sample. The typical RNA-seq experimental procedure involves isolation of RNA from the sample of interest, construction of a cDNA library that reflects the starting RNA population, and the sequencing of the library.

Using commercially available kits RNA is extracted from a tissue or cells of interest. Standard cDNA library preparation protocols require about 0.1 - 10 μg of starting total RNA depending on the application and sequencing platform [19]. After extraction and quality assessment, the total RNA can be enriched for mRNA by using *polydeoxythymidine* (poly-dT or oligo-dT) magnetic beads to bind to the poly(A) tail of mRNA or by using reagents that can digest rRNA (which accounts for about 95% of the total RNA [2]). Please see Figure 1.2. The mRNA sample is then purified to remove any remaining rRNA or contamination from DNA. The purified mRNA is randomly fragmented and primed with random hexamers (single stranded DNA of length 6 nucleotides; the sequence is random), that have special tags attached to them. The fragmented mRNAs are reverse transcribed using the enzyme reverse transcriptase to obtain a single stranded complementary DNA (cDNA). A second primer based on the tag attached to the first primer is added in order to synthesize the second strand of DNA to obtain a double stranded cDNA (ds cDNA). After purification and size selection (size depends on platform and experimental goals), adapters (special platform dependent sequences) are ligated to the ends of the ds cDNA. At this point a few rounds of *polymerase chain reaction* (PCR), typically 12-16 rounds [19] are used to amplify the cDNA library (this step is optional). The adapters can be indexed for a given cDNA library. In this way samples can be pooled and sequenced together. In some cases before sequencing the cDNA library, the samples are normalized to have an equal amount of cDNA in each library. This step is not performed by all labs as it can have an adverse effect on the expression levels of the transcripts [20]. Also see the discussion in section 1.4 on normalization. The library preparation steps described above are generic and may be done in a slightly different manner depending on the goal of the experiment.

At this point the cDNA library is now ready to be sequenced. There are many sequencing platforms currently available. Some of the common ones are the Illumina platform and the Roche 454 platform. These platforms differ in their chemistry and the length of sequences that they produce. However they both output text files that contain a unique name for each cDNA molecule in the library and its sequence. These sequences are aligned to the known genome sequence of the target organism [21, 22]. The aligned reads are summarized by assigning to a gene the number of reads (sequences) aligned within its boundaries (or near its boundaries, since gene annotation is not perfect) [23, 24]. Alternatively, there are other methods that can quantify mRNA expression levels without aligning to a reference genome [25]. The counts assigned to each gene in each sample is typically stored in the form of a matrix which we will denote as $\mathbf{X}$ in this dissertation.

1.3.2 Notation

Throughout this dissertation we will let x_{gi} represent the total number of reads aligned to gene g in sample i (unless we specify otherwise). When not specified we will assume that $g \in \{1, 2, \cdots, G\}$ and $i \in \{1, 2, \cdots, n\}$; where G is the number of genes and n is the number of samples. We will denote all the gene counts within a sample as the column vector $\mathbf{X}_i = [x_{1i},\ x_{2i},\ \cdots,\ x_{Gi}]^T$. As mentioned above the entire count matrix will be denoted as $\mathbf{X} = [\mathbf{X}_1,\ \mathbf{X}_2,\ \cdots,\ \mathbf{X}_n]$. Order statistics will play a significant role in this thesis so we will introduce our notation for order statistics here and remind the audience about their meaning again when the need arises. Let

$$\mathbf{Q}_i = [x_{(1)i},\ x_{(2)i},\ ,\cdots,x_{(G)i}]^T \tag{1.1}$$

where $x_{(1)i} \leq x_{(2)i} \leq \cdots \leq x_{(G)i}$. That is $\mathbf{Q}_i$ is the sorted counts vector for sample i.

1.3.3 External RNA Control Consortium Spike-in Mixes

The External RNA Control Consortium (ERCC) is a collaborative group of academic, private, and public organizations hosted at the National Institutes of Standard and Technology (NIST) [26, 27]. The ERCC has developed a set of 92 mRNA controls (20-mer poly(A) tails) that can be used in gene expression platforms such as RNA-seq, DNA microarrays, and quantitative real-time reverse transcriptase PCR (qRT-PCR). See Figure 1.3. The unique sequence of each of the mRNAs are largely random. They have been compared to multiple databases including human, fruit fly, mouse, bacteria, mosquito, and other species [28]. The 92 mRNA transcripts are divided into 4 groups labelled A, B, C, and D. Each group contains 23 mRNA transcripts spanning a 10^6-fold concentration range (see Figure 1.5). The mRNA lengths and nucleotide composition are similar across the 4 groups (see Figure 1.4). There are two ERCC control spike-in mixes: mix 1 and mix 2. The molar concentration ratios of mix 1 to mix 2 are 4, 1, 0.67, and 0.5 for group A, B, C, and D respectively (see Figure 1.3). When the ERCC spike-in mix is used as a control in the experiment its measurements can be used as part of the data normalization process [20,29].

1.4 Normalization of High-throughout Genomics Data

Normalization of the count matrix is a critical step that is taken prior to any statistical analysis. Normalization is defined as the removal of systematic experimental bias and technical variation with the goal of improving the identification of gene expression changes across biological conditions [30]. One obvious factor that needs to be accounted for when comparing the expression of two genes is their lengths. A longer gene will tend to have a higher read count than an equally expressed shorter gene. In comparing the same gene across different treatments, as is done in this dissertation, normalizing for gene length is not necessary since the genes being compared have equal lengths. However, normalization to remove the effect of sequencing depth is required. In this section we will summarize some of the current techniques used to normalize RNA-seq data. We begin by discussing some of the assumptions that are implicitly used behind normalization procedures. Next we discuss normalization methods that scale every gene in each sample with a single sample specific scalar. Numerous algorithms are available in the literature [11,12,31] for finding the sample specific scalars. Finally we discuss quantile normalization [14], a technique that was initially developed to normalize DNA microarray data.

1.4.1 Assumptions on Global Gene Expression

Most normalization strategies that are based solely on the counts matrix without any external information make the assumption that: for each cell or tissue under study only a few genes change expression levels or that an equivalent number of genes increase and decrease across the different biological conditions [30]. In this dissertation we will refer to this assumption as the *global transcriptome similarity.* This assumption can be interpreted in different ways leading to different normalization procedures. For example, the mean expression level across genes within each sample should be the same across biological conditions [12]. Or that on average the distribution of gene expression within each sample should be same across biological conditions [14]. While these assumptions may be reasonable in certain experiments, they may not always hold. For example, mRNA content has been shown to fluctuate significantly during zebrafish early developmental stages [30]. It has also been shown that cells with high levels of c-Myc can amplify their global gene expression two to three times more than their low c-Myc counterparts [20]. Other normalization methods are based on *housekeeping genes.* These are genes that are believed to be play a critical role in basic cellular pathways and as such should be expressed all the time at an equal rate independent of biological conditions [32].

1.4.2 Scaling Normalization Methods

In this section we describe three normalization methods that scale each gene count in a sample by a single sample specific constant. We will collectively call these terms scaling methods in this dissertation when we wish to discuss their generic features. In this dissertation we will discuss: total library size scaling (or mean scaling or counts per million (cpm)), median scaling, trimmed mean scaling, and the Anderson and Huber (AH) [15] scaling method. The mean scaling method simply divides each gene count in a given sample by the average gene count in that sample. Similarly the trimmed mean and the median scaling methods divide each gene count for a given sample respectively by the trimmed mean and the median of that sample. All three of these methods make the assumption that the location parameter (e.g. mean, trimmed mean, median) of each sample should be the same across biological conditions. The techniques only differ by the choice of location parameter (i.e. mean or median) and/or the estimation procedure (i.e. mean or trimmed mean).

The AH method assumes that majority of the genes are not differentially expressed. So if a ratio of two samples were taken then we should expect the distribution of the these ratios to be centered around a dominant scale, in particular the median. The method is generalized to more than two samples by first computing a reference sample (the geometric mean of each gene across samples). This reference sample is then compared to each individual sample in the dataset:

$$s_i = \text{median}_{g \in (1,2,\ldots,G)} \{x_{gi}/x_g^{ref}\} \tag{1.2}$$

where $x_g^{ref} = (\prod_{i=1}^{n} x_{gi})^{1/n}$ (i.e. geometric mean).

1.4.3 Quantile normalization

Quantile normalization assumes that the distribution of gene expression levels within a cell or tissue should be approximately the same [14]. Based on this assumption it modifies the observed distribution of counts within each sample to be the same for the entire experiment. Although it was originally designed for DNA microarray data the idea behind it is general. It has been applied to different types of high-throughput genomics data including RNA-seq data [33], DNA methylation data [34] and high-throughput qRT-PCR data [35]. The algorithm is as follows: (1) sort each sample vector in the counts matrix $\mathbf{X}$ to get $\mathbf{Q} = [\mathbf{Q}_1,\ \mathbf{Q}_2,\ \cdots,\ \mathbf{Q}]$, where $\mathbf{Q}_i = [x_{(1)i},\ x_{(2)i},\ ,\cdots, x_{(G)i}]^T$. (2) Compute the average counts across rows of $\mathbf{Q}$ to get the reference quantile:

$$\bar{\mathbf{Q}}_{\cdot\cdot} = (1/n)\sum_{i=1}^{n} \mathbf{Q}_i \tag{1.3}$$

where n is the number of samples. (3) Replace each column of $\mathbf{Q}$ with $\bar{\mathbf{Q}}_{\cdot\cdot}$ to obtain $\mathbf{Q} = [\bar{\mathbf{Q}}_{\cdot\cdot},\ \bar{\mathbf{Q}}_{\cdot\cdot},\ \cdots,\ \bar{\mathbf{Q}}_{\cdot\cdot}]$. (4) Get the normalized counts by re-ordering $\mathbf{Q}$ according to the original order of $\mathbf{X}$.

1.5 Differential Expression Analysis

The identification of genes that are expressed in different quantities under different biological conditions is one of the main uses of RNA-seq data [15,31,36]. Due to the interdependent nature of metabolic pathways in living organisms one can expect that the expression of genes in a sample of interest will be correlated. However differential expression analysis is typically performed for one gene at a time with mechanisms that allows information to be borrowed across genes. There are currently many statistical methods available and they can be categorized into 2 groups: those that work with counts; and those that work with a log transformation of counts. In this dissertation we will discuss two of the most popular methods in both camps: (1) DESeq2 [31] in the counts camp and, (2) voom-limma [36] in the log counts camp. For the remainder of this dissertation, we will assume that the observed counts for a given gene under a given biological conditions are independent and identically distributed. We will also assume that genes are independent.

1.5.1 DESeq2

The DESeq2 method [31] is based on a hierarchical generalized linear model. It assumes that counts of a gene across replicates follow a negative binomial (NB) distribution:

$$x_{gi}|\beta_g, \alpha_g \sim \text{NB}(\text{mean} = \mu_{gi}, \text{dispersion} = \alpha_g) \tag{1.4}$$

where $\mu_{gi} = s_i q_{gi}$. A logarithmic link is assumed: $\log(q_{gi}) = x_i^T \beta_g$, where x_i^T is an indicator of which group sample i belongs to, and β_g is the vector of group specific effects. On top of this likelihood model Love et al. [31] assume that:

$$\log(\alpha_g) \sim \text{N}(\log\alpha_{tr}(\bar{\mu}_g), \sigma_d^2), \quad \bar{\mu}_g = (1/n)\sum_i^n \frac{x_{gi}}{s_i} \tag{1.5}$$

where $\alpha_{tr}(\bar{\mu}) = \frac{a_1}{\bar{\mu}} + \alpha_0$ describes the mean-dispersion trend. And $\beta_g \sim \text{N}(\mathbf{0}, \sigma_r \mathbf{I})$. The hyper-parameters $\alpha_0, a_1, \sigma_d^2, \sigma_r^2$ and $\alpha_{tr}(\cdot)$ are estimated from the data and combined with the gene-specific (i.e. using data from the given gene only) maximum likelihood estimates of β_g and α_g to obtain the final moderated versions (of β_g and α_g) for use in inference. See Love et al. [31] for details.

1.5.2 Voom-limma

The limma [16] package designed for DNA microarrays has been one of the most successful use of the empirical Bayes estimation procedure [37]. At a time when experimental replicates were typically small (2-3), the limma package was able to borrow information across all genes when estimating the gene-specific variance. This was done by assuming that the gene-specific variances were generated from a prior distribution with a single hyper-parameter:

$$\sigma_1^2, \sigma_2^2, \ldots, \sigma_G^2 \sim \text{prior}(\sigma_0^2) \tag{1.6}$$

where $\sigma_g^2,\ g \in (1, 2, \ldots, G)$ is the gene-specific variance and σ_0^2 is the hyper-parameter to be estimated using information in the entire dataset. Using this framework limma estimates the gene-specific variance with the mean of the posterior distribution. The result (based on the appropriate selection of the distribution models for the prior and likelihood) turns out to be a weighted average of the gene-specific variance estimate (using the only the gene's data) $\hat{\sigma}_g^2$ and the estimate using the entire dataset $\hat{\sigma}_0^2$:

$$\text{limma estimate: } \tilde{\sigma}_g^2 = w\hat{\sigma}_g^2 + (1-w)\hat{\sigma}_0^2, \quad w \in (0,1) \tag{1.7}$$

where the weight w tends to 1 as the number of replicates increases. That is, the gene-specific estimate based on gene-specific information only, is shrunk towards the experiment-wise variance. The log of the intensities (log-intensities) of DNA microarrays is typically approximated by a Gaussian model [16]. As was observed in many datasets the mean and variance of log-intensities do not exhibit any trend [16]. Hence shrinking the gene-specific variance estimate towards a common experiment-wise estimate is reasonable.

The voom-limma pipeline [36] accepts as input log counts per million (log-cpm). Unlike log-intensities log-cpm values have been observed to exhibit a mean-variance trend; low log-cpm values tend to have a higher variance as compared to high log-cpm values [15, 31, 36]. The voom-limma package estimates the mean-variance trend from the data and shrinks the gene-specific variance towards the trend. This is achieved by weighting each log-cpm expression value by the inverse value (voom-weights) of the mean-variance trend evaluated at that value. This is, if the log-cpm value is low it will be penalized more than if it were high. The differential expression analysis then proceeds as a weighted Gaussian linear model. The limma package also allows the user to enter his or her own weights.

1.6 Dissertation Outline

This dissertation is outlined as follows. In chapter 2, I introduce alternative frameworks for describing distributional shape (skewness and kurtosis) of data: (1) Tukey's g-and-h distribution [38] and (2) L-moments [39]. Both of these frameworks enjoy theoretical and practical advantages over traditional moments [39–42]. I will primarily use Tukey's g-and-h distribution as a means of generating data that deviate from the Gaussian distribution (in terms of skewness and kurtosis) in a smooth manner. L-moments are linear combinations of order statistics that summarize the location, spread, skewness and kurtosis of a distribution. L-moments statistics will form the basis of methods that I propose in this dissertation. In chapter 3, I illustrate how to use L-moments ratios to characterize the distribution of RNA-seq datasets (SO-plot). Based on these characterizations I approximate how well Gaussian linear models will perform on log RNA-seq data. In chapter 4, I propose a test for assessing the global transcriptome similarity assumption. I illustrate the utility of the method on RNA-seq counts and log RNA-seq counts. In chapter 5, I introduce a new normalization method that is a modification of quantile normalization. In chapter 6, I demonstrate the software that I have developed to implement some of the ideas in this dissertation. Finally, I conclude this dissertation with a summary and discussion in chapter 7.

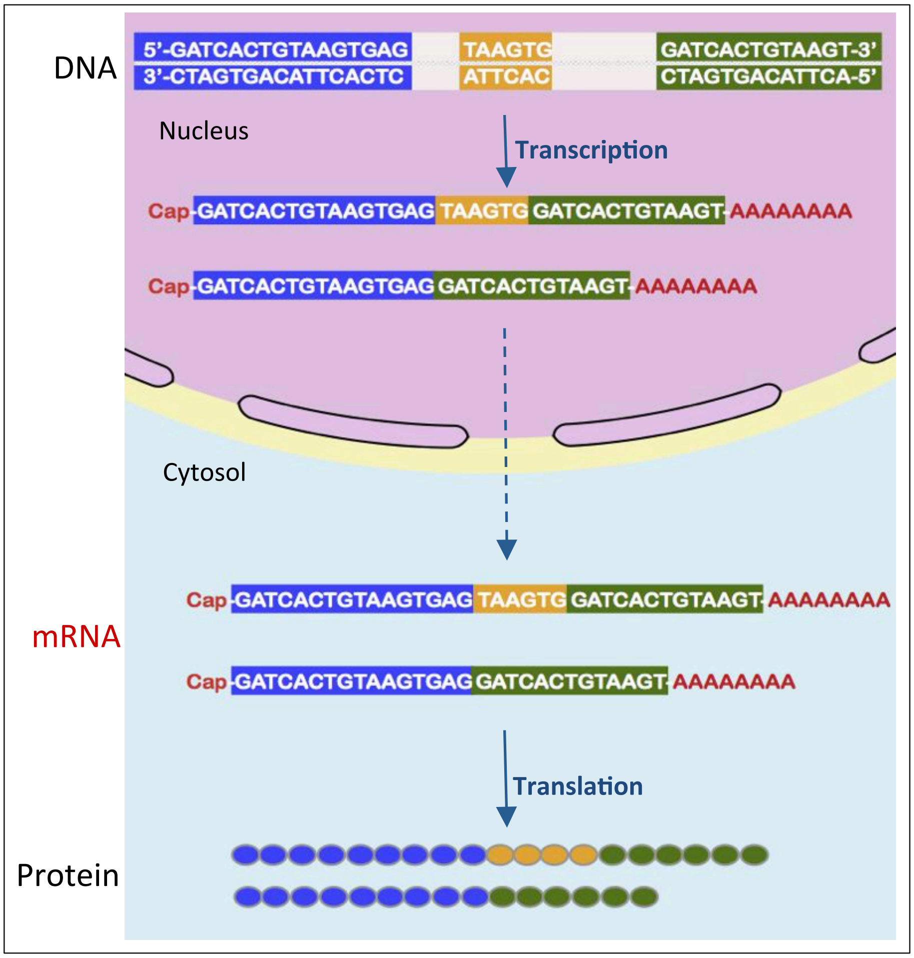

DNA
5'-GATCACTGTAAGTGAG
3'-CTAGTGACATTCACTC
TAAGTG
ATTCAC
GATCACTGTAAGT-3'
CTAGTGACATTCA-5'
Nucleus
Transcription
Cap-GATCACTGTAAGTGAG TAAGTG GATCACTGTAAGT-AAAAAAAA
Cap-GATCACTGTAAGTGAG GATCACTGTAAGT-AAAAAAAA
Cytosol
mRNA
Cap-GATCACTGTAAGTGAG TAAGTG GATCACTGTAAGT-AAAAAAAA
Cap-GATCACTGTAAGTGAG GATCACTGTAAGT-AAAAAAAA
Translation
Protein

Figure 1.1: **The central dogma of molecular biology.** The central dogma of molecular biology describes the flow of information from DNA to mRNA to Proteins. DNA is transcribed into mRNA in the nucleus of the eukaryotic cell. The mRNA is modified with a poly(A) tail at the 3′ end and a cap at the 5′ end. The poly(A) tail in mRNA makes it possible for it to be separated from total RNA. As the figure shows a single gene can generate different types of mRNA transcripts known as isoforms.

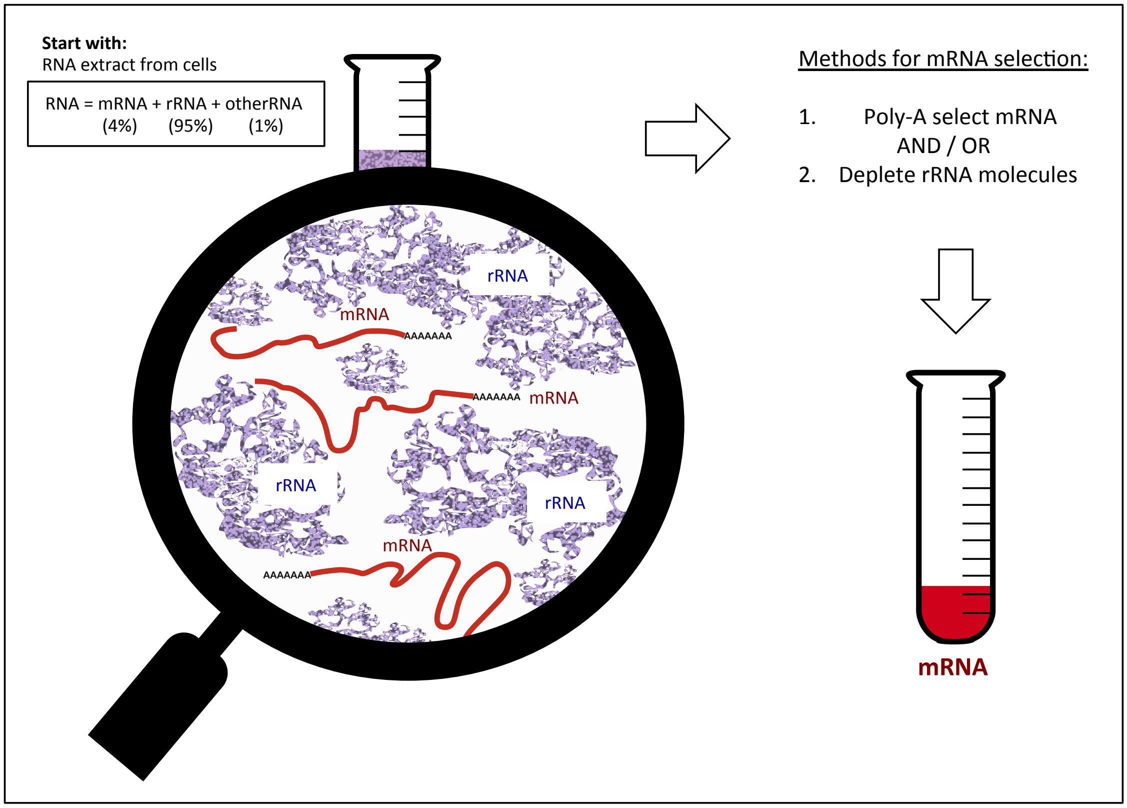
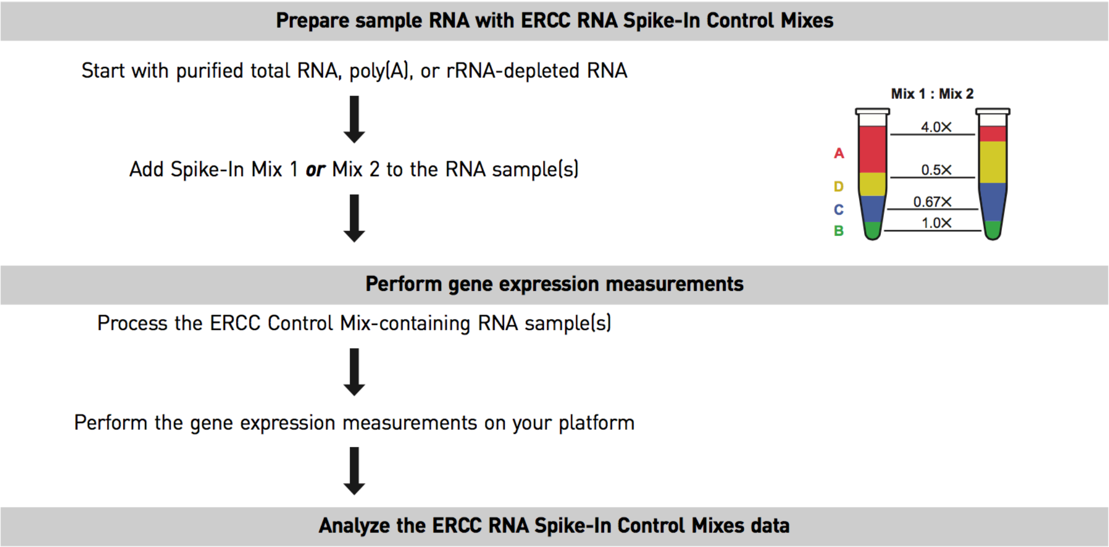


Figure 1.2: **The Poly(A) tails of mRNA.** The entire set of RNA molecules found in a cell is called the transcriptome. A typical cell contains approximately 20 to 30 pg of RNA, which is about 1% of the cell's total mass; while mRNA forms only about 4% of the total RNA in a cell. During library preparation total RNA must be enriched for mRNA by using ploy-dT magnetic beads to bind to the poly(A) tail of mRNA or by using reagents that can digest rRNA.



Prepare sample RNA with ERCC RNA Spike-In Control Mixes

Start with purified total RNA, poly(A), or rRNA-depleted RNA

Add Spike-In Mix 1 ***or*** Mix 2 to the RNA sample(s)

Perform gene expression measurements

Process the ERCC Control Mix-containing RNA sample(s)

Perform the gene expression measurements on your platform

Analyze the ERCC RNA Spike-In Control Mixes data

Figure 1.3: **Ambion ERCC spike-in control mix.** The ERCC has developed a set of 92 mRNA controls (20-mer poly(A) tails) that can be used in gene expression platforms such as RNA-seq, DNA microarrays, and quantitative real-time reverse transcriptase PCR (qRT-PCR). The 92 mRNA transcripts are divided into 4 groups labelled A, B, C, and D. Each group contains 23 mRNA transcripts spanning a 10^6-fold concentration range. The mRNA lengths and composition are similar across the 4 groups. There are two ERCC control spike-in mixes: mix 1 and mix 2. The molar concentration ratios of mix 1 to mix 2 are 4, 1, 0.67, and 0.5 for groups A, B, C, and D respectively. See Ambion's user guide [28] for more information. This figure was obtained as a screen capture from Ambion's user guide [28].

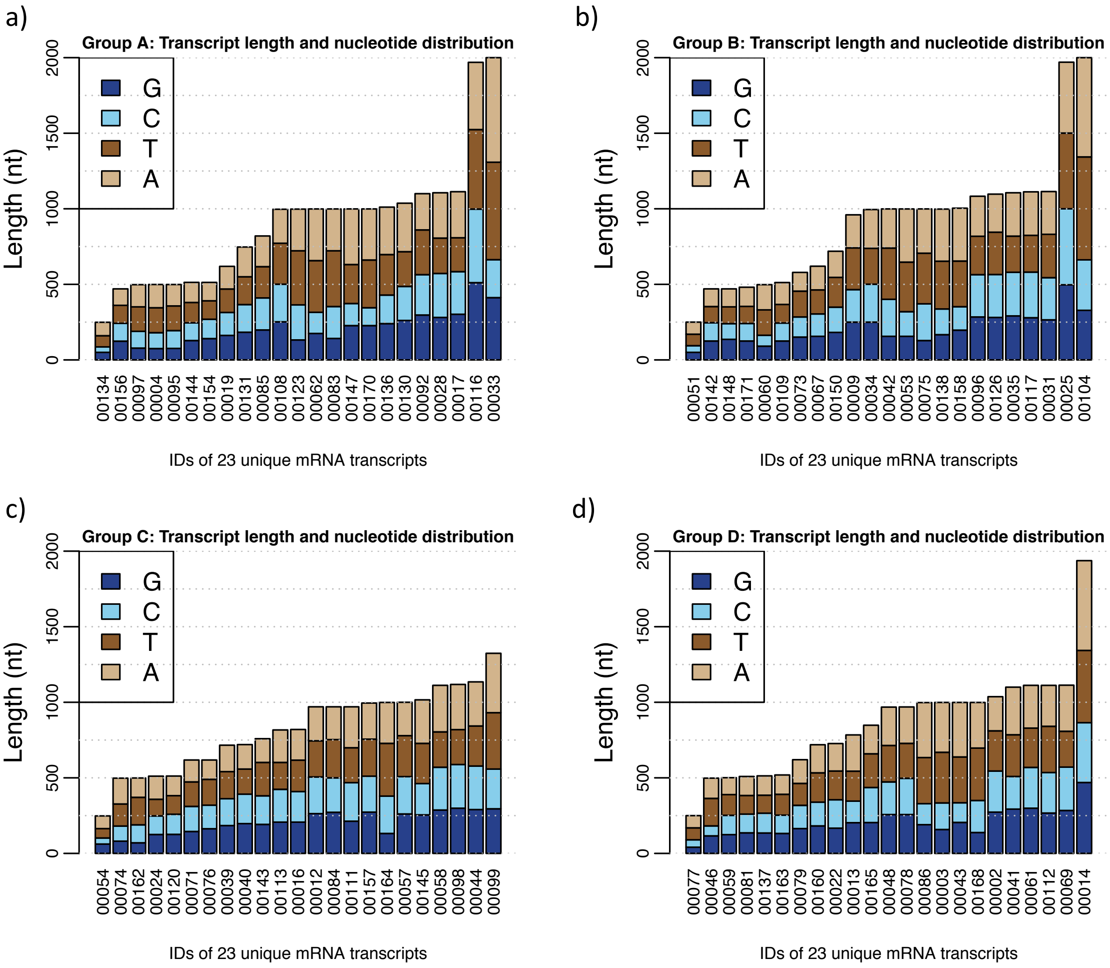


Figure 1.4: **ERCC spike-in composition.** A summary of the nucleotide composition of the ERCC spike-in mix. The 92 mRNA transcripts are divided into 4 groups labelled A, B, C, and D. The mRNA lengths and composition are similar across the 4 groups.

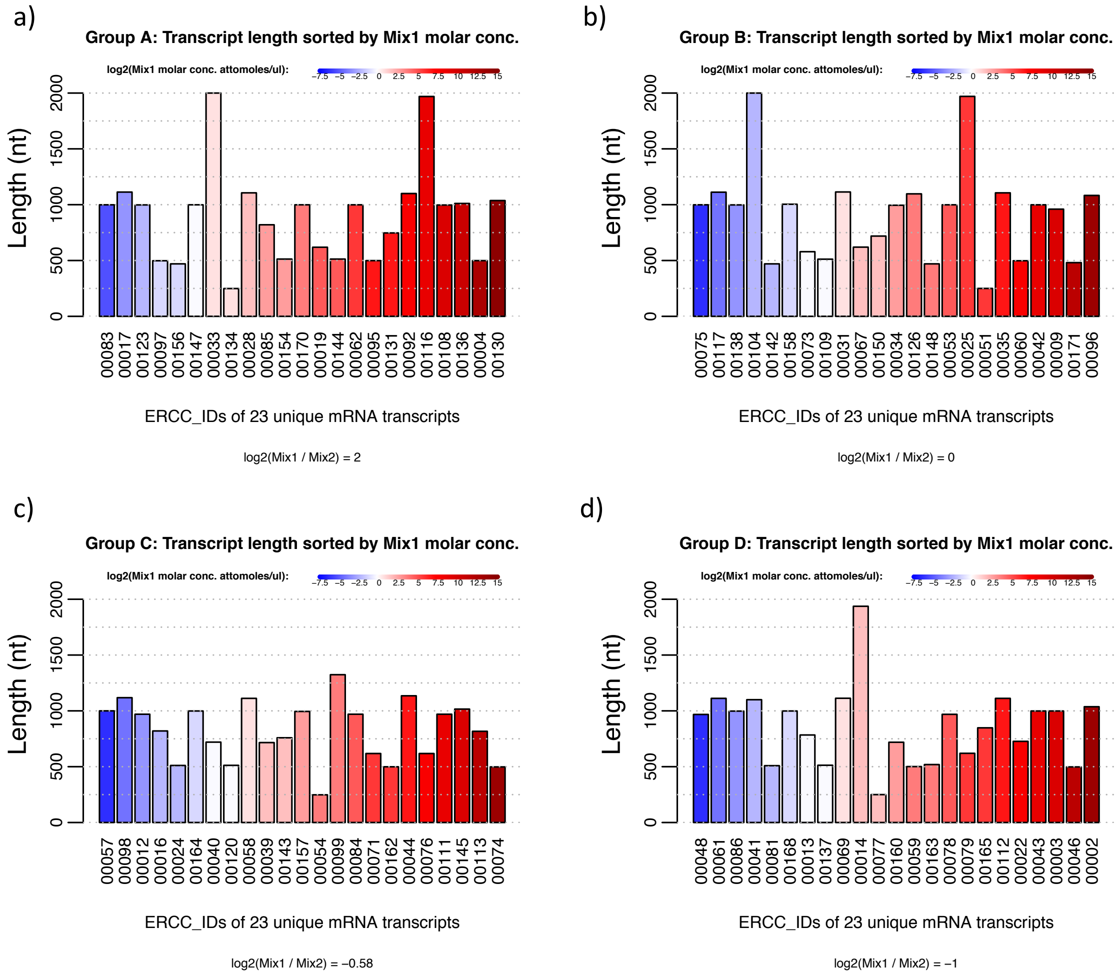


Figure 1.5: **ERCC spike-in intensity range.** The 92 mRNA transcripts in the ERCC spike-in are divided into 4 groups labelled A, B, C, and D. Each group contains 23 mRNA transcripts spanning a 10^6-fold concentration range.

Chapter 2: Shape Analysis of High-throughput Genomics Data

2.1 Introduction

Given a dataset we first try to characterize its central value (location) with the sample mean (or sample median) and its spread around the location (scale) with the sample standard deviation (or sample range). When the sample size, n, is sufficiently large we can begin to assess the shape of the data in some meaningful way. Distributional shape is often characterized by two features (1) skewness: a measure of how far the shape of the distribution deviates from symmetry around its location and (2) kurtosis: a measure of how much weight is at the tails of the distribution relative to the weight around the location. Unlike skewness which has a natural standard (symmetry) there is no standard for kurtosis. Often the Gaussian description of kurtosis is used as a standard. Traditional methods for describing distributional shape are through moments; theoretical moments for the postulated random variable and sample moments for the observed data. Higher sample moments have notoriously high variances, and are non-robust against contamination in the far tails [42, 43]. Kirby [44] points out that the sample moments based method for measuring the coefficient of variation (CV), skewness, and kurtosis of data have algebraic bounds that depend solely on sample size. Even for a relatively large sample size such as 100, and in some cases 1000, sample moments based methods can yield unreliable results [45, 46].

Alternative frameworks for describing distributional shape have been proposed. Two of these are (1) Tukey's g-and-h distribution [38] and (2) L-moments [39]. Both of these frameworks enjoy theoretical and practical advantages over traditional moments [39–42]. The g-and-h distribution is a simple transformation of the standard Gaussian distribution. It has two shape parameters, the skewness parameter (g) and kurtosis parameter (h). The key advantage of this framework lies in the simplicity of the transformation and the ease with which finely calibrated data can be generated. L-moments have several advantages over traditional moments and have been used extensively in Regional Frequency Analysis: the study of statistical methodologies that combine regional data (eg. the monthly maximum amount of rainfall in specific geographic locations) in order to provide better estimates of quantities of interest [47]. The definition of L-moments and some of its properties are formally described in Hosking [39].

In this chapter we use the g-and-h distribution to generate data; and the theory of L-moments along with the parallel nature of high-throughput data to provide a framework for summarizing the shape (skewness and kurtosis) of transcriptome (RNA-seq/microarray) data. From these summaries one can assess how the data deviates from hypothesized distributional assumptions and assess how certain genes deviate from the typical gene in a given dataset. These deviant genes may be classified as volatile (or outliers) or interesting genes. Perhaps they may be genes affected by batch effects, random technical effects, or unknown systematic biological effects [48]. While the current number of samples from high-throughput genomic experiments typically range between 6 and 25, they have the advantage of generating multiple “samples” (e.g. 10,000 genes) in one experiment. This can allow us to say something meaningful about the shape of the data on a global level. This chapter is organized as follows. First we briefly state the definition of the traditional moments and their sample moments estimators. We then introduce the g-and-h distribution and the theory of L-moments showing where they are analogous to (and differ from) traditional moments. We then introduce the SO-plot for the assessment of the shape high-throughput transcriptomic data. Finally we describe an algorithm for detecting volatile genes at a specified cutoff.

2.2 Numerical assessment of shape

2.2.1 Traditional moments

Suppose that we observe the data, $(x_1, x_2, \ldots, x_n)$, assumed to have been generated from a random variable X. For example $(x_1, x_2, \ldots, x_n)$ can be thought of as the measurements of a single gene obtained from n replicates (biological or technical). The traditional method for describing the shape of X is through its moments: $\mu_1 = \mathrm{E}(X)$ and $\mu_r = \mathrm{E}\{(X - \mu_1)^r\},\ r = 2, 3, \ldots$; when they exist. The mean (μ), the standard deviation (σ), the skewness (γ), and the kurtosis (κ) of X are defined as follows: $\mu = \mathrm{E}(X)$, $\sigma = \sqrt{\mathrm{E}[(X-\mu)^2]}$, $\gamma = \mathrm{E}[(X-\mu)^3]/\sigma^3$, and $\kappa = \mathrm{E}[(X-\mu)^4]/\sigma^4$. The coefficient of variation is defined as $CV = \sigma/\mu$ ($\mu \neq 0$). The skewness (γ), kurtosis (κ) and CV are all dimensionless (without units) quantities and are used to compare or characterize various distributional shapes. The measures of shape are not independent. For example, one would expect a highly skewed distribution to have a high kurtosis. The general relationship between the moments measure of skewness and kurtosis is $\kappa \geq \gamma^2+1$ and the range of possible values are $-\infty < \gamma < \infty$ for skewness and $1 \leq \kappa < \infty$ for kurtosis [49]. The corresponding sample moments estimators are [47]:

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2.1}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \tilde{\mu})^2}, \tag{2.2}$$

$$\tilde{\gamma} = \frac{n^2}{(n-1)(n-2)}\Big\{\frac{1}{s^3}\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{\mu})^3\Big)\Big\}, \tag{2.3}$$

$$\tilde{\kappa} = \frac{n^2}{(n-2)(n-3)}\Big\{\Big(\frac{n+1}{n-1}\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{\mu})^4\Big) - 3\Big(\frac{n-1}{n}\Big)^2 s^4\Big\}. \tag{2.4}$$

The CV is estimated with $s/\tilde{\mu}$. The sample moments estimators of skewness and kurtosis are biased and have poor sampling properties when the data are skewed with tails heavier than the Gaussian. In particular, their bias can increase with sample size. See Figure 2.3. As mentioned in the introduction, these estimators have algebraic bounds that depend solely on the sample size. In particular, $|\tilde{\gamma}| \leq \sqrt{n}$ and $|\tilde{\kappa}| \leq n+3$ and for nonnegative data $0 \leq CV \leq \sqrt{n-1}$ [44,47]. For example (page 18 of Hosking and Wallis [47]) the moments based skewness estimate for the log-normal (LN) distribution, which has moments skewness $\gamma = 6.91$, cannot exceed 4.47 for a sample of size 20 ($\sqrt{20} \approx 4.47$). These characteristics make the sample moments a poor candidate for assessing plausible distributional shapes.

2.2.2 Tukey's g-and-h distribution

The g-and-h distribution was introduced by Tukey [38] as a means of numerically summarizing the shape of data. It is formally described in Hoaglin et al. [50]. The g-and-h distribution can take on a wide variety of shapes. It can approximate most of the members of the Pearson system of distributions [40]; which includes the family of chi-square distributions, the family of student's t-distributions, the exponential distribution, and the family of beta distributions just to mention a few. The ability to create diverse shapes with fine control makes the g-and-h appealing for testing the distributional assumptions of statistical methods and algorithms, model selection, and robust exploratory data analysis [50, 51]. See Figure 2.1. The definition of the g-and-h distribution is as follows: let $Z \sim N(0,1)$ and consider the transformation

$$\mathrm{T}_{g,h}(Z) = \frac{\exp{(gZ)} - 1}{g} \exp\left(\frac{hZ^2}{2}\right). \tag{2.5}$$

The random variable $\mathrm{T}_{g,h}(Z)$ is said to have the standard g-and-h distribution. The g parameter controls skewness in both magnitude and direction. When $g = 0$ the distribution is symmetric. When $g > 0$ ($g < 0$) the distribution is skewed to the right (left) with the magnitude of skewness described by $|g|$. For a given g, the parameter h describes how much more ($h > 0$) probability weight is at the tails relative to the Gaussian distribution ($h = 0$). When $h < 0$ the tails are lighter than the Gaussian. For example the t-distribution with 10 degrees of freedom (df) is approximated by ($g = 0, h = 0.058$), and the uniform distribution is approximated by ($g = 0, h = -0.244$) [40]. Once the shape of the standard g-and-h distribution has been specified we can scale and shift it as desired: $X = A + B\mathrm{T}_{g,h}(Z)$, where A is a location parameter and $B > 0$ is a scale parameter. We will denote this distribution as $\mathrm{GH}(A, B, g, h)$. The Gaussian and log-normal (LN) family of distributions are exact cases of the g-and-h distribution; $\mathrm{N}(\mu, \sigma) = \mathrm{GH}(A = \mu, B = \sigma, g = 0, h = 0)$

and $\mathrm{LN}(\mu, \sigma) = \mathrm{GH}(A = \sigma^{-1}, B = \sigma e^{\mu}, g = \sigma, h = 0)$. More examples of g-and-h approximations are: (1) the chi-square with 4 df and 10 df are approximated by $(g = 0.502, h = -0.046)$ and $(g = 0.303, h = -0.017)$ respectively; (2) the exponential distribution is approximated by $(g = 0.760, h = -0.098)$; and (3) the Cauchy distribution is approximated by $(g = 0, h = 0.97)$ [40]. The nth moment of the g-and-h distribution only exists when $h < 1/n$; the moments have been derived by Martinez and Iglewicz [40] and are stated below.

2.2.3 Moments of g-and-h distribution

Below we state the moments of the g-and-h distribution (see Martinez and Iglewicz [40] for the derivation). Suppose that $g = 0$ (ie. the g-and-h distribution is symmetric). If n is odd then the n^{th} sample moment is 0. If n is even: $\mathrm{E}(\mathrm{T}^n_{0,h}) = n!/((n/2)!\sqrt{2^n})(1 - nh)^{-(n+1)/2}$; $h < 1/n$. When $g \neq 0$:

$$\mathrm{E}(\mathrm{T}^n_{g,h}) = \frac{1}{g^n\sqrt{1 - nh}} \sum_{k=0}^{n} (-1)^k \binom{n}{k} \exp\left\{\frac{[(n-k)g]^2}{2(1 - nh)}\right\}; \quad h < \frac{1}{n}. \tag{2.6}$$

2.2.4 L-moments

2.2.4.1 Definition

The theory of L-moments has been used heavily in the study of Regional Frequency Analysis and Hydrology for estimation and exploratory data analysis purposes [47]. The r^{th} L-moment of a random variable X is defined as

$$\lambda_r = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} \mathrm{E}(X_{r-k:r}), \quad r = 1, 2, \ldots \tag{2.7}$$

where $X_{i:n}$ is the i^{th}-order statistic from a sample of size n. The “L” in L-moments is based on the fact that they are defined as (L)inear combinations of order statistics [47]. Analogous to traditional moments the first four L-moments describe the location, scale, skewness, and kurtosis of a random variable X: $\lambda_1 = E(X)$, $\lambda_2 = (1/2)\mathrm{E}(X_{2:2} - X_{1:2})$, $\lambda_3 = (1/3)\mathrm{E}(X_{3:3} - 2X_{2:3} + X_{1:3})$, and $\lambda_4 = (1/4)\mathrm{E}(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4})$. L-moments uniquely characterize any random variable whose first moment is finite [39].

2.2.4.2 Interpretation of the first four L-moments

We shall now describe an interpretation of the first four L-moments that will play a crucial role in the way in which we understand our results. First, note that λ_1 and λ_2 (L-scale) are known methods of describing location and scale respectively. In particular, λ_1 is the expected value of X and λ_2 is the average range of a conceptual sample of size 2. As described in Hosking and Wallis [47], when we rewrite $X_{3:3} - 2X_{2:3} + X_{1:3}$ as $(X_{3:3} - X_{2:3}) + (X_{1:3} - X_{2:3})$ we see that λ_3 is a measure of skewness. If X is symmetric then $(X_{3:3} - X_{2:3})$ will tend to equal $(X_{1:3} - X_{2:3})$ and $\lambda_3 \approx 0$; on the other hand if X is skewed to the right (left) $(X_{3:3} - X_{2:3})$ will tend to be bigger (smaller) than $(X_{2:3} - X_{1:3})$ making λ_3 positive (negative). For kurtosis, write $X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}$ as $\{(X_{4:4} - X_{3:4}) + (X_{2:4} - X_{1:4})\} - 2(X_{3:4} - X_{2:4})$. The term $\{(X_{4:4} - X_{3:4}) + (X_{2:4} - X_{1:4})\}$ measures how far the extremes differ from the third and second quartiles, whereas the term $2(X_{3:4} - X_{2:4})$ measures the middle spread between the third and second quartiles. If a distribution has heavy tail we would on average expect $\{(X_{4:4} - X_{3:4}) + (X_{2:4} - X_{1:4})\}$ to be greater than $2(X_{3:4} - X_{2:4})$, making λ_4 positive. If X has a relatively uniform shape we would on average expect $\{(X_{4:4} - X_{3:4}) + (X_{2:4} - X_{1:4})\}$ to be similar to $2(X_{3:4} - X_{2:4})$, making $\lambda_4 \approx 0$.

In fact $\lambda_4 = 0$ for all symmetric distributions. If a distribution has a light tail (eg. Beta(0.25, 0.25)) we would on average expect $\{(X_{4:4} - X_{3:4}) + (X_{2:4} - X_{1:4})\}$ to be less than $2(X_{3:4} - X_{2:4})$, making λ_4 negative (see Figure 2.4). Hosking [39] defines unit-free coefficients analogous to the traditional moment's coefficient of variation (CV), skewness, and kurtosis as follows: $\tau = \lambda_2/\lambda_1$ (L-CV), $\tau_3 = \lambda_3/\lambda_2$ (L-skew), and $\tau_4 = \lambda_4/\lambda_2$ (L-kurt). Some examples of L-moments are: $(\lambda_1 = 1/2, \lambda_2 = 1/6, \tau_3 = 0, \tau_4 = 0)$ for the uniform distribution; $(\lambda_1 = 0, \lambda_2 = 1/\sqrt{\pi}, \tau_3 = 0, \tau_4 \approx 0.1226)$ for the Gaussian family; and $(\lambda_1 = 1, \lambda_2 = 1/2, \tau_3 = 1/3, \tau_4 = 1/6)$ for the standard exponential distribution. L-skew and L-kurt are constrained in $-1 < \tau_3 < 1$ and $0.25(5\tau_3^2 - 1) \leq \tau_4 < 1$ and when $X \geq 0$ the L-CV is constrained in $0 < \tau < 1$ [39]. The finite bounds provide a key advantage over the g-and-h distribution framework and the traditional moments measures for skewness and kurtosis since they $(g, h, \gamma,$ and $\kappa)$ are not bounded and can take arbitrarily large values. It is usually easier to graphically summarize (and interpret) a diagnostic number that is bounded versus one that is not, for example the correlation coefficient versus covariance.

2.2.4.3 Estimation

Below we describe the estimation procedure for finding L-moments from data [39,52]. Suppose that we sort our sample, $(x_1, x_2, \ldots, x_n)$, to obtain $x_{1:n} \leq x_{2:n} \leq \ldots \leq x_{n:n}$. The estimate of λ_r is defined as a linear combination of the sorted sample; $l_r = (1/n)\sum_{j=1}^{n} w_{j:n}^{(r)} x_{j:n}$, where the weights are

$$w_{j:n}^{(k)} = \sum_{i=0}^{min\{j-1,k-1\}} (-1)^{k-1-i} \binom{k-1}{i} \binom{k-1+i}{i} \binom{j-1}{i} / \binom{n-1}{i}. \tag{2.8}$$

The l_rs are called l-statistics. They are unbiased and their exact variance covariance structure have been derived by Elamir and Seheult [53]. Whereas the r-th traditional sample moment raises the sample to the r-th power, the r-th l-statistic only takes a linear combination of the sorted sample, and as such provides more stable results. The L-CV (τ), L-skew (τ_3), L-kurt (τ_4) are estimated as follows: (1) $\hat{\tau} = l_2/l_1$, (2) $\hat{\tau}_3 = l_3/l_2$, and (3) L-kt $\hat{\tau}_4 = l_4/l_2$. These ratios like their traditional moments counterparts are biased. However they have relatively little bias for small samples (see Figure 2.3). Given a g-and-h random variable X we can compute its traditional moments and L-moments. See Figure 2.2. The theory of L-moments holds a lot of potential for high-throughput transcriptome experiments data. The reader is encouraged to read Hosking [39] for a more thorough and formal introduction.

2.2.5 L-moments of the g-and-h distribution

For a continuous random variable, X, with quantile function $Q(u),\ u \in (0,1)$. The n^{th} L-moment can be defined as: $\lambda_n = \int_0^1 Q(u)P^*_{n-1}(u)du, \quad n = 1, 2, \ldots,$ where

$$P^*_r(u) = \sum_{k=0}^{n} (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} u^k, \quad n = 0, 1, 2, \ldots \tag{2.9}$$

is the n^{th} shifted Legendre polynomial [39]. We numerically integrate λ_n to find the n^{th} L-moment of the g-and-h distribution. See Figure 2.2.

2.3 Symmetry-Outlier Plot (SO-plot)

The L-moment ratio diagram is the analog of the conventional moment ratio diagram [54]. It was introduced by Hosking [39] and has been used as a graphical component in the process of selecting a model for observed hydrological data [43, 45].

The L-moment ratio diagram depicts the theoretical relationship between the L-skew and the L-kurt of a distribution. See Figure 2.4 and Figure 2.5 Families of distributions with a fixed shape such as the uniform and the Gaussian are depicted as single points. The shape of these distributions are constant as a function of their parameters. Distributions whose shape change with their parameters are depicted as curves or two dimensional regions on the L-moment ratio diagram depending on the variety of shapes that they can take. For example the generalized Pareto distribution (GPA) and log-normal distribution show up as curves and the five-parameter Wakeby (WA5) distribution shows up as a filled parabola see [45].

We have renamed the L-moment ratio diagram as the symmetry-outlier plot (SO-plot) to emphasize its utility within the context of high-throughput transcriptome data. It depicts the theoretical relationship between L-skew (τ_3) and L-kurt (τ_4) of the g-and-h family of distributions as a guide. In particular we show the curves: $\tau_4 = f_{g,h}(\tau_3),$ for $g \in (-\infty, \infty)$ and $h \in \{0, 0.25, 0.5\}$, where $f_{g,h}(.)$ is the mathematical dependence of τ_4 on τ_3. That is, the expected relationship between skewness and kurtosis (Figure 2.4). For a given h, the SO-plot diagram summarizes the mapping of $g \in (-\infty, \infty)$ into $\tau_3 \in (-1, 1)$. On each curve $f_{g,h}$ we have shown the points $g \in \{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$ for reference. Given a high-throughput dataset we plot each genes's L-skew and L-kurt estimate on the SO-plot and summarize the estimates with box-plots. (see Figures 3.1 and 3.3).

Within the context of log RNA-seq data or microarray intensities we have partitioned the L-skew range into 4 regions to aide in interpretation: (1) Minor skew $|\tau_3| \leq 0.05$, (2) Moderate skew $0.05 < |\tau_3| \leq 0.2$, (3) Large skew $0.2 < |\tau_3| \leq 0.35$, and (4) Extreme/volatile skew $0.35 < |\tau_3| < 1$. Due to the dependence of kurtosis on skewness, an L-kurt estimate should be interpreted based on the its corresponding L-skew estimate. In general if L-skew is large the we expect L-kurt to be large as well. Genes with a large (in absolute value) L-kurt estimate observed at a relatively small L-skew estimate (minor skew - moderate skew) tend to be volatile/outliers.

2.4 Detection of volatile genes

We now present a method for finding genes within a dataset whose sample distribution is markedly different from the majority of genes in the same dataset. Suppose that we have performed an RNA-seq experiment where the reads have been aligned, summarized into a count matrix, and normalized for differences in library size. We also assume that known systematic sources of biological variation have been removed from the data.

We first summarize the shape of each gene by computing its L-skew (τ_3) and L-kurt (τ_4) estimates to obtain: $\tau_{3_g},\ g = 1, 2, \ldots G$ and $\tau_{4_g},\ g = 1, 2, \ldots G$. Due to the general dependence of L-kurt on L-skew we adjust each τ_{4_g} by assuming that it is a sum of a skewness component and a skew-adjusted kurtosis component ($\tau^*_{4_g}$). The skewness component is large when L-skew is large (close to 1 or -1), and small when L-skew is small (close to 0). We fit a model to estimate the skewness component and adjust L-kurt as follows: $\tau^*_{4_g} = \tau_{4_g} - \text{lowess}(\tau_{4_g} \sim \tau_{3_g})$, where $\text{lowess}(\tau_{4_g} \sim \tau_{3_g})$ denotes a lowess fit of the scatter plot (τ_{3_g}, τ_{4_g}) (Figures 2.6 and 2.7 b). The adjusted pair $(\tau_{3_g}, \tau^*_{4_g})$ is assumed to follow a bivariate Gaussian distribution [39]. For each gene we compute its statistical square distance from the center:

$\mathrm{D}_g = \mathbf{u}_g^T \mathbf{A}^{-1} \mathbf{u}_g$, where $\mathbf{u}_g = (\tau_{3_g} - (1/G)\sum_{g=1}^{G} \tau_{3_g},\ \tau_{4_g}^{*} - (1/G)\sum_{g=1}^{G} \tau_{4_g}^{*})$ and $\mathbf{A}$ is the sample variance-covariance matrix of $(\tau_{3_g}, \tau_{4_g}^{*}),\ g = 1, 2, \ldots, G$. For each squared distance we compute the corresponding outlier score (d-values): $d_g = 1 - F_{\chi_2^2}(D_g),\ g = 1, 2, \ldots, G$ where $F_{\chi_2^2}(\cdot)$ is the distribution function of chi-square with 2 df. A gene g is called volatile (or an outlier) if $d_g \leq \alpha$ where $\alpha \in [0, 1]$. Usually we set α to 0.01% or 0.1%. Note that we use the term outlier here in the sense that the sample shape (independent of location and scale) of the gene is markedly different from the majority of the genes in the dataset.

2.5 Volatile genes can show unknown systematic effects

We applied the outlier detection algorithm to the Pickrell dataset [55] (Figure 2.6 and 2.7). The Pickrell dataset is part of the International HapMap Project. RNA samples were extracted from the lymphoblastoid cell lines of 69 unrelated Nigerian individuals, 29 males and 40 females. The aligned and summarized count matrix was obtained from the bioconductor (http://www.bioconductor.org) tweeDEseqCountData package. Genes with at least one count per million (cpm) in 29 or more samples were kept and normalized for library size (cpm).

Without adjusting for gender we found 18 genes (cutoff $\alpha = 0.01\%$) to be "outliers". After adjusting for gender (subtracting gender means) 10 out of the 18 genes were no longer called outliers. These 10 genes showed systematic differences with respect to gender. In Figure 2.7 we have indicated the Hugo Gene Nomenclature Committee (HGNC) gene symbol and chromosome number where available. Note that we use the term outliers in the sense that these genes were markedly different from the majority of the genes in the dataset with respect to their shape (L-skew and L-kurt estimate); which is independent of location (expression level) and scale (spread).

2.6 Discussion and Conclusions

In this chapter we have introduced L-moments statistics, in particular the L-skew and L-kurt ratios. We use the summarize the shape of individual genes, and when taken together they can give is a global view of our the shape of our dataset (gene-wise). In the next chapter we will use L-skew and L-kurt estimates of the samples (sample-wise) to test some of the global distributional assumptions used in normalization techniques. To summarize, we have built on the sound statistical properties of the L-moments ratio estimators to provide a framework for exploring the distributional shapes of genes and the detection of genes (volatile/outlier genes) with shapes that are markedly different from the majority in a given high-throughput transcriptome dataset (SO-plot). It is our hope that the SO-plot will become part of the repertoire of the RNA-seq data analyst.

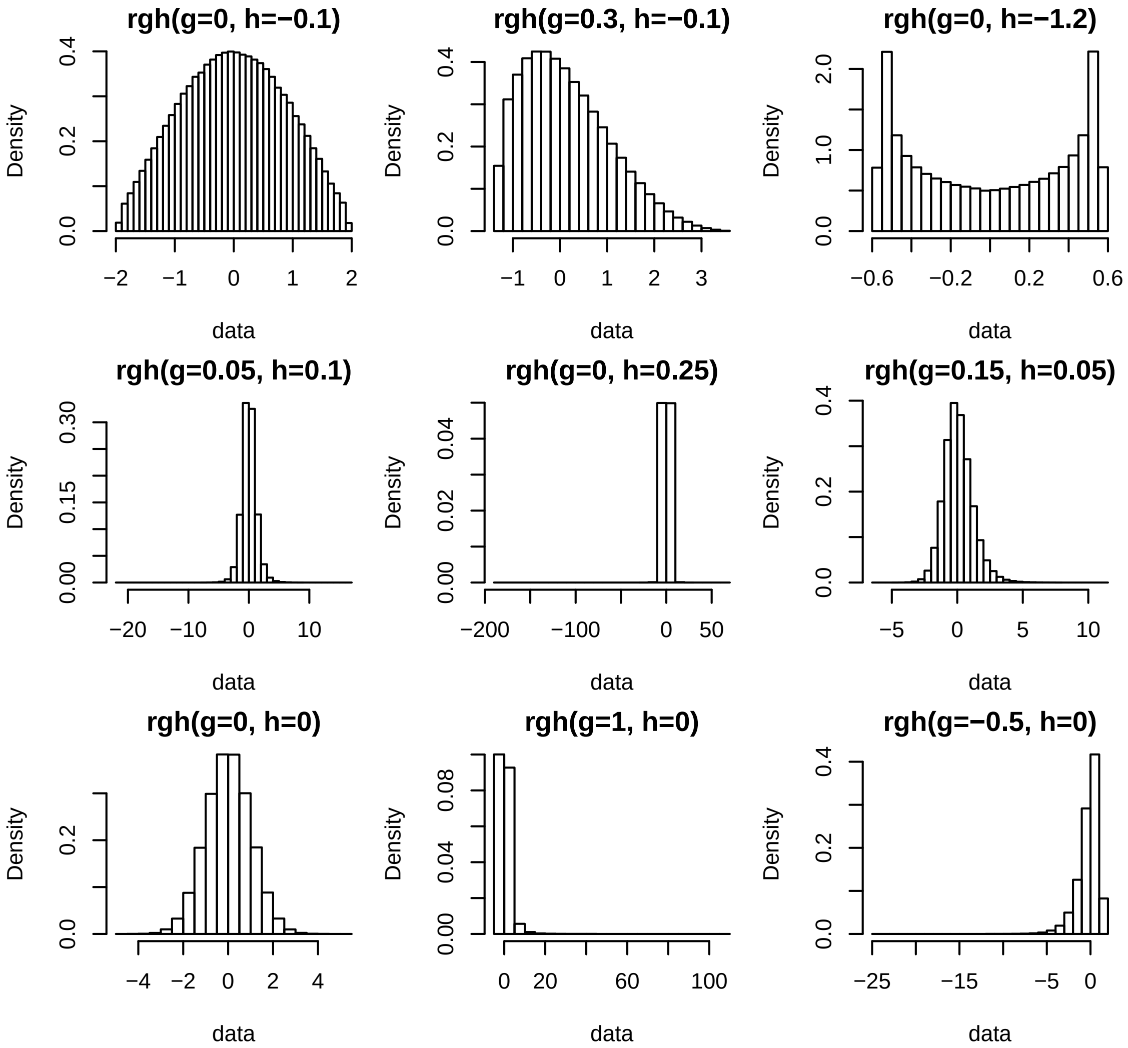


Figure 2.1: **The g-and-h distribution can take on various shapes.** The g-and-h distribution can take on a variety of shapes. This makes it very easy to generate data in a finely calibrated fashion.

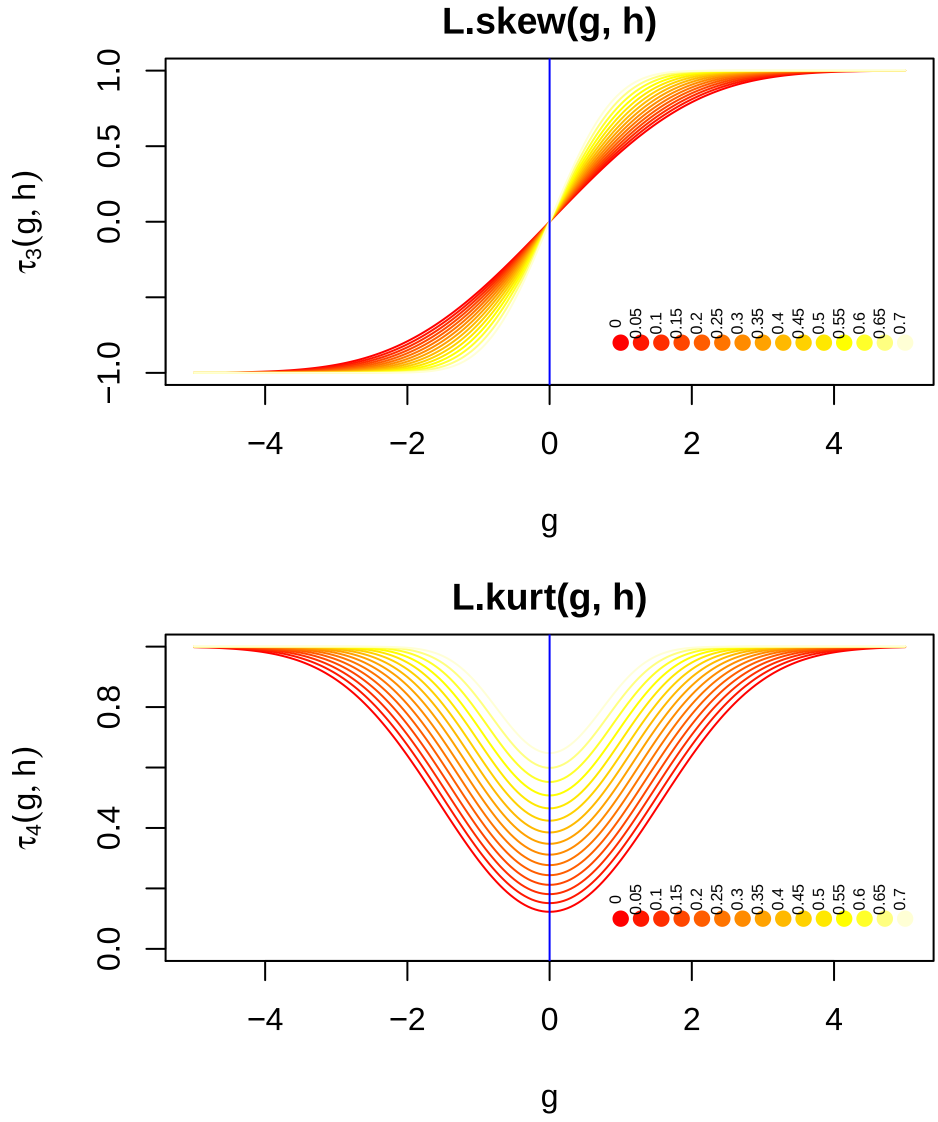


Figure 2.2: **L-moments ratios as a function of** g **and** h**.** For a given g-parameter and h-parameter we can compute the corresponding L-skew, $\tau_3(g,h)$, and L-kurt, $\tau_4(g,h)$, parameters. We have indicated the h-parameter by color.

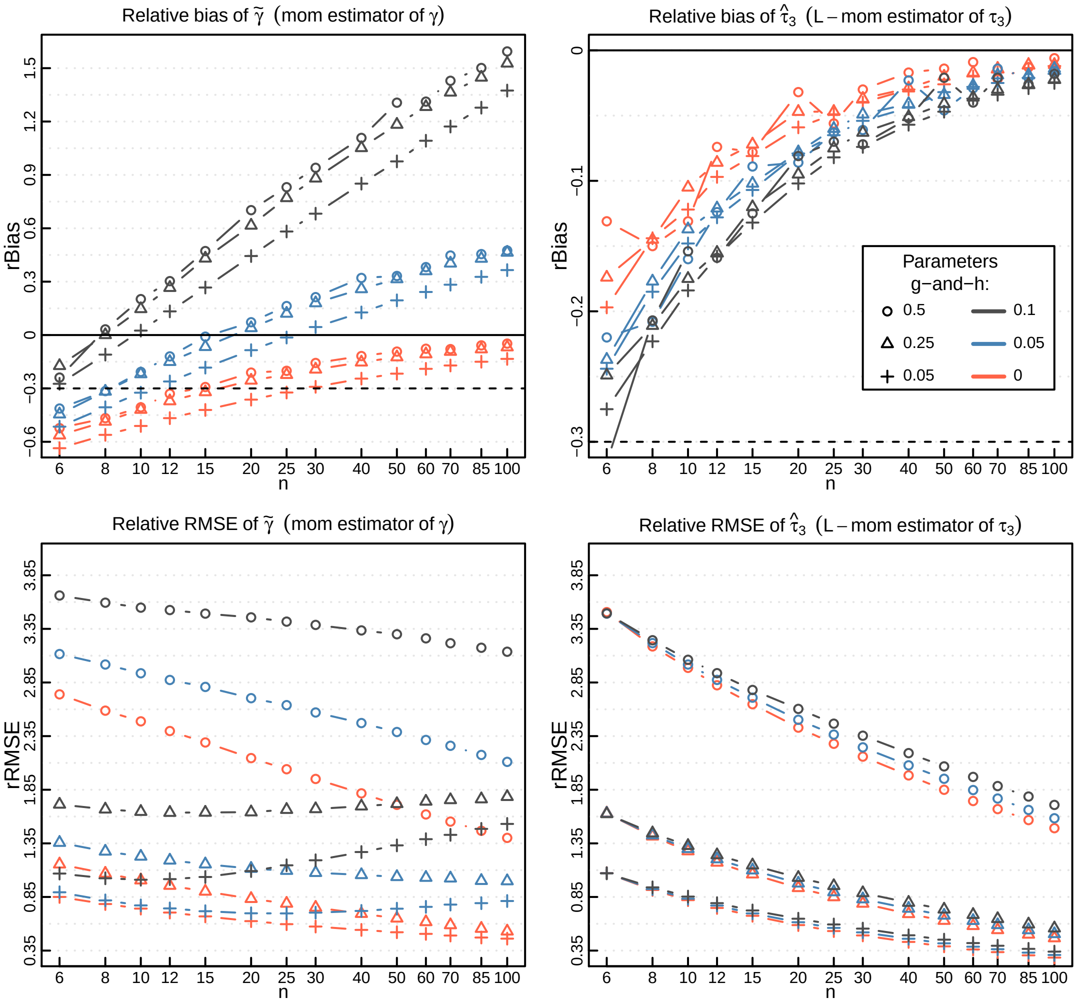


Figure 2.3: **The sampling properties of the traditional moments estimator for skewness ($\tilde{\gamma}$) and L-moments estimator for skewness ($\hat{\tau}_3$).** In the first column we have shown the relative bias (rBias) and relative RMSE (rRMSE) of $\tilde{\gamma}$ based on 100,000 simulations at $(g, h) \in \{0.05, 0.25, 0.5\} \times \{0, 0.05, 0.1\}$. In the second column we have done the same for $\hat{\tau}_3$. Note that $\hat{\tau}_3$ has less bias than $\tilde{\gamma}$ for each (g, h) (indicated by the broken horizontal line). For relatively mild tails (Gaussian tail, $h = 0$) $\tilde{\gamma}$ is more efficient (has less variance) than $\hat{\tau}_3$. However its bias increases with sample size when $h > 0$; underscoring its excessive sensitivity to changes in the underlying distribution.

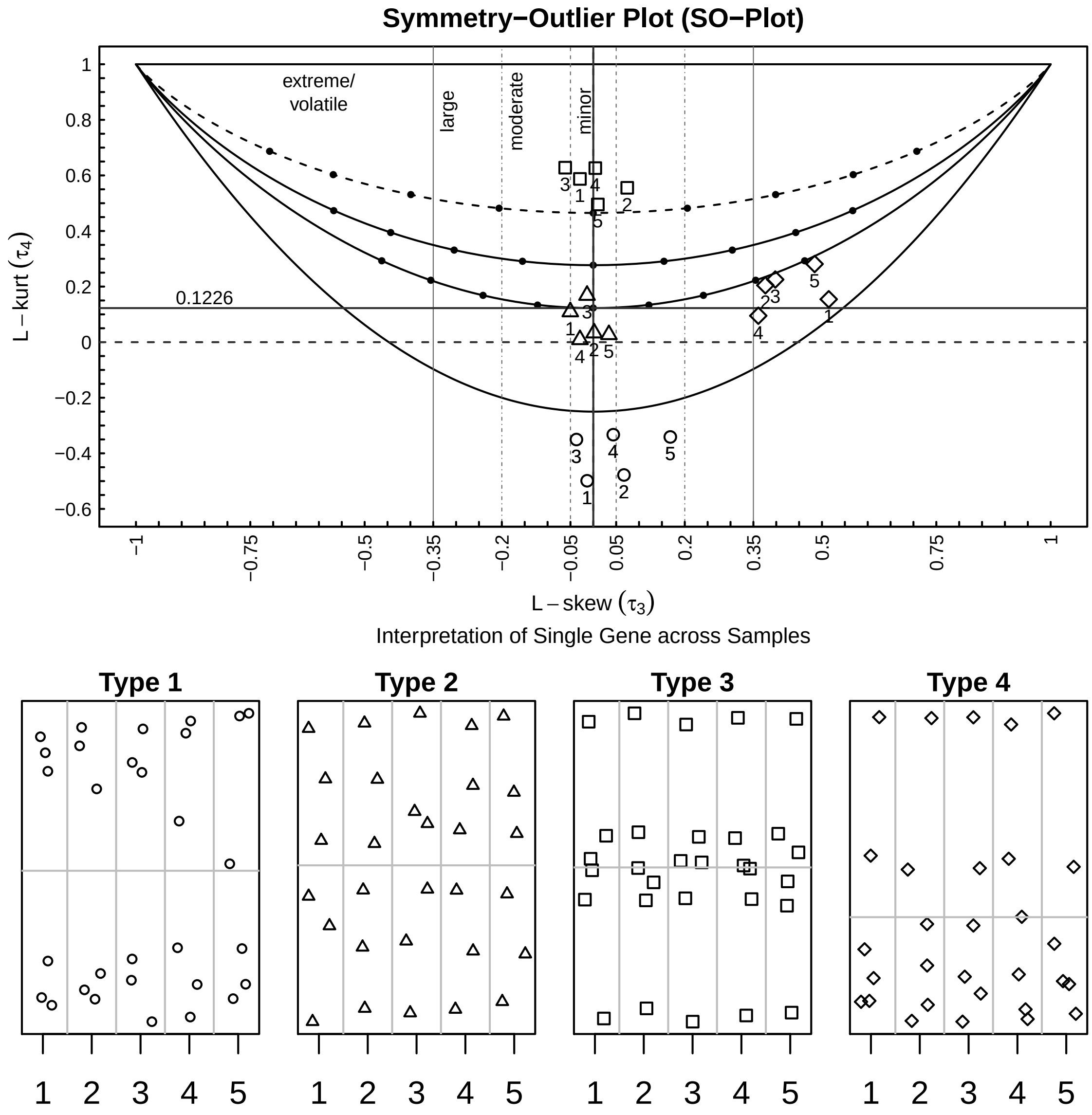
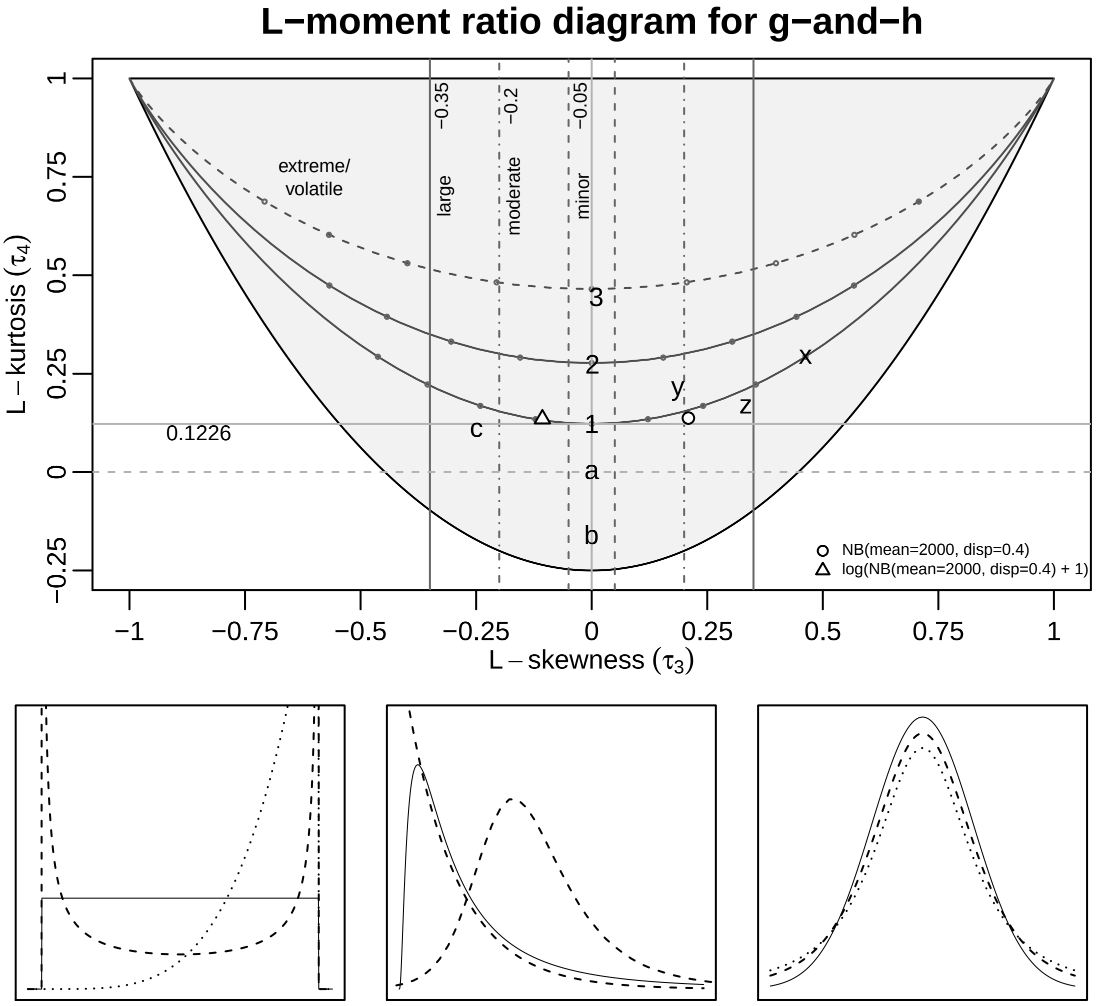


Figure 2.4: **Interpretation of the SO-plot.** We have shown examples, based on a sample of size six, of four main types of sample shape (bottom row) and where they occur on the SO-plot (top row). On the SO-plot we have shown 4 parabolas for reference. Starting from below: (1) the theoretical lower bound for L-kurt, in terms of L-skew, $\tau_4 = 0.25(5\tau_3^2 - 1)$. (2) The curve that indicates all possible shapes of the g-and-h distribution with h fixed at 0, (3) with h fixed at 0.25, and (4) with h fixed at 0.5. On the curves (2), (3) and (4) we have indicated the points at which $g \in \{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$ from left to right. See Figure 2.5 for a theoretical perspective.



a:Unif(0, 1), b:Beta(0.25, 0.25), c:Beta(5, 1)

x:LN(0, 1), y:GH(0.35, 0.1), z:Exp(1)

1:N(0, 1), 2:t(3), 3:t(2)

Figure 2.5: **L-moment ratio diagram for the g-and-h family.** In the bottom row panels we have shown various distributional shapes (three in each plot). The solid curve is the first distribution indicated below the plot (from left to right). For example in the first plot the solid curve is Unif(0, 1). The dashed curve is Beta(0.25, 0.25), and the dotted curve is Beta(5, 1). In the top row we have shown where these 9 distributions fall on the L-moments ratio diagram. Also included are NB(μ=2000, ϕ=0.4) and log(NB(μ=2000, ϕ=0.4)+1). The theoretical region of possible L-moment ratios is indicated by the shaded area and is bounded above by 1 and below by the parabola $0.25(5\tau_3^2 - 1)$. Also shown are some members of the g-and-h family. Specifically $(g, h) \in (-\infty, \infty) \times \{0, 0.25, 0.5\}$. On these curves we have also shown $g \in \{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$ (in knots). For example the point X is $(g = 1, h = 0)$ ($h = 0.5$ is shown in a dashed curve).

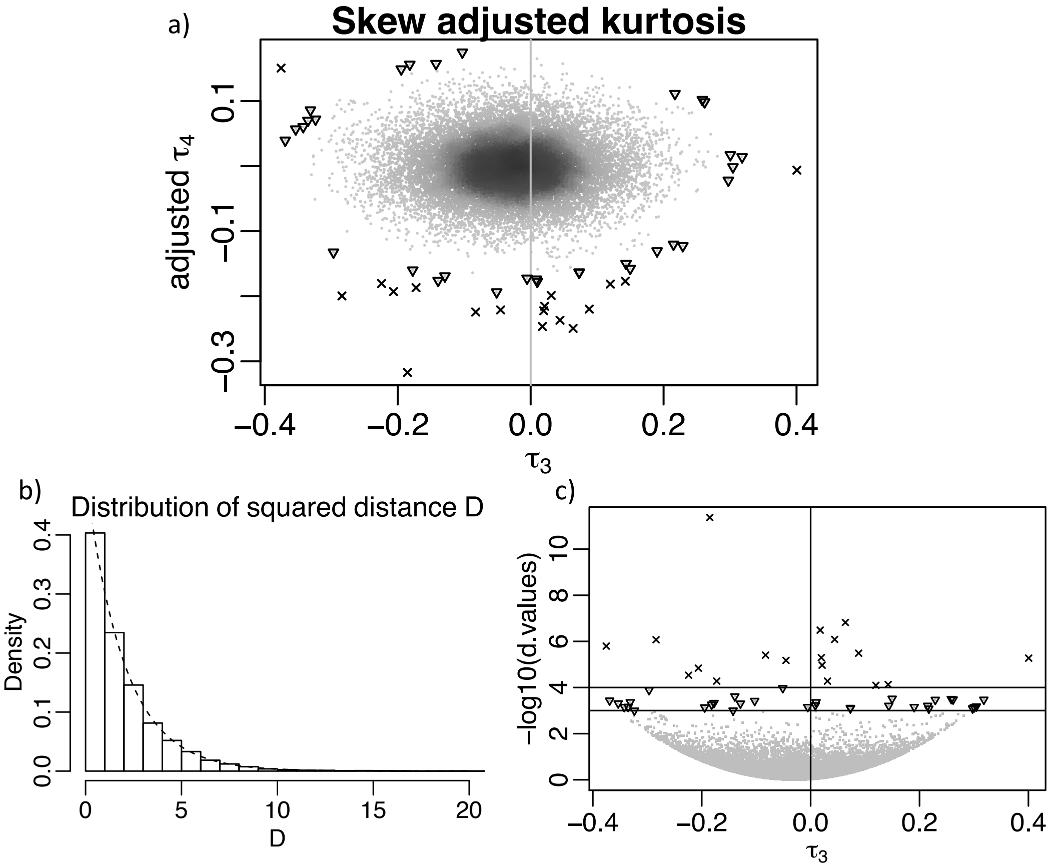


Figure 2.6: **Outlier detection algorithm (Pickrell dataset).** Above we have shown the steps in computing outliers. First we compute the L-moment ratio estimates for each gene and fit a lowess curve to adjust the L-kurt (τ_4) estimates for their dependence on L-skew (τ_3). The adjusted estimates are shown in panel (a). In panel (b) we plot the histogram of the squared distances. We have overlaid the density (broken curve) of the chi-square with 2 degrees of freedom. In the panel (c) we have shown the d-values as a function of L-skew. The cutoffs 0.01%, and 0.1% are shown as horizontal lines.

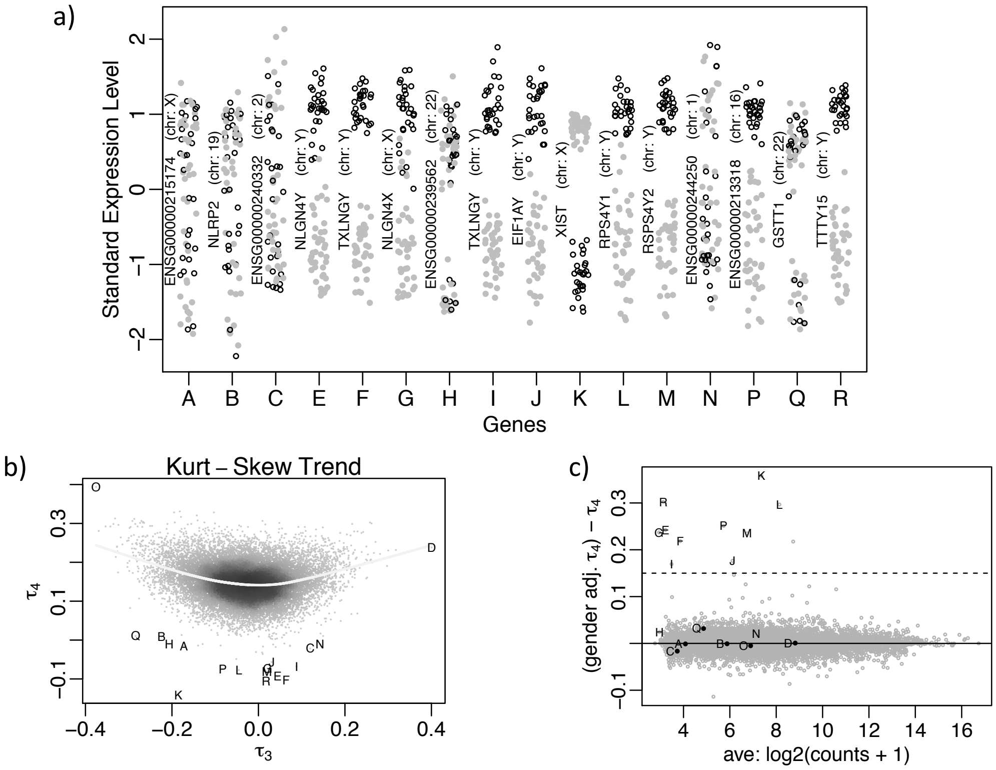


Figure 2.7: **Volatile/outliers genes.** In the plot above we demonstrate the outlier detection algorithm on the Pickrell dataset. The level of detection was set at 0.01%. Eighteen genes were called outliers. In panel (a) we have shown 16 of these genes (see Supplementary Figure 10 for genes O and D). We have shown the standardized expression level (subtract sample mean and divide by sample standard deviation) for the selected genes. We have indicated gender by color. Gray is female and black is male. In panel (b) we have indicated the location of the 18 genes on the SO-plot. In panel (c) we have shown the change in τ_4 after we account for gender. The genes (K, L, P, M, J, R, G, E, F, and I) above the broken line were gender specific. They were not called outliers (at 0.01%) after adjustment. Genes that remained outliers (at 0.01%) after gender adjustment were C, A, Q, B, O, and D.

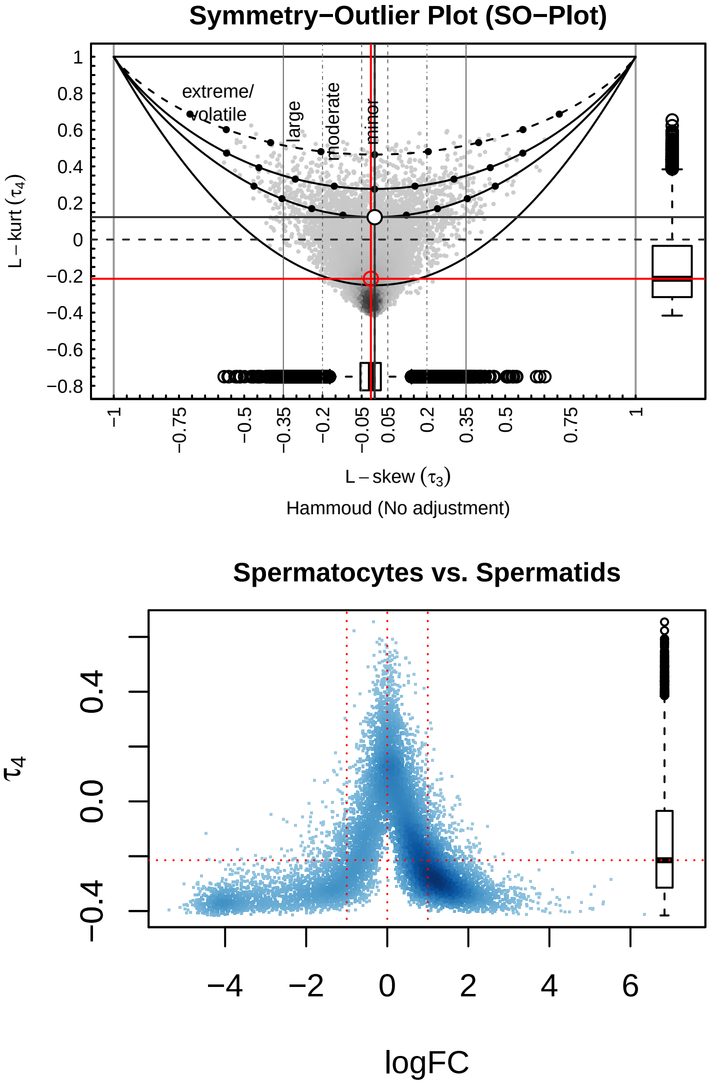


Figure 2.8: **Two group information in the SO-plot.** We have shown the SO-plot for the Hammoud datast without adjusting for cell type. Majority of the genes that are differentially expressed between the two cell types (Spermatocytes vs. Spermatids) are interpreted by the SO-plot to a have high negative L-kurt (τ_4) estimate. In the bottom panel we have shown τ_4 estimates as a function of the log-fold change between the Spermatocytes samples (5) and the Spermatids samples (5).

Chapter 3: A universal framework for understanding robustness

3.1 Introduction

Assays based on massively parallel next-generation sequencing platforms [56] have become the technology of choice for a large variety of transcriptomic studies in recent years due to its decreasing cost and measurement advantages over microarray platforms; including increased dynamic range [57]. These technological improvements in measurements have been accompanied by the development of new algorithms and statistical methodologies to analyze the data they produce. Chief among these are methods designed to detect differentially expressed genes between two or more groups of interest. Two commonly used frameworks for solving this kind of problem have emerged: (1) those based on the assumption that the counts generated by the sequencing process follow a negative binomial (NB) distribution, for example DESeq [15] and edgeR [17]; and (2) those assuming that statistics based on log-transformed counts follow a Gaussian distribution, for example the voom [36] transformation in limma [16]. Some recent articles have compared these two frameworks, for example Soneson et al. [58] and Rapaport et al. [59]. Both articles report that modified microarray (ie. linear Gaussian models) based methods when applied to log-transformed RNA-seq data yield comparable results to methods based on counts, even when the data are simulated from a negative binomial distribution.

In this chapter we demonstrate how the SO-plot can be used as a universal framework for assessing the robustness of methods for analyzing RNA-seq datasets that are based on the Gaussian linear model. To demonstrate the advantages of the Gaussian linear model framework we simulate data from the NB distribution and show that with the appropriate weights, we can compete with NB based models, and even do better. In particular we compute weights from L-moments statistics and use them in limma as user specified weights. By doing so we improve the power (compared with voom-limma) and in some cases outperform NB binomial models, while at the same time controlling the observed false discovery rate (FDR) below the specified nominal rate.

3.2 The SO-plot reveals extent of data agreement with distributional assumptions

In Figures 3.2 and 2.7 we have shown the SO-plot for the log-transformed counts of four RNA-seq datasets: Bottomly [60], Hammoud [61], MAQC [62], and Pickrell [55] and the Geng [16] (microarray) dataset. (Please see appendix A for a description of the datasets; when the need arise in this dissertation will give the necessary details about a dataset.) We observe that the Bottomly, MAQC, and Pickrell datasets have a minor negative skew; the Hammoud dataset is fairly symmetric; and the SO-plot description of the Geng (microarray) dataset supports the widely held assumption that the log-intensities of microarray chips are approximately Gaussian [16]. See Figure 3.1 for an example of an SO-plot simulated from Gaussian ($g = 0, h = 0$) data.

We summarize the distribution of the L-skew estimates for a given dataset as follows: (25% quantile, median, 75% quantile) of the τ_3 estimates. For example, for the ten C57BL/6J samples in the Bottomly dataset the L-skew summary is (-0.18, -0.04, 0.1). In Figure 3.3 we show the SO-plot for the eleven DBA/2J samples in the Bottomly dataset, as well as the SO-plots for a random subsample from the same eleven DBA/2J samples with sizes 9, 8, and 6. For all these sample sizes ($n = 11, 9, 8, 6$) the median L-skew estimate is -0.05; demonstrating the SO-plots ability to provide a consistent measure of skewness across various sample sizes.

3.3 T-test based methods are robust for log negative binomial data

In order to assess the loss of power and potential increase in the FDR, under non-Gaussian data with small samples, We performed two simulations. First, we simulated data using the g-and-h distribution for 10,000 genes under 2 conditions (5 samples in each). In this simulation the scale parameter B was set to 0.53 (the observed average of standard deviations of each gene in the Bottomly dataset). Out of the 10,000 genes we randomly selected 77% to be null, 10% to be differentially expressed at log fold change 0.58 (both up and down regulated), 5% to be differentially expressed at log fold change 1 (both up and down regulated), 3% to be differentially expressed at log fold change 1.32 (both up and down regulated), 2% to be differentially expressed at log fold change 1.58 (both up and down regulated), 2% to be differentially expressed at log fold change 1.81 (both up and down regulated), and 1% to be differentially expressed at log fold change 2. The resulting p-values from the simple t-test were adjusted for multiple testing using the BH [63] method and the nominal FDR was set at 10%. These simulations were performed at $(g = 0, h = 0)$ which corresponds to the standard normal distribution, $(g = -0.11, h = 0.05)$, which corresponds to (L-skew=-0.06, L-kurt=0.15), and $(g = -0.25, h = 0.15)$. which corresponds to (L-skew=-0.14, L-kurt=0.22).

At each g and h we replicated the experiment 100 times. The median observed FDR was approximately 9%, 7%, and 5% for $(g = 0, h = 0)$, $(g = -0.11, h = 0.05)$, and $(g = -0.25, h = 0.15)$ respectively. As expected the observed power for the $(g = 0, h = 0)$ simulation was uniformly better (across the log fold changes) than the $(g = -0.11, h = 0.05)$ simulation, and the observed power for the $(g = -0.11, h = 0.05)$ simulation was uniformly better than the $(g = -0.25, h = 0.15)$ simulation. See Figure 3.6 for a summary of these simulation results. For example at log fold change 1.32 the observed median observed powers for $(g = 0, h = 0)$, $(g = -0.11, h = 0.05)$, and $(g = -0.25, h = 0.15)$ were approximately 70%, 65%, and 45% respectively. The median L-skew estimates observed in the 4 RNA-seq datasets were between -0.04 and 0.00.

In the second simulation we generated RNA-seq counts from the negative binomial distribution. The mean of the counts were computed from taking the average across the DBA/2J samples from the Bottomly dataset. The dispersion parameter was fixed at 0.1. The number of simulated genes, the number of conditions, the number of samples within conditions, and the distribution of null genes and differentially expressed genes were the same as described in the previous paragraph (ie. the g-and-h simulations). Four methods were tested on this dataset: (1) the simple t-test, (2) λ_2-weighted limma model (see explanation below), (3) voom weighted limma model, and (4) the DESeq2 negative binomial model [31]. By λ_2-weighted model we mean using limma with externally obtained weights; where the weights were obtained from the λ_2 estimates for each gene in the simulated dataset after accounting for group specific information. Similar to voom's weights we fit a lowess function between the average log counts and fitted λ_2 estimates. The weight for each individual log count is the inverse of the lowess predicted λ_2 estimate. See Figure 3.5.

For each method the obtained p-values were adjusted for multiple testing using the BH method at nominal FDR 10%. The simulation was performed 100 times to obtain a distribution of observed powers and observed FDR. See Figure 3.7 for a summary of the results. Although the data were simulated from a negative binomial model all the four methods performed sufficiently well with the simple t-test performing least favorably. The λ_2-weighted limma model uniformly out performed the voom-limma model and even outperforms the DESeq2 negative binomial model for log fold change 1.32 and up; while maintaining an observed FDR under 10%.

3.4 Discussion and Conclusions

The SO-plot provides a universal plot for assessing the distributional assumptions of high-throughput genomics data. Given a dataset one can construct the SO-plot and determine whether a t-test based method is appropriate. Based on our simulations and analysis of publicly available datasets we conclude that datasets with absolute median l-skew ($|\tau_3|$) estimates within 0.1 and l-kurt estimates within 0 and 0.2 can be analyzed with t-test based methods. Our simulations based on the negative binomial distribution, a common model for RNA-seq counts, shows that using limma on the log transformed counts with external weights based on l_2 estimates (l_2-limma) provides as much power as negative binomial based methods while controlling the FDR below the specified nominal rate.

The SO-plot (symmetry-outlier) is informative for samples sizes as little as $n \geq 6$. This makes the SO-plot a very powerful tool for exploratory purposes. Using the SO-plot we demonstrate in an indirect manner that log RNA-seq counts have a range of distributional shapes for which t-test based methods are robust (reasonable power and controlled FDR). These results support recent studies that have compared the performance of differential expression methods based on the Gaussian assumption of log RNA-seq counts versus count based methods [36,58,59]. Although RNA-seq data are inherently discrete, and as such some have advocated for modeling the counts directly with discrete distributions such as Poisson or negative binomial; statistics based on log-transformed counts tend to be more stable and robust [51, 64].

Although we analyzed RNA-seq and microarray data other types of high-throughput data can benefit from this kind of analysis. For example in methylation analysis where statisticians have defined complex probability models [65, 66] but t-tests are also commonly in use [67]. It would also be worth it to explore the use of these methods for exploration and analysis of differential variability in gene expression [68].

We advocate the use of the SO-plot because its general and universal. In particular is is unit free and its has finite bounds which makes it easier for comparing different datasets.

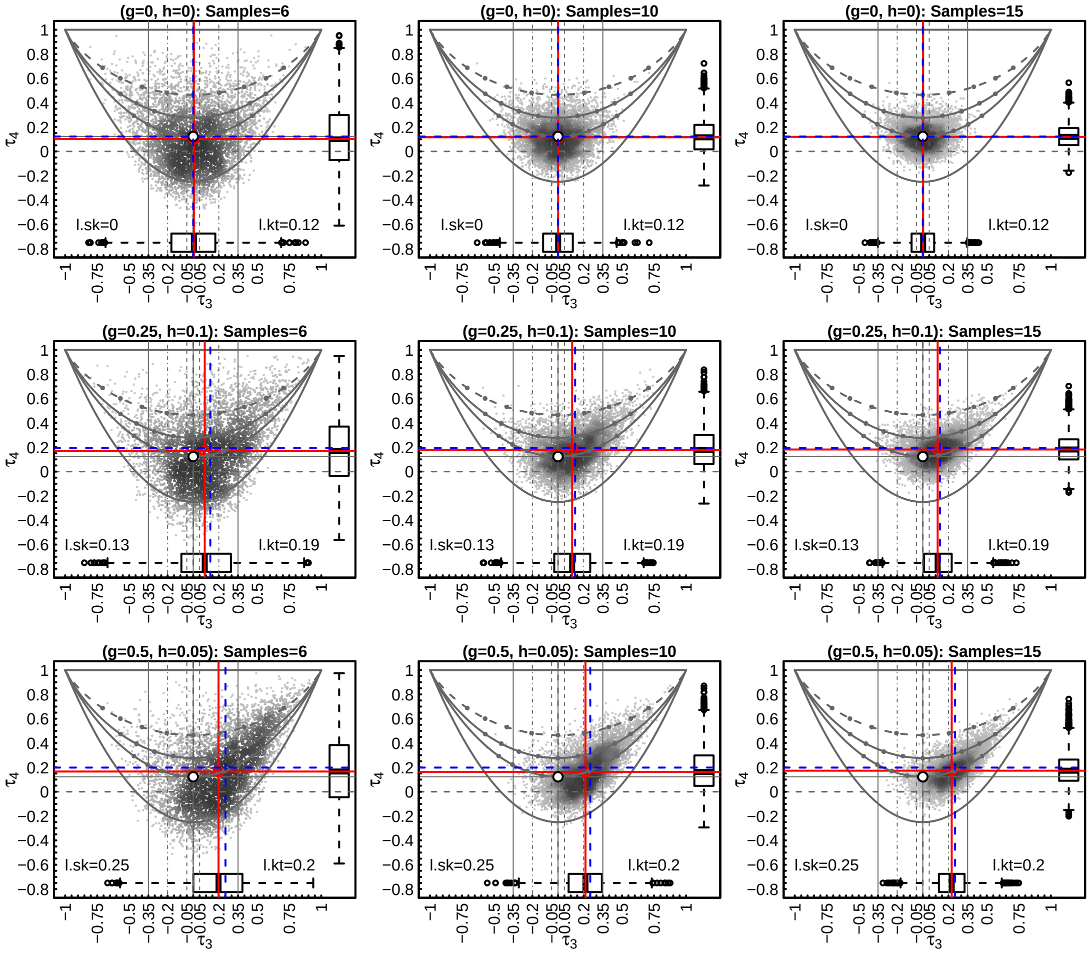
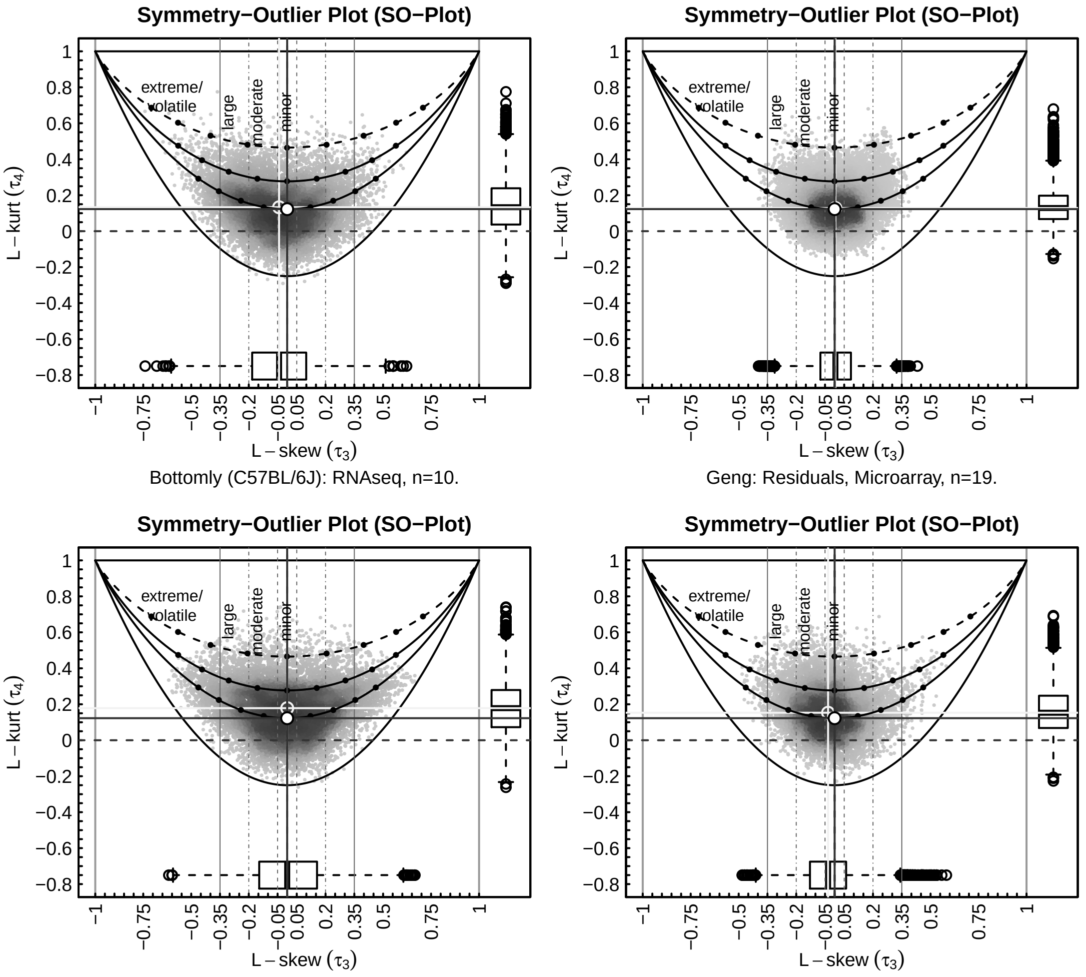


Figure 3.1: **L-skew and L-kurt estimates are informative for small samples.** We have shown the distribution of 5,000 L-skew and L-kurt estimates from data generated from the g-and-h distribution. The blue broken lines are the true parameters and the red lines are the medians of their corresponding estimators. The L-skew and L-kurt estimates provide useful information for samples sizes as little as 6. Observe the consistency in the distribution of the L-skew estimates across the different g and h parameters (ie. across various underlying distributional shapes).



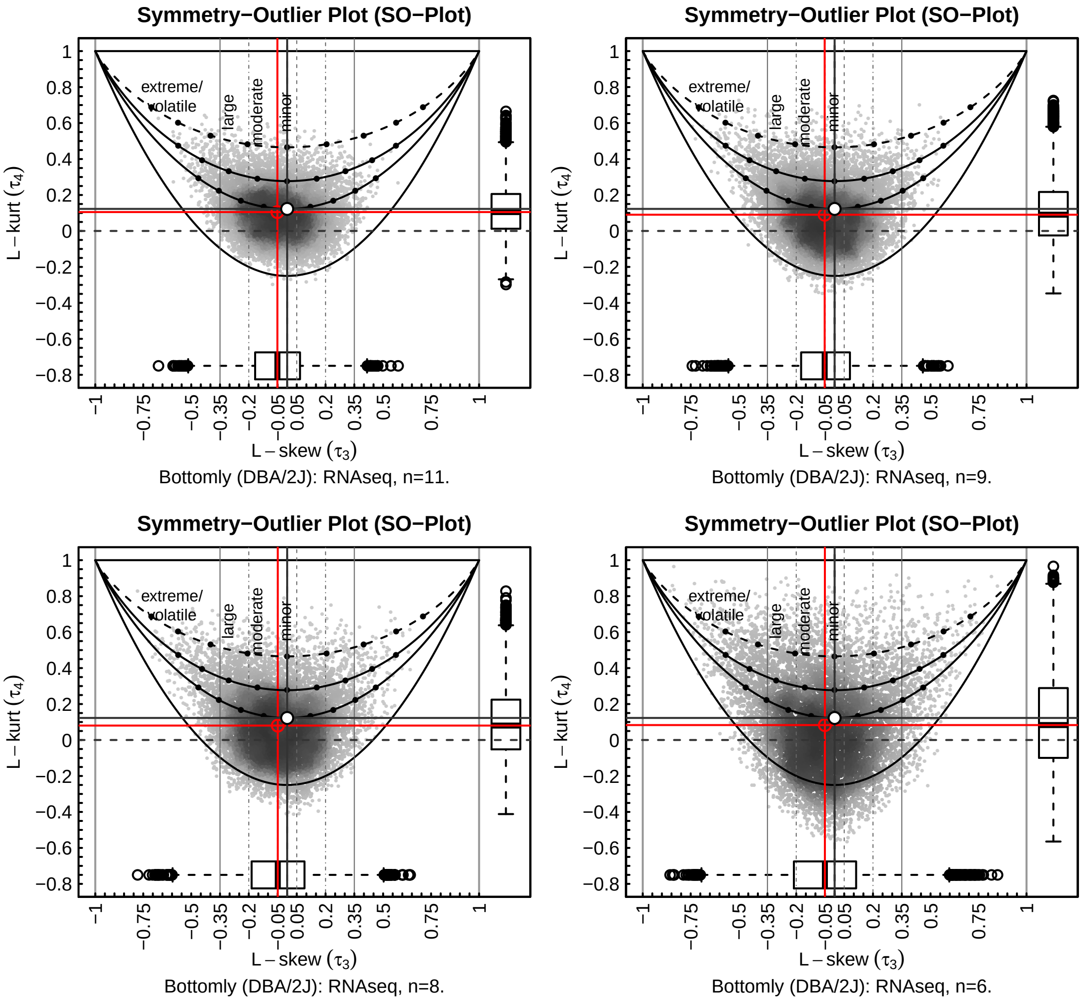
Bottomly (C57BL/6J): RNAseq, n=10.

Geng: Residuals, Microarray, n=19.

Hammoud: Residuals, RNAseq, n=10.

MAQC: Residuals, RNAseq, n=14.

Figure 3.2: **SO-plot examples.** We have described four datasets with the SO-plot. (1) Bottomly (C57BL/6J): a subset of the Bottomly dataset. It contains the 10 samples from the C57BL/6J mouse strain. (2) Geng (microarray): this contains the residuals (19 samples) after the data were adjusted for dosage (ie. group means subtracted). (3) Hammoud: this contains all 10 samples from the Hammoud dataset after adjustment for cell type (residuals). (4) MAQC: this contains all 14 samples of the MAQC dataset after adjusting for RNA source (residuals). The SO-plot provides a means of summarizing the shape (independent of location and scale) of a given high-throughput dataset. The SO-plot is similar to a histogram in the sense that it describes sample shape (regardless of sample size). The more samples we have the clearer the structure of the histogram/SO-plot.



Bottomly (DBA/2J): RNAseq, n=11.

Bottomly (DBA/2J): RNAseq, n=9.

Bottomly (DBA/2J): RNAseq, n=8.

Bottomly (DBA/2J): RNAseq, n=6.

Figure 3.3: **Bottomly (DBA/2J replicates).** In the top left panel we show the SO-plot for the 11 DBA/2J replicates of the Bottomly dataset. The L-skew summary is (-0.17, -0.05, 0.07). The L-skew summary refers to the (25th, 50th, 75th)-quantiles of the L-skew estimates. In the top right we show the SO-plot for a random sample of 9 out of the 11 samples (L-skew summary: (-0.17, -0.05, 0.08)). Similarly, in the bottom row we show the SO-plot for a random sample of 8 and a random sample of 6 from the 11 samples. The L-skew summary for these plots are (-0.19, -0.05, 0.09) and (-0.21, -0.05, 0.11) respectively.

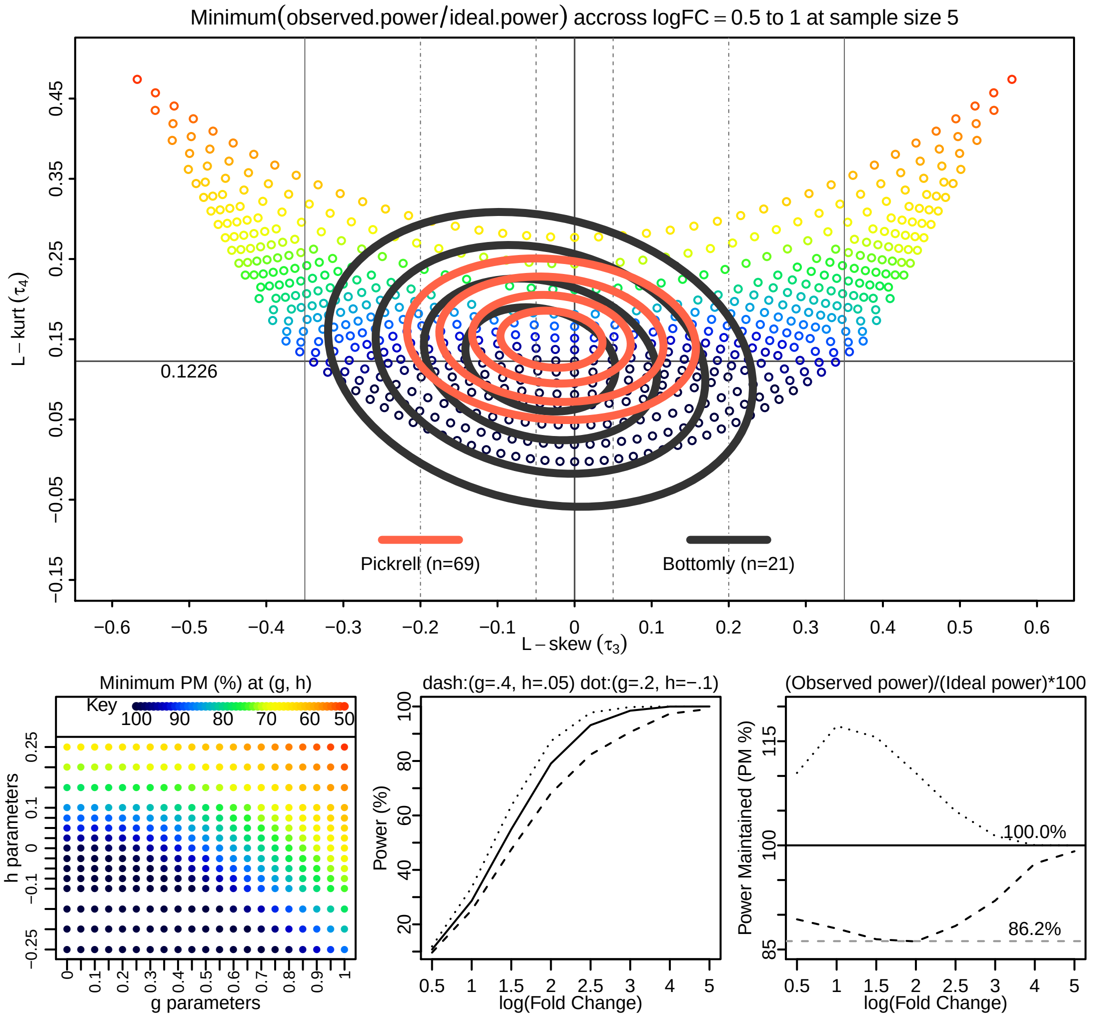


Figure 3.4: **Proportion of power maintained (% PM).** For the following g-and-h parameters, $(g, h) \in \{0 \text{ to } 1\} \times \{-0.25 \text{ to } 0.25\}$ and scale $B = 1$, we performed a simple two (5 samples each) sample t-test at log fold-changes, $logFC = 0.5$, 1, 1.5, 2, 2.5, 3, 4, 5 at level 5%. The observed power at each point was based on 100,000 simulations. In the the bottom left panel we have summarized the simulation results (at each g and h) across $logFC$. For example (bottom middle panel), we compare the observed power at $(g = 0.4, h = 0.05)$ to the ideal power (ie. Gaussian) across $logFC$. The proportion of the power maintained, (observed power / ideal power)×100, is computed in the bottom right panel. To summarize the results at $(g = 0.4, h = 0.05)$ we report the minimum of the power maintained (min-PM) across $logFC$. That is, if you perform a t-test based on data with $(g = 0.4, h = 0.05)$ error distribution (5 samples each and B=1) then you can expect a power of 86.2% or more of the ideal Gaussian power. We have also shown the results at $(g = 0.2, h = -0.1)$. In this case we actually outperform the Gaussian. This is due to the relatively light tail (h=-0.1) and the robustness of the t-test against asymmetry. In the plot above we have shown the min-PM results at the corresponding L-skew $(\tau_3(g, h))$ and L-kurt $(\tau_4(g, h))$ in the SO-plot. We have summarized the Pickrell $(n = 69$, no gender adjustment) and Bottomly $(n = 21$, cell type adjustment) L-skew and L-kurt estimates by fitting them to a bivariate Gaussian and showing the level contours at 25%, 50%, 75% and 90%.

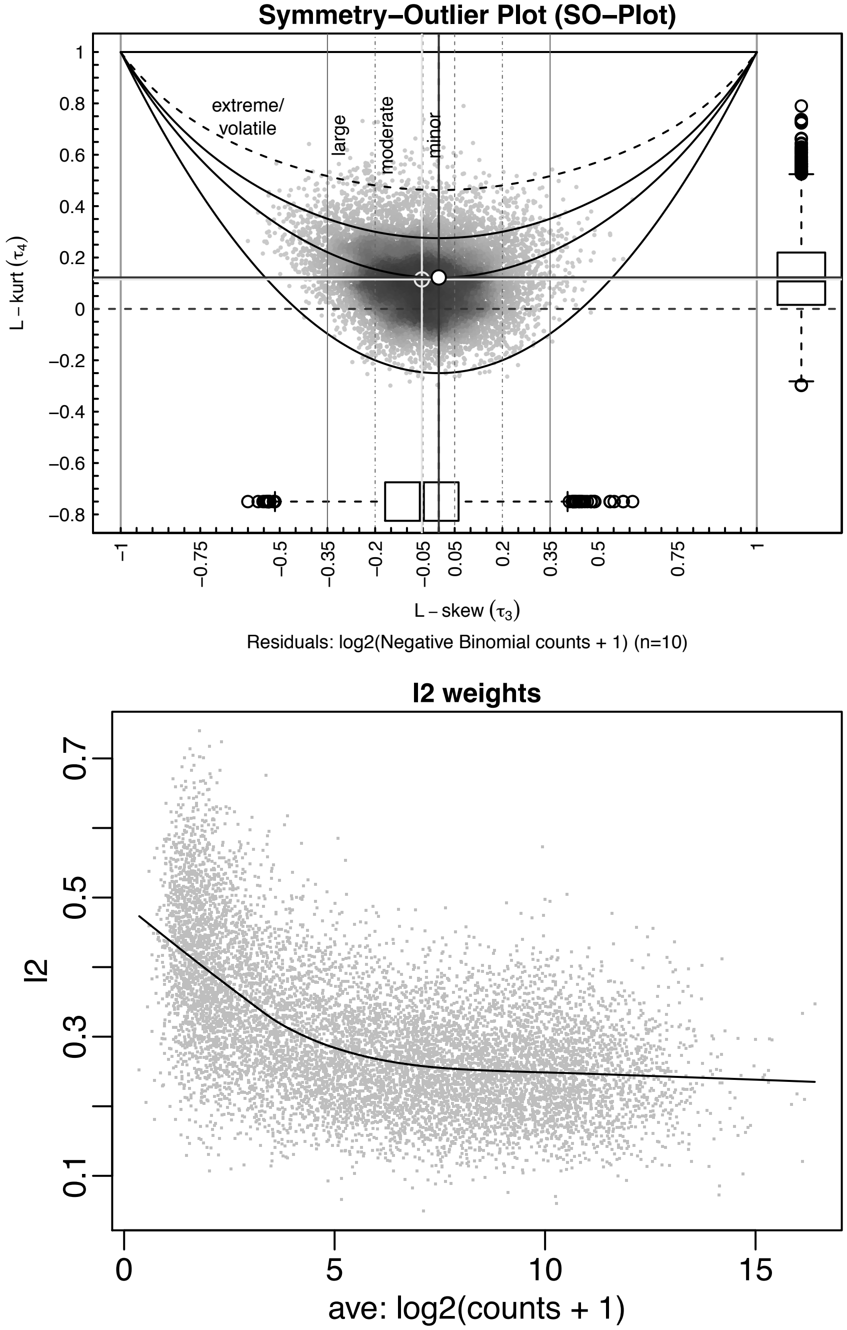


Figure 3.5: **Negative binomial simulation.** The top panel shows the SO-plot for log2 negative binomial counts. In the bottom panel weights are computed for each expression level. The weights are based on the mean-λ_2 trend. λ_2 estimates are highly correlated with standard deviation however they are unbiased and less sensitive to outliers. The computed weights are fed into limma directly.

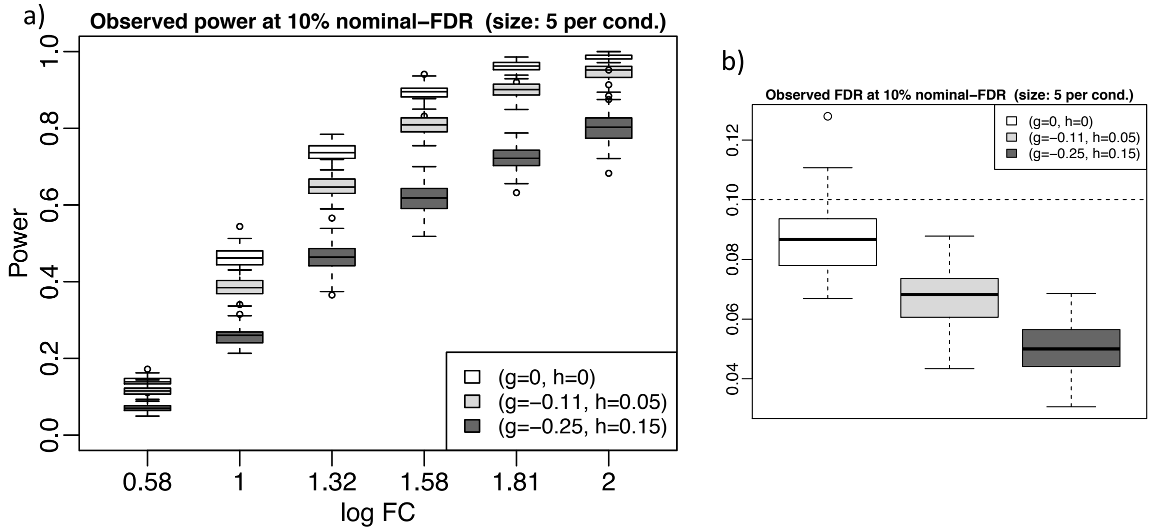


Figure 3.6: **Multiple testing adjusted p-values.** The data were generated from the g-and-h distribution. The scale parameter was set at B=0.53 (based on the average standard deviation observed in the Bottomly dataset). A simple t-test was performed for each of the 10,000 genes in the dataset. 77% of the genes were chosen to be null, the remaining 23% were differentially expressed with logFC ranging from 0.58 to 2. The p-values have been adjusted for multiple testing using the BH method. The nominal FDR rate is 10%.

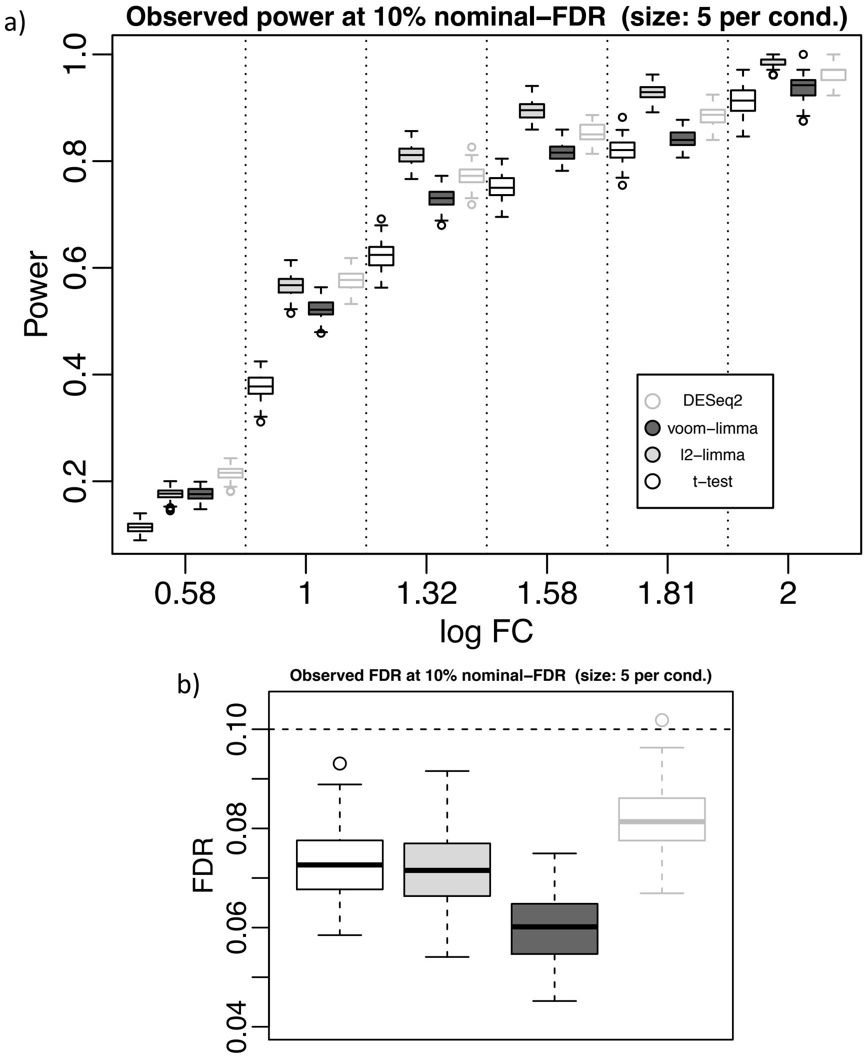


Figure 3.7: **Power at 10 % nominal FDR.** In this figure we compare the results of a simulation study performed to understand the scope of t-test based methods for differential expression. The data were generated according to the negative binomial distribution (ie. DESeq2's assumptions). (a) Observed power at various log fold changes after p-values have been adjusted for multiple testing (BH method). (b) The observed FDR when the nominal is 10%.

Chapter 4: Testing Global Transcriptome Similarity

4.1 Introduction

This chapter describes a method for testing the global transcriptome similarity assumption discussed in Chapter 1 (see 1.4.1). This assumption usually means two things: (1) the number of genes differentially expressed across the biological groups of interest are small or (2) an equivalent number of genes increase and decrease across biological groups [69]. Statistically this may imply that some parameter (e.g. mean, median, 0.75-quantile) of the sample expression distributions are the same across biological groups in the experiment. This is the basis of scaling normalization procedures like the AH [31] method and the median scaling method. Or that on average the sample expression distributions are the same across biological groups. This is the basis of quantile normalization [14].

Hicks and Irizarry [70] have recently proposed a statistical procedure to test the global similarity assumption. Their procedure can be summarized as follows. First, test whether the sample medians within the biological groups are different using a one-way anova test. Second, scale each sample by its median (or by some other scaling method) and test whether the empirical distributions within groups are significantly different. The second test is based on the *quantro* statistic which can be summarized as follows:

$$\mathrm{F}_{quantro} = \frac{\mathrm{SS}_{between}(J-1)}{\mathrm{SS}_{within}(n-J)} \tag{4.1}$$

where

$$\mathrm{SS}_{between} = \sum_{j=1}^{J} n_j \int_0^1 (\bar{\mathbf{Q}}_{.j} - \bar{\mathbf{Q}}_{..})^2 \tag{4.2}$$

and

$$\mathrm{SS}_{within} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \int_0^1 (\mathbf{Q}_{ij} - \bar{\mathbf{Q}}_{.i})^2. \tag{4.3}$$

In the notation above $\mathbf{Q}_{ij} = [x_{(1)ij},\ x_{(2)ij},\ ,\cdots, x_{(G)ij}]^T$ where $x_{(1)ij} \leq x_{(2)ij} \leq \cdots \leq x_{(G)ij}$. That is, $\mathbf{Q}_{ij}$ is the sorted counts vector for sample i under biological condition $j \in (1, 2, \ldots, J)$. The sample group average quantiles are $\bar{\mathbf{Q}}_{.j} = (1/n_j)\sum_{i=1}^{n_i} \mathbf{Q}_{ij}$ and the overall sample quantile average is the average of the group quantiles: $\bar{\mathbf{Q}}_{..} = (1/J)\sum_{j=1}^{J} \bar{\mathbf{Q}}_{.j}$. The notation $\int_0^1 (\bar{\mathbf{Q}}_{.j} - \bar{\mathbf{Q}}_{..})^2$ represents the sum of the squared element wise difference between the quantiles (without loss of generality). To assess significance Hicks and Irizarry [70] obtain a bootstrap null distribution of $\mathrm{F}_{quantro}$ by permuting the group labels B times. The p-value is then obtained as follows:

$$p = (1/B)\sum_{b=1}^{B} I_{[F^b_{quantro} > F_{quantro}]}. \tag{4.4}$$

For more details on the quantro procedure please see Hicks and Irizarry [70]. In this chapter we propose a parametric procedure for testing the same assumption (i.e. are the empirical distributions confounded with biological condition ?) based on L-moments statistics.

4.2 Wilk's shape manova

The idea behind our procedure is very simple. Summarize each sample with its L-ratio shape estimate:

$$\tau_{ij} = [\tau_{3ij} \ \tau_{4ij}]^T \tag{4.5}$$

where τ_{3ij} and τ_{4ij} are the L-skew and L-kurt estimates of sample i under condition j. Based on the fact that τ_{ij} is a linear combination of order statistics generated from samples with very large sample sizes (typically the number of genes G in high-throughput experiment is around 10,000 or more) [39] we assume that:

$$\tau_{ij} \sim \mathrm{N}_2([\tau_{3j}, \tau_{4j}]^T, \Sigma), \quad i \in (1, 2, \ldots, n_j) \tag{4.6}$$

where Σ is a symmetric semi-positive definite matrix. We perform a Wilk's one-way manova (multivariable analysis of variance) on τ_{ij} [71] to test the hypothesis:

$$\mathrm{H}_0 : [\tau_{31}, \tau_{41}]^T = [\tau_{32}, \tau_{42}]^T = \ldots = [\tau_{3J}, \tau_{4J}]^T. \tag{4.7}$$

That is, all the group shapes are the same. The Wilk's test statistic is:

$$\Lambda^* = \frac{\det(\mathbf{W})}{\det(\mathbf{B} + \mathbf{W})} \tag{4.8}$$

where $\mathbf{B} = \sum_{j=1}^{J} n_j(\bar{\tau}_{.j} - \bar{\tau}_{..})(\bar{\tau}_{.j} - \bar{\tau}_{..})^T$ and $\mathbf{W} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (\tau_{ij} - \bar{\tau}_{.j})(\tau_{ij} - \bar{\tau}_{.j})^T$; and $\bar{\tau}_{.j} = (1/n_j) \sum_{i=1}^{n_j} \tau_{ij}$ and $\bar{\tau}_{..} = (1/J) \sum_{j=1}^{J} \bar{\tau}_{.j}$. Under the null hypothesis [71]:

$$\Big(\frac{n-j-1}{j-1}\Big)\Big(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\Big) \sim F_{2(j-1),2(n-j-1)} \tag{4.9}$$

where $F_{2(j-1),2(n-j-1)}$ is the F-distribution with numerator degrees of freedom $2(j-1)$ and denominator degrees of freedom $2(n-j-1)$. Note that in this test we do not need to first adjust the samples for scale (as was done in quantro) since the L-ratio statistics are independent of scale.

4.3 Wilk's test statistic is robust

In order to check the distributional behavior of the Wilk's Fstat (equation 4.9) under the null distribution. We performed a bootstrap experiment using the Pickrell [55] dataset. The Pickrell dataset is part of the International HapMap Project. RNA samples were extracted from the lymphoblastoid cell lines of 69 unrelated Nigerian individuals, 29 males and 40 females. The aligned and summarized count matrix was obtained from the bioconductor (http://www.bioconductor.org) tweeDEseqCountData package. Genes with at least one count per million (cpm) in 29 or more samples were kept.

To assess the null distribution of Wilk's $Fstat$ (equation 4.9) we sampled 20 out of the 40 (without replacement) female samples and randomly assigned them into two groups (10 each) and computed the Wilk's $Fstat$. We performed this experiment 500 times in order to obtain a null distribution of $Fstat$. The results of this simulation experiment are summarized in Figure 4.1. The distribution of the p-values (Figure 4.1 (d)) is fairly random (uniformly distributed). Indicating that the Wilk's $Fstat$ has an F-distribution under the null hypothesis.

4.4 Wilk's shape manova and quantro results are consistent

We used quantro and Wilk's shape manova to test the global transcriptome similarity assumption for 4 RNA-seq datasets; the Pickrell dataset [55], Bottomly dataset [60], the SEQC dataset (samples A and B only) [62], and the Zebrafish dataset [29]. All four datasets consists of two biological groups. Please see appendix A for the details on these datasets. We set $B = 1000$ for the quantro permutation test. The p-values were (0.151, 0.121), (0.471, 0.417), (0, 0), (0.03, 0) for Pickrell, Bottomly, SEQC, and Zebrafish data respectively; where $(p_1, p_2) =$ (Wilk's p-value, quantro p-value). As we can see the results are consistent. Please see Figures 4.2, 4.3, and 4.4 for a graphical inspection. The ERCC spike-in mix 1 was added to the Zebrafish samples during library preparation. We tested the ERCC spike-in for global similarity as well; quantro reported a p-value of 0.096, while Wilk's test reported a p-value of 0.537 (see Figure 4.4 (c)). A visual inspection suggests that the the test should not be rejected. Finally we assessed global similarity of the data generated from DNA microarrays. The p-values for these test were similar for both quantro and Wilk's shape manova test. See Figure 4.5 for a summary of these results.

4.5 Shape Plot Reveals Library Preparation Bias

In order to explain the results in this section we will need more information about the SEQC dataset. Below is a brief description. The SEQC dataset is part of the Microarray Quality Control (MAQC) project [62]. The aims of the project is to assess the technical performance of high-throughput genomics technology, including RNA-seq. The dataset contains four tissue types: A, B, C, and D. Tissue A is Stratagene's universal human reference RNA and tissue B is Ambion's human brain reference RNA. Tissues C and D are mixtures of A and B in the ratios of 3:1 and 1:3 respectively. For each of the total RNA from the four tissue samples, 4 libraries were constructed. Making 4 technical replicates per tissue type. Each of these 4 technical replicates were divided into 2 parts (to be sequenced on two flow-cells).

Each of the 2 parts were further separated into 8 parts (to be sequenced in each of the 8 lanes in a flow-cell). The ERCC spike-in mix 1 was spiked into the total RNA of tissue A and mix 2 was spiked into the total RNA of tissue B prior to the creation of tissue C and tissue D. The results in this section are based on library preparation differences. In summary each of the 4 different library preparations for each of the 4 tissue types were split into 16 parts.

We summarized the raw counts of the SEQC dataset by the shape (L-skew, L-kurt) estimate of each sample. This is displayed on Figure 4.6. We see that the four different tissue types clearly cluster into 4 groups. At this point we are not able to ascertain whether these differences in shape (independent of location and scale) are due to biology or technology. However for tissue types A and C there is a clear formation of two clusters within each tissue type. In Figure 4.7, we show each shape plot of the 4 tissue types separately; and color each sample point by its library type. In all the four tissue specific shape plots we can see that the samples cluster by library type. The shape plot is able to detect structural differences in the data that may otherwise not be possible. In Figure 4.8, we show the four tissue type specific plots again, however this type the data are on log scale. Again the clustering by library type appear. However it not is not as strong as the case when the data are counts. See Figures 4.9, 4.10, and 4.11 for similar plots using the ERCC spike-in counts.

4.6 Discussion and Conclusions

We have introduced the Wilk's shape manova procedure for checking whether the empirical assumptions sample distributions of RNA-seq data are the same. We have also shown that the assumptions of the test are reasonable under the Pickrell RNA-seq dataset. The shape plot is a simple and convenient tool for assessing distributional assumptions about RNA-seq data; both gene-wise (SO-plot) and sample-wise (shape plot). It is both sensitive and accurate; and its statistics are computationally easy to compute. Unlike principal component analysis that require a mixing of the samples in order to show possible hidden structures; the shape plot is sample specific. Each point depends only of the sample it represents. In addition the shape plot is scale invariant while principal components analysis plots are not. We think of the shape plot as another addition to the many tools available for exploring and analyzing high-throughput genomics datasets.

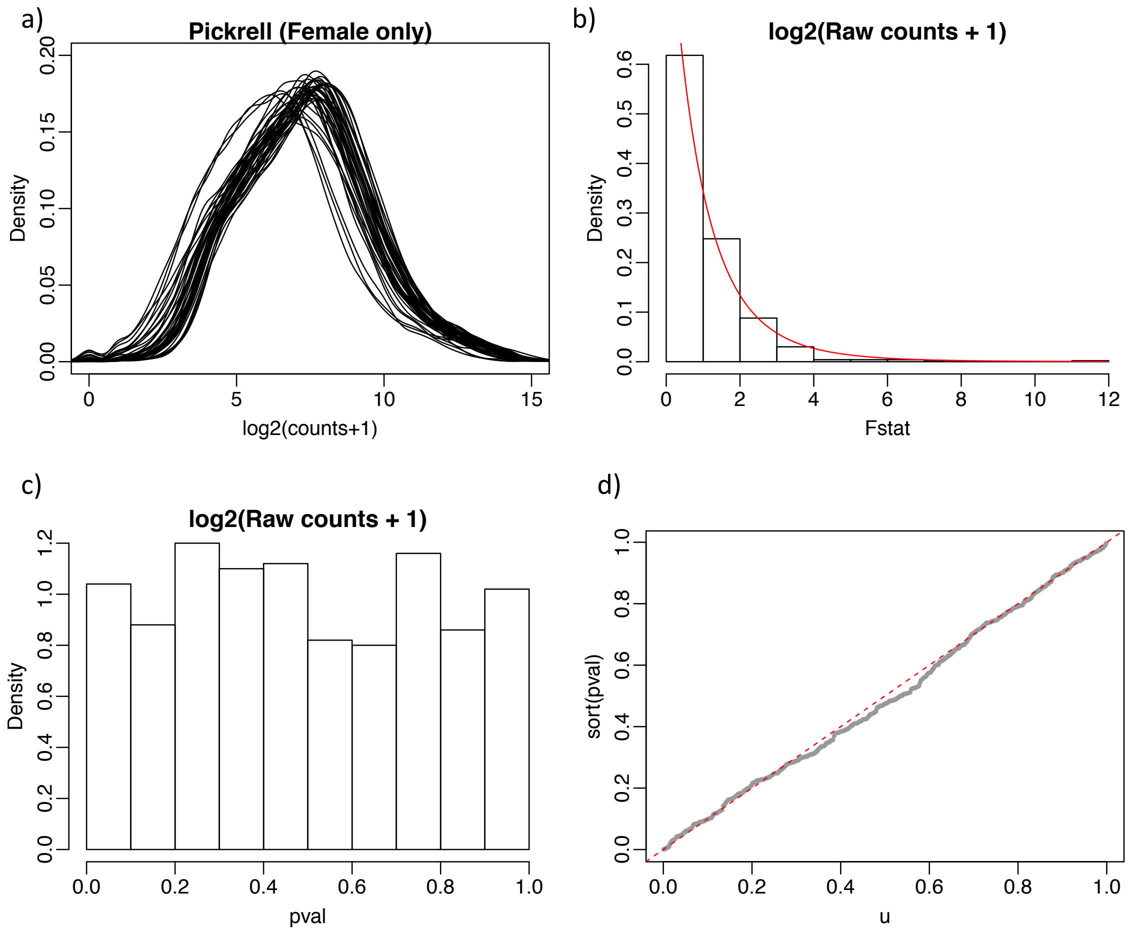


Figure 4.1: **Wilk's shape manova is robust for RNA-seq Data.** (a) Density plots of log2 transformed raw counts (+1). These 40 female samples were subsampled 500 times. Each sample selected 20 females out the 40 without replacement. The 20 samples were randomly assigned to two groups (10 in each) group. Wilk's Fstat was computed each time to generate a null distribution. (b) The null distribution of Wilk's Fstat. The red curve indicates the density of the corresponding *F*-distribution. (c) The distribution of p-values for each observed Fstat. As we can see this distribution is fairly uniform. Indicating that our Gaussian assumptions of the shape estimates are reasonable. (d) The Q-Q plot of the observed p-values.

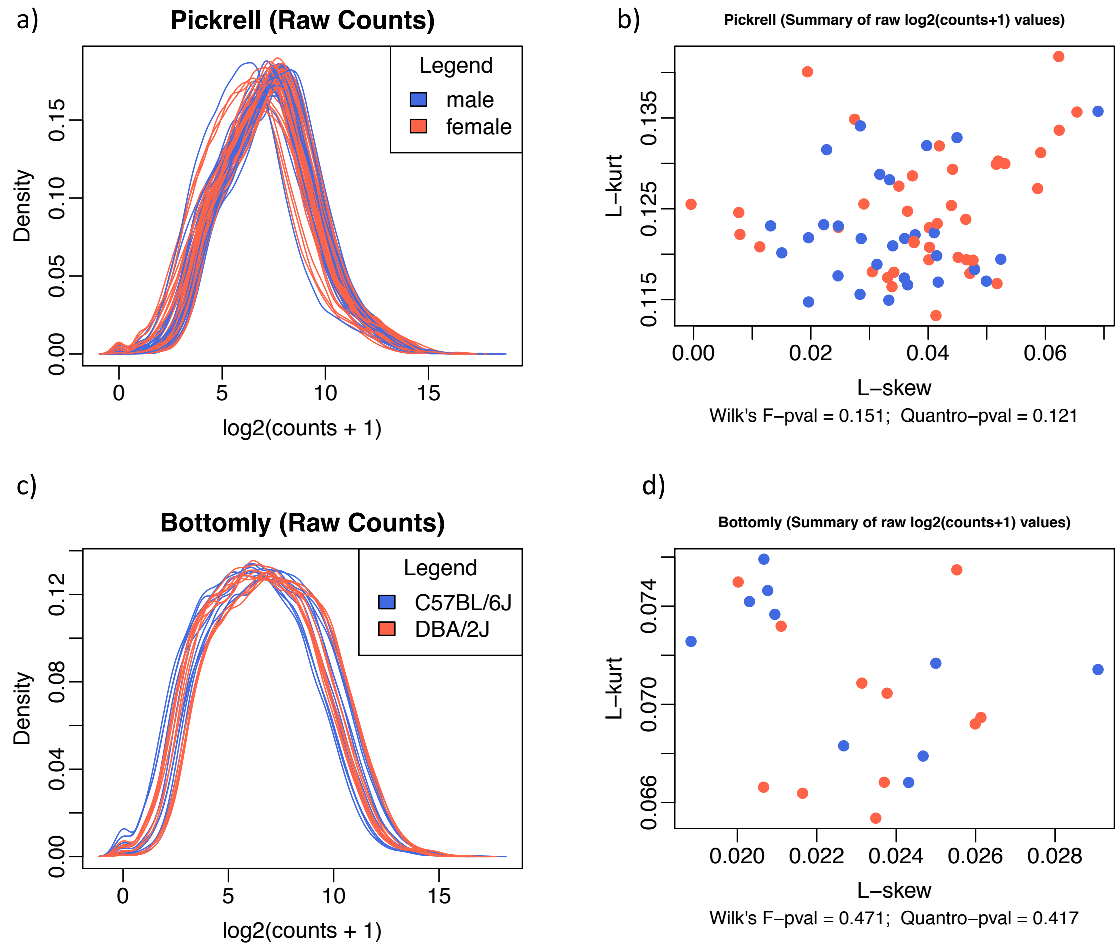


Figure 4.2: **Test for global transctiptome similarity: Pickrell and Bottomly Datasets.** Results of the shape manova and quantro test for global transcriptome similarity after accounting for scale. quantro was performed using $(B = 1000)$ permutations.

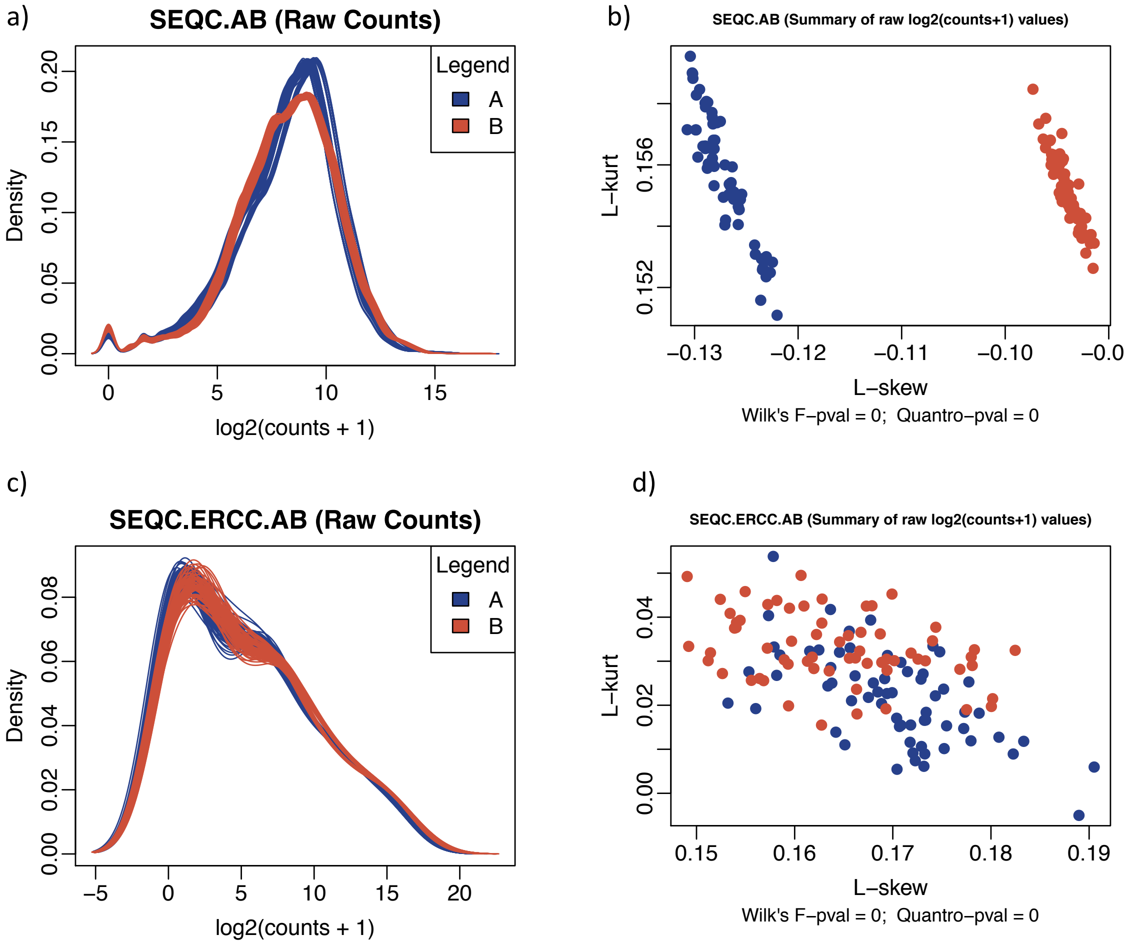


Figure 4.3: **Test for Global Transctiptome Similarity: SEQC Samples A and B only.** Results of the shape manova and quantro test for global transcriptome similarity after accounting for scale. quantro was performed using ($B = 1000$) permutations.

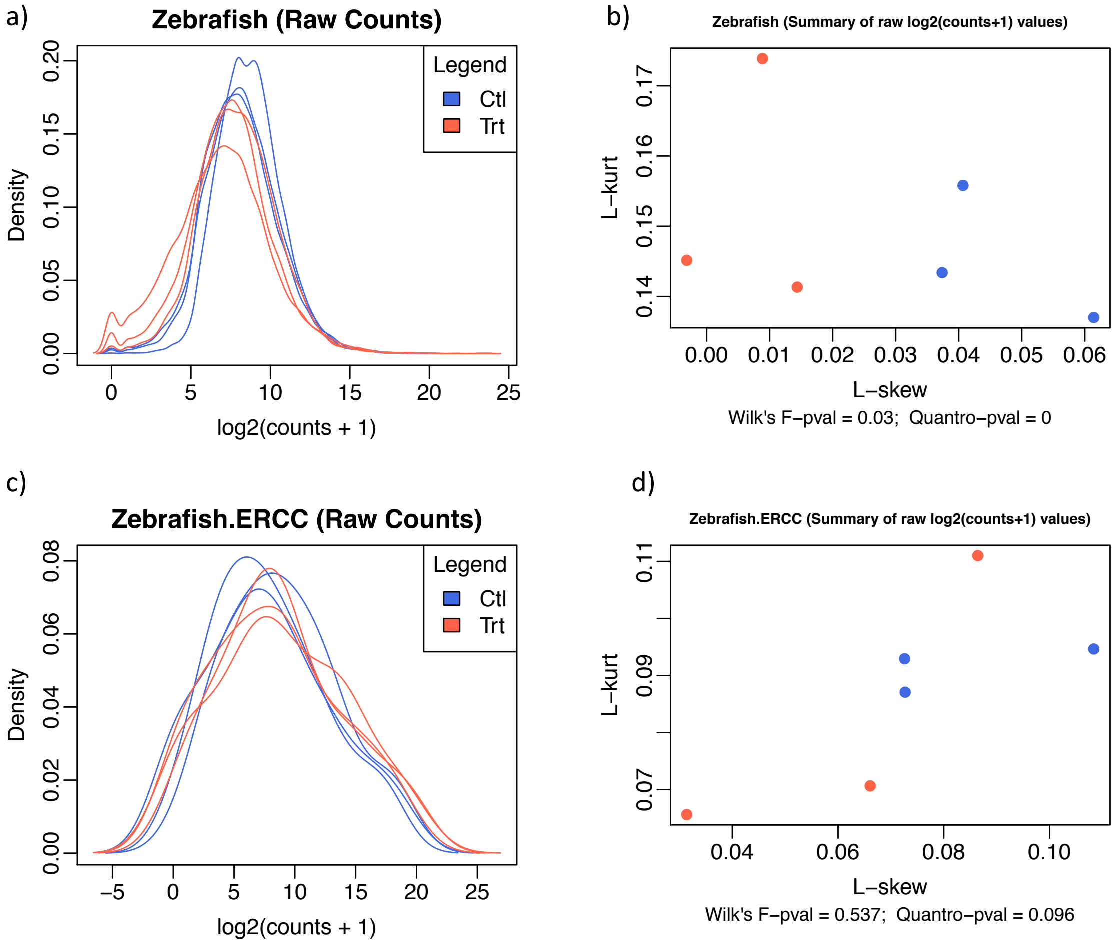


Figure 4.4: **Test for Global Transctiptome Similarity: ERCC Samples A and B only.** Results of the shape manova and quantro test for global transcriptome similarity after accounting for scale. quantro was performed using ($B = 1000$) permutations.

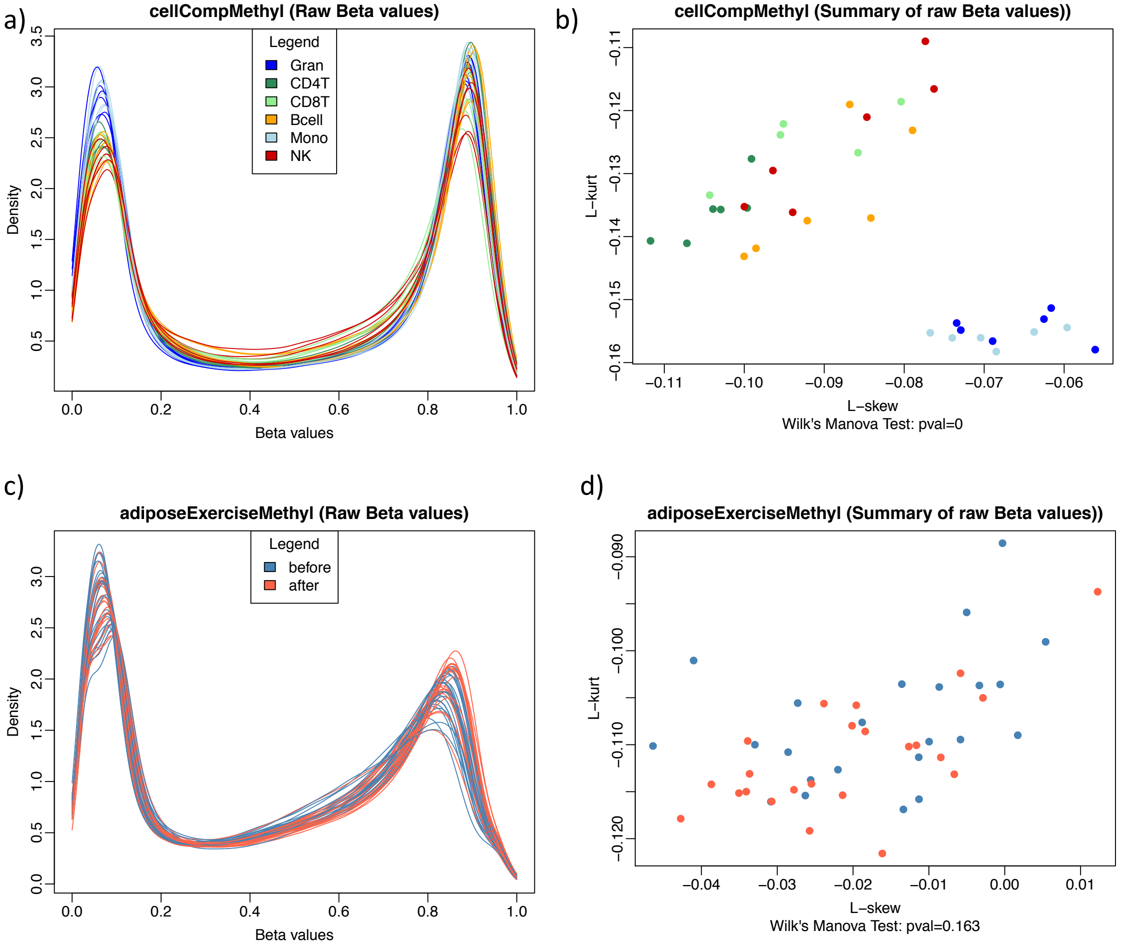


Figure 4.5: **Test for Global Methylome Similarity: DNA Methylation Array Data.** Results of the shape manova test for global methylome similarity after independent for scale.

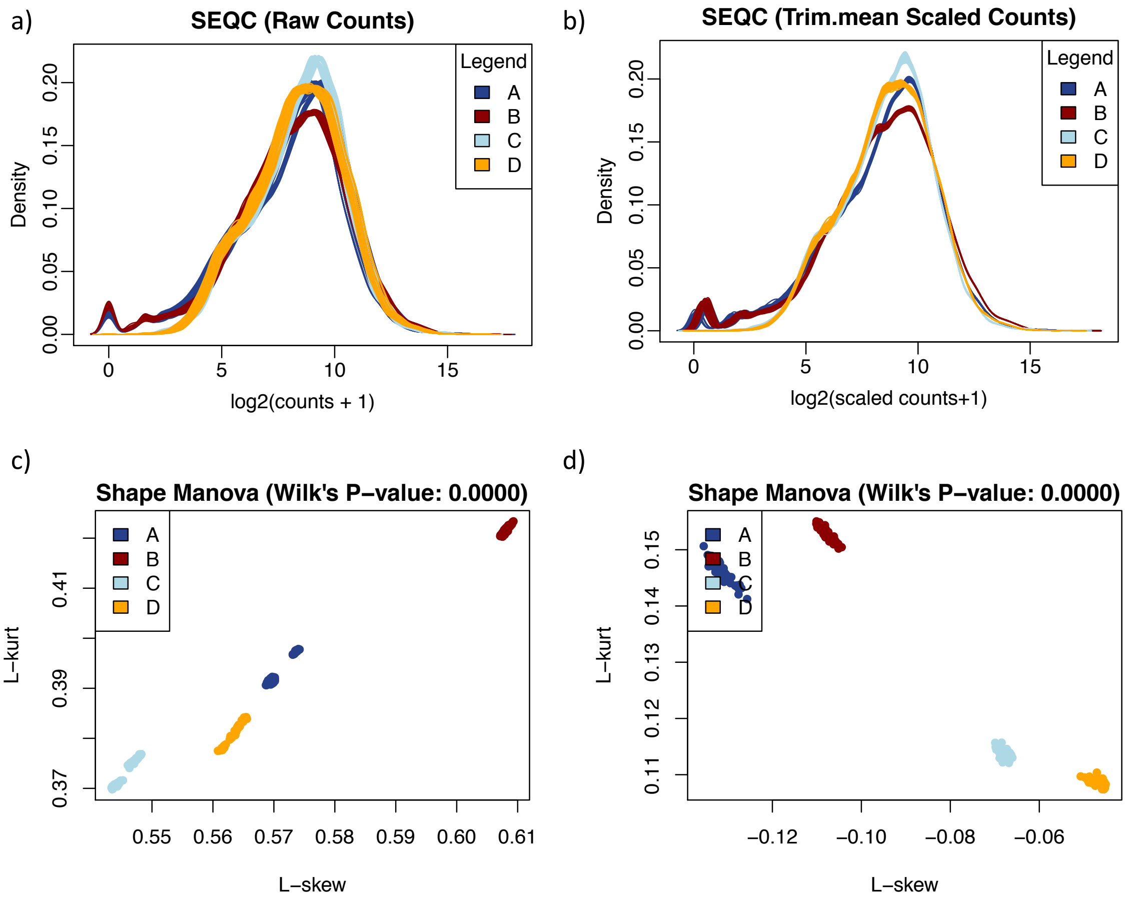


Figure 4.6: **Shape manova works on counts and log counts.** Results of the shape manova test for global methylome similarity after independent for scale. (a) Density plots of the log transformed raw counts. (b) Density plots of the log transformed trimmed mean scaled counts (trim=0.25). (c) Shape plot summarizing samples on the counts scale. The count densities are not shown in this figure. Note that sample A and sample C cluster into two groups. See Figure 4.7 for a more detailed view. (d) Shape plot summarizing the density plots in both (a) and (b). (Shape plots are the same regardless of the scaling in (b)).

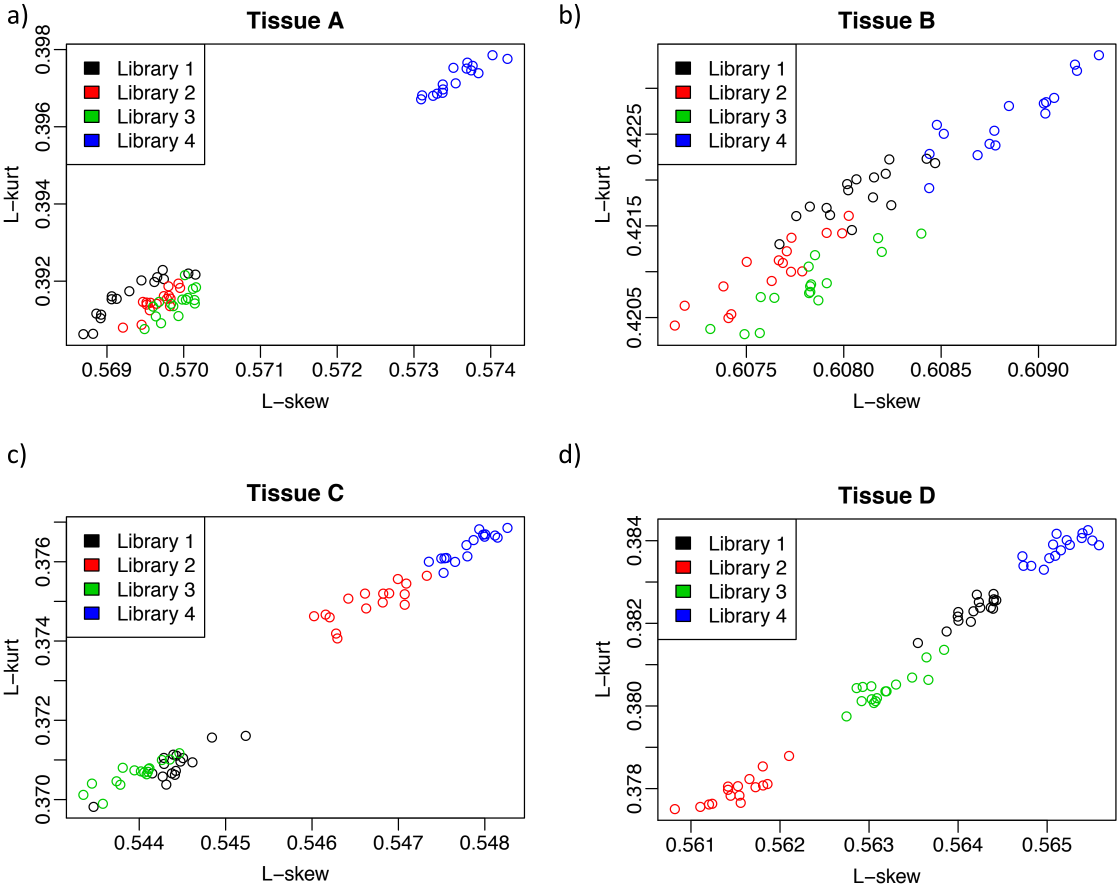


Figure 4.7: **Shape plot shows library preparation bias on counts scale.** Above we summarize each of the technical replicates in each of the 4 tissue types by the shape plot. We indicate the each of the 4 distinct library constructions by color. We can see that the samples group by library. Each of the L-ratio estimates we constructed from the raw counts.

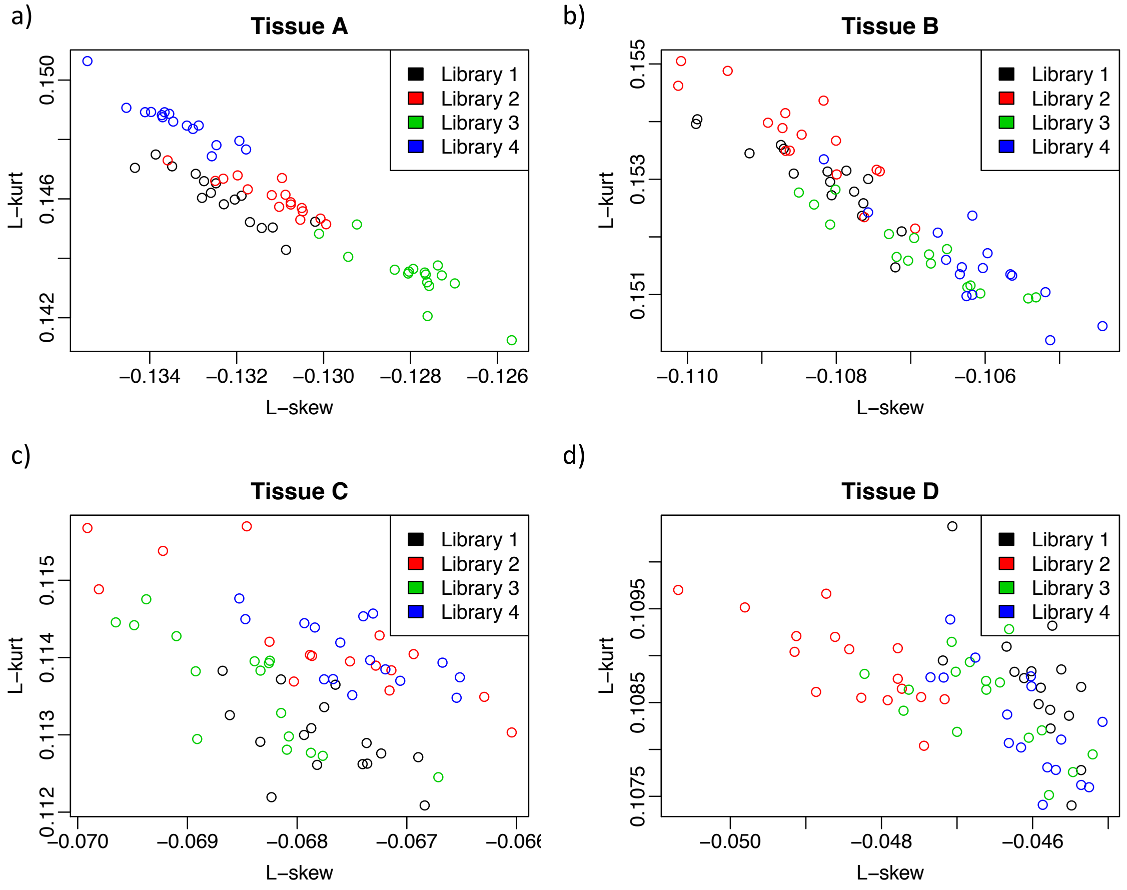


Figure 4.8: **Shape plot shows library preparation bias on log-counts scale.** Above we summarize each of the technical replicates in each of the 4 tissue types by the shape plot. We indicate the each of the 4 distinct library constructions by color. We can see that the samples group by library. Each of the L-ratio estimates we constructed from log of the raw counts (+1).

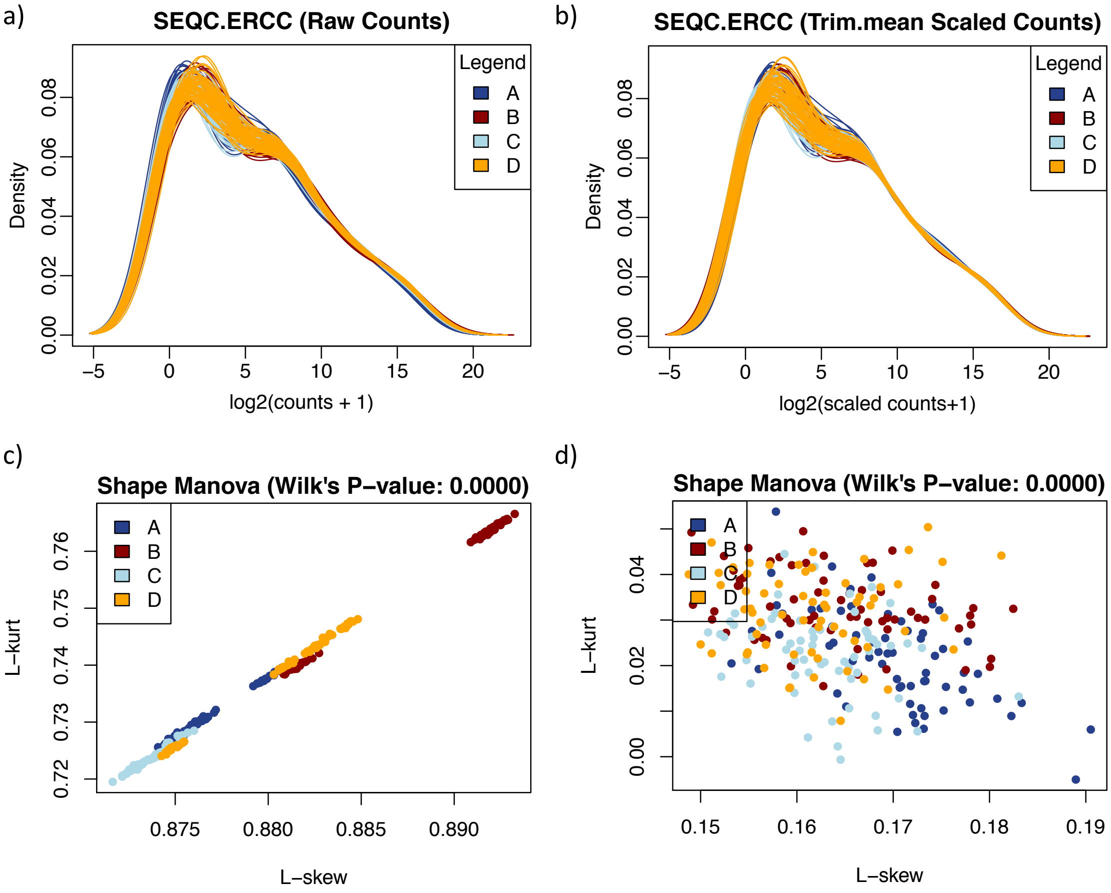


Figure 4.9: **SEQC ERCC spike-in shape plot.** Shape plots for the SEQC ERCC spike-in dataset. (a) Density plots of log raw counts (+1). (b) Density plots of log trimmed mean scaled counts (trim = 0.25).

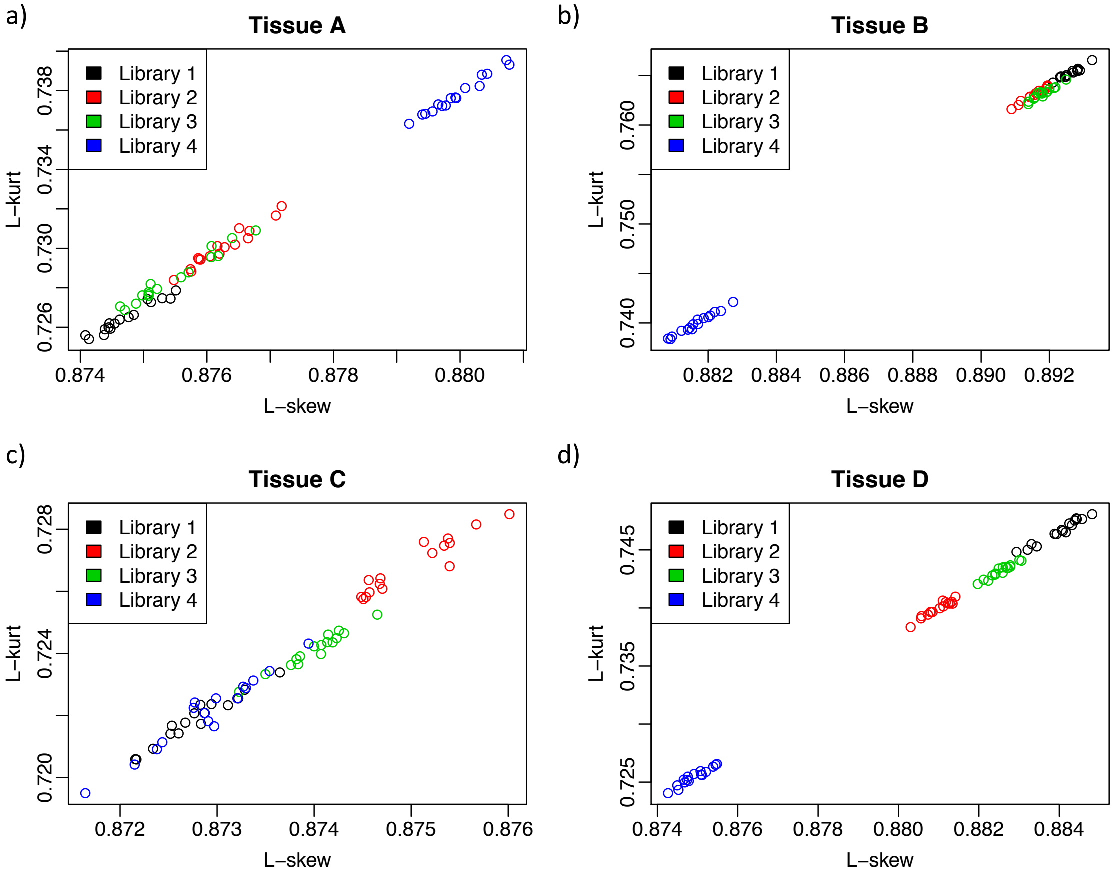


Figure 4.10: **Shape plot shows library preparation bias on counts scale (Spike-in).** Above we summarize each of the technical replicates in each of the 4 tissue types by the shape plot. We indicate the each of the 4 distinct library constructions by color. We can see that the samples group by library. Each of the L-ratio estimates we constructed from the raw counts.

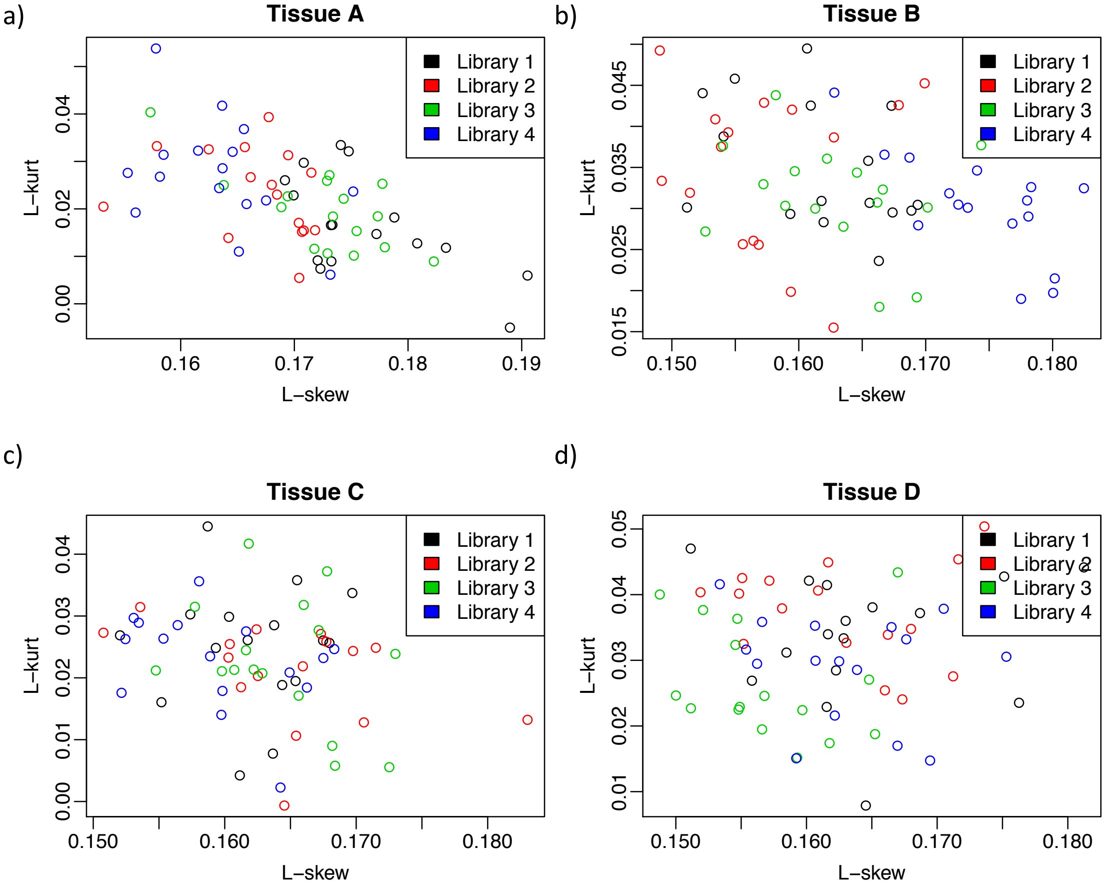


Figure 4.11: **Shape plot shows library preparation bias on log-counts scale (Spike-in).** Above we summarize each of the technical replicates in each of the 4 tissue types by the shape plot. We indicate the each of the 4 distinct library constructions by color. We can see that the samples group by library. Each of the L-ratio estimates we constructed from log of the raw counts (+1).

Chapter 5: Smooth Quantile Normalization

5.1 Introduction

Normalization is a key step in the analysis of high-throughput genomics data including RNA-seq. The normalization procedure chosen can have a huge impact on the list of differentially expressed genes and down stream analysis if not chosen properly. Normalization procedures are based on assumptions about the data generation process and as such it is important to understand the assumptions that they make and how robust they are with respect to deviations from these assumptions. In the previous chapter we introduced shape manova, a statistical procedure for testing where sample distributions within groups differ in as statistically significant way. Scaling normalization methods such as the trimmed mean try to make the within sample means equal for all samples. Quantile normalization [14] on the other hand makes the some what stronger assumption that all the quantiles within each sample (sample global distribution) should be the same across biological conditions.

In this chapter we introduce, smooth quantile normalization (qsmooth), a modification of quantile normalization that makes the assumption that: all samples within the same biological group should have the same shape. qsmooth first performs quantile normalization within each biological group and then shrinks the group quantiles towards the overall reference quantile depending on the variation between the group quantiles and the variation of quantiles within the groups.

5.2 Smooth Quantile Normalization

5.2.1 Motivation

The idea behind qsmooth is a simple generalization of quantile normalization. Let $\mathbf{Q}$ denote the log of the sorted counts. We add 1 to each count before the log transformation in order to avoid taking logs at 0. Suppose that we perform a regression on each row of $\mathbf{Q}$ using the biological groups as the design matrix. Then the group estimates will just be the group averages. If we do this for every row then we end up with group specific quantiles for each sample. This is equivalent to performing quantile normalization within groups. Performing quantile normalization across all the samples is equivalent to performing a regression without using the group specific information. Qsmooth is weighted average of these two extremes (full quantile normalization and between groups quantile normalization). The weight of the average is chosen to depend on the variance between the groups and the variance within the groups.

5.2.2 Algorithm

Let x_{gij} represent the total number of reads aligned to gene g in sample i under biological condition j. Let $g \in (1, 2, \cdots, G)$; $j \in (1, 2, \cdots, J)$; and $i \in (1, 2, \cdots, n_j)$ where $n = \sum_{j=1}^{J} n_j$. G is the number of genes and n_j is the number of samples in group j. We will denote a sample as the column vector $\mathbf{X}_i = [x_{1i},\ x_{2i},\ \cdots,\ x_{Gi}]^T$. The entire count matrix will be denoted as $\mathbf{X} = [\mathbf{X}_1,\ \mathbf{X}_2,\ \cdots,\ \mathbf{X}_n]$. Denote y_{gij} as the log of the scale normalized version of x_{gij}. Sort each sample, $\mathbf{Y}_i =$

$[y_{1ij},\ y_{2ij},\ \cdots,\ y_{Gij}]^T$, to obtain:

$$\mathbf{Q}_{ij} = [y_{(1)ij},\ y_{(2)ij},\ \cdots,\ y_{(G)ij}]^T \tag{5.1}$$

where $y_{(1)ij} \leq y_{(2)ij} \leq \ldots \leq y_{(G)ij}$. Compute the group reference quantiles:

$$\bar{\mathbf{Q}}_{.j} = (1/n_j) \sum_{i=1}^{n_j} \mathbf{Q}_{ij} = [\bar{y}_{(1).j},\ \bar{y}_{(2).j},\ \cdots,\ \bar{y}_{(G).j}]^T. \tag{5.2}$$

Compute the overall reference quantile:

$$\bar{\mathbf{Q}}_{..} = (1/J) \sum_{j=1}^{J} \bar{\mathbf{Q}}_{.j} = [\bar{y}_{(1)..},\ \bar{y}_{(2)..},\ \cdots,\ \bar{y}_{(G)..}]^T. \tag{5.3}$$

Compute $\sigma^2 = [\sigma^2_{(1)},\ \sigma^2_{(2)},\ \cdots,\ \sigma^2_{(G)}]^T$:

$$\sigma^2 = (1/n) \sum_{j=1}^{J} n_j \mathbf{S}^2_{.j} \tag{5.4}$$

where $\mathbf{S}^2_{.j} = (1 - n_j)^{-1} \sum_{i=1}^{n_j} (\mathbf{Q}_{ij} - \bar{\mathbf{Q}}_{.j})^2$

Compute $\tau^2 = [\tau^2_{(1)},\ \tau^2_{(2)},\ \cdots,\ \tau^2_{(G)}]^T$:

$$\tau^2 = (J - 1)^{-1} \sum_{j=1}^{J} (\bar{\mathbf{Q}}_{.j} - \bar{\mathbf{Q}}_{..})^2. \tag{5.5}$$

Compute individual quantile reference weights:

$$\nu = \sigma^2 / (\sigma^2 + \tau^2) \tag{5.6}$$

where division is performed component-wise.

Finally smooth weights:

$$\begin{aligned} \nu^* &= \text{RollingMedian}\{\nu_{(1)},\ \nu_{(2)},\ ,\ldots, \nu_{(G)} |\ w_0\} && (5.7) \\ &= [\nu^*_{(1)},\ \nu^*_{(2)},\ ,\ldots, \nu^*_{(G)}] && (5.8) \end{aligned}$$

where w_0 is the length of the window. The default is $w_0 = 99$.

The shrunken quantiles are:

$$\begin{aligned} \mathbf{Q}^*_{ij} &= [\nu^*_{(1)}\bar{y}_{(1)..},\ \nu^*_{(2)}\bar{y}_{(2)..},\ \cdots,\ \nu^*_{(G)}\bar{y}_{(G)..}]^T && (5.9) \\ &= [(1-\nu^*_{(1)})\bar{y}_{(1).j},\ (1-\nu^*_{(2)})\bar{y}_{(2).j},\ \cdots,\ (1-\nu^*_{(G)})\bar{y}_{(G).j}]^T && (5.10) \end{aligned}$$

The normalized are expression matrix is obtained by ordering each $\mathbf{Q}^*_{ij}$ according to the original order.

5.3 Qsmooth adapts to data

We applied the qsmooth algorithm to the Bottomly and Pickrell datasets. See Figures 5.5 and 5.4. We observed that the results were similar to quantile normalization, as depicted by the density plots. For the Bottomoly dataset we can reasonably assume that biologically the transcriptomes are similar across biological conditions and as such quantile normalization can be performed. This is because the biological groups consist of a wild-type and a mutant and as such we only expect a few genes to be differentially expressed. The Pickrell dataset consists of samples taken from the somatic cells of randomly selected individuals. Hence we can also assume that the transcriptomes are similar across gender; justifying the use of the quantile normalization procedure here as well.

5.4 Functions of up-regulated expressed genes

Differential expression analysis was performed on the *T.cruzi* dataset [72]. The criteria for selecting differentially expressed genes were an absolute log2 fold-change greater than 1 and an FDR less than 0.05. The differential expression analysis was performed for the 0.25-trimmed mean scaled data, the qsmooth normalized data, and the quantile normalized data, using the λ_2-limma model (see Section 3.3). See Figures 5.9 and 5.10. The list of differentially expressed genes were organized into 2 groups of genes. Genes that were up-regulated in the trypomastigote (infective form) stage as compared to the epimastigote stage, and genes that were down-regulated in the trypomastigote stage as compared to the epimastigote stage.

To assess the list of differentially expressed genes obtained from the 3 normalization procedures we performed a gene function enrichment analysis via a hypergeometric test. We tested for the enrichment of gene function in the set of genes that were differentially up-regulated in the trypomastigote stage in all the 3 normalization methods (1553 genes). Some of the functions that were found to be significant ($\mathrm{FDR} < 0.05$) are surface protease (GP63) and mucin-associated surface protein (MASP). These genes have been shown to play an important role in the lifecycle of *T.cruzi*, especially when it is in the infective stage [73]. We also tested for the enrichment of gene function in the set of genes that were differentially up-regulated in the trypomastigote stage in only the qsmooth and 0.25-trimmed mean normalization methods (474 genes). Again, GP63 and MASP came up as significant. We did not find GP63 and MASP to be significant when the set of differentially expressed genes (87 genes) obtained only in the quantile normalization method was used for function enrichment testing.

5.5 Discussion

We have presented an algorithm for normalization that is a modification of quantile normalization. This algorithm implicitly tests the global similarity assumption at each quantile and decides how much to shrink towards the quantile reference. We have implemented this algorithm on a few publicly available dataset and find that it performs similarly to the full quantile normalization in datasets where the assumption of global similarity is reasonable. These include the Bottomly dataset and the Pickrell dataset (see Figures 5.5 and 5.4).

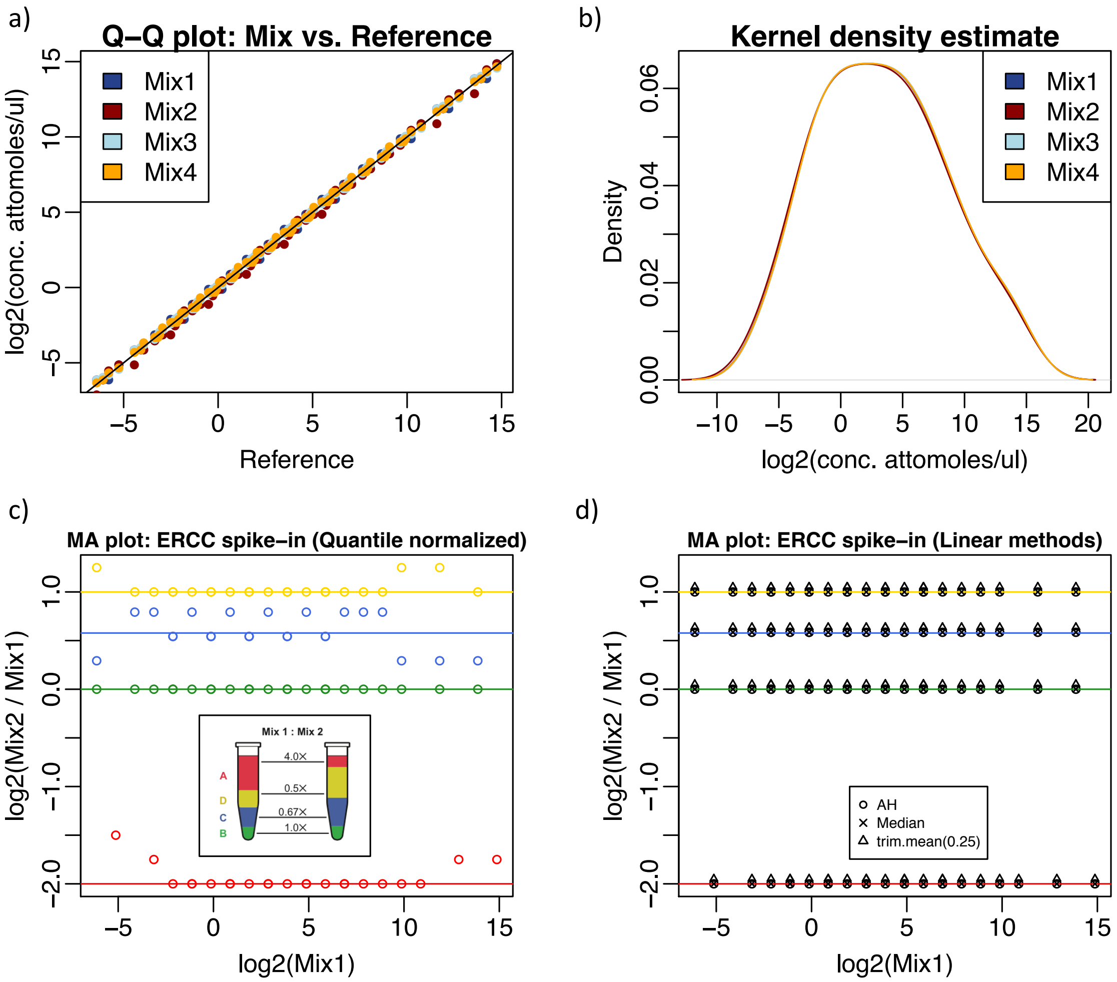


Figure 5.1: **Known normalization bias.** (a) Q-Q plot for ERCC spike-in mix using true concentrations. (b) Density estimates of ERCC spike-in using true concentrations. (c) MA-plot of the quantile normalized true concentration. (d) MA-plot of the scale normalized true concentration.

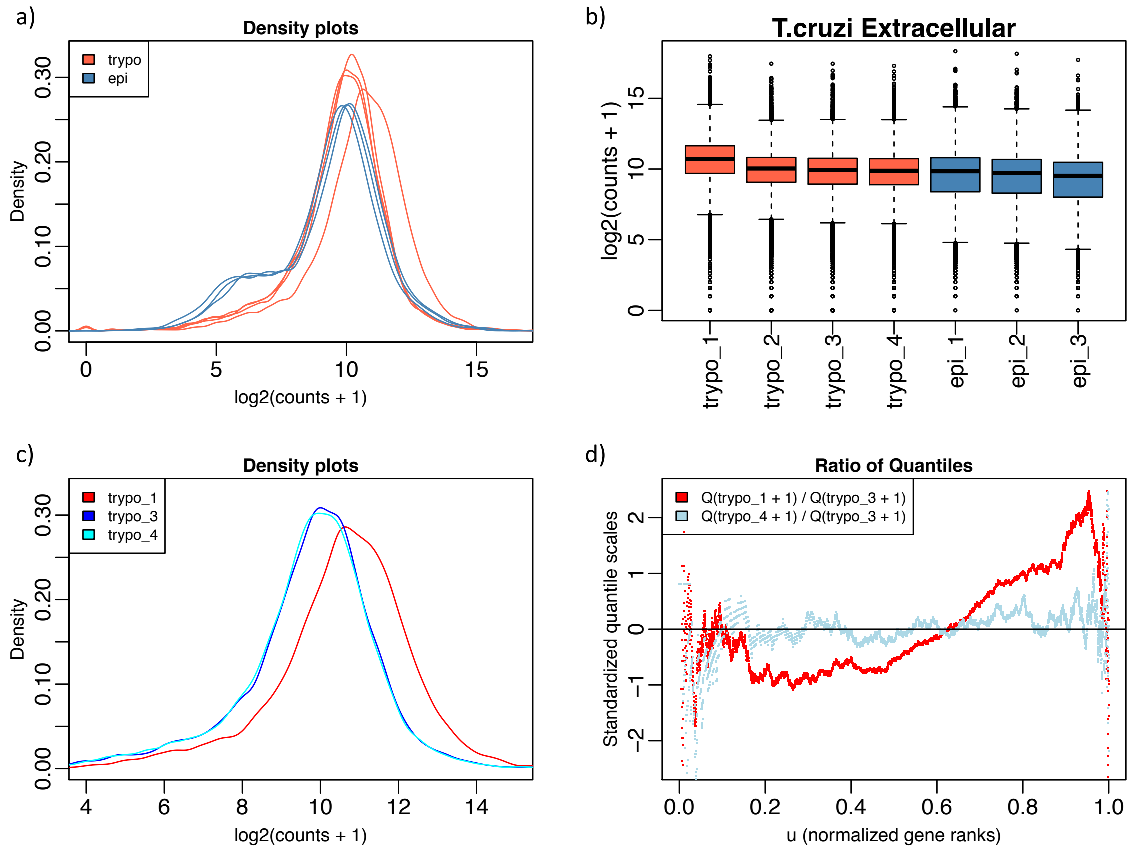


Figure 5.2: **Ratio of quantiles.** If two samples have the same distributions the we would expect the ratio or their quantiles to be approximately constant. (a) Density plots of T.cruzi samples. (b) Boxplot of T.cruzi samples. (c) We have selected three type samples. trypo_3 and trypo_4 have almost the same shape (distribution) but both are very different from trypo_1. (d) The ratio of quantiles. The ratio if trypo_4 to trypo_3 is almost constant. the ratio of trypo_1, to trypo_3 is non-linear suggesting that quantile normalization would work better.

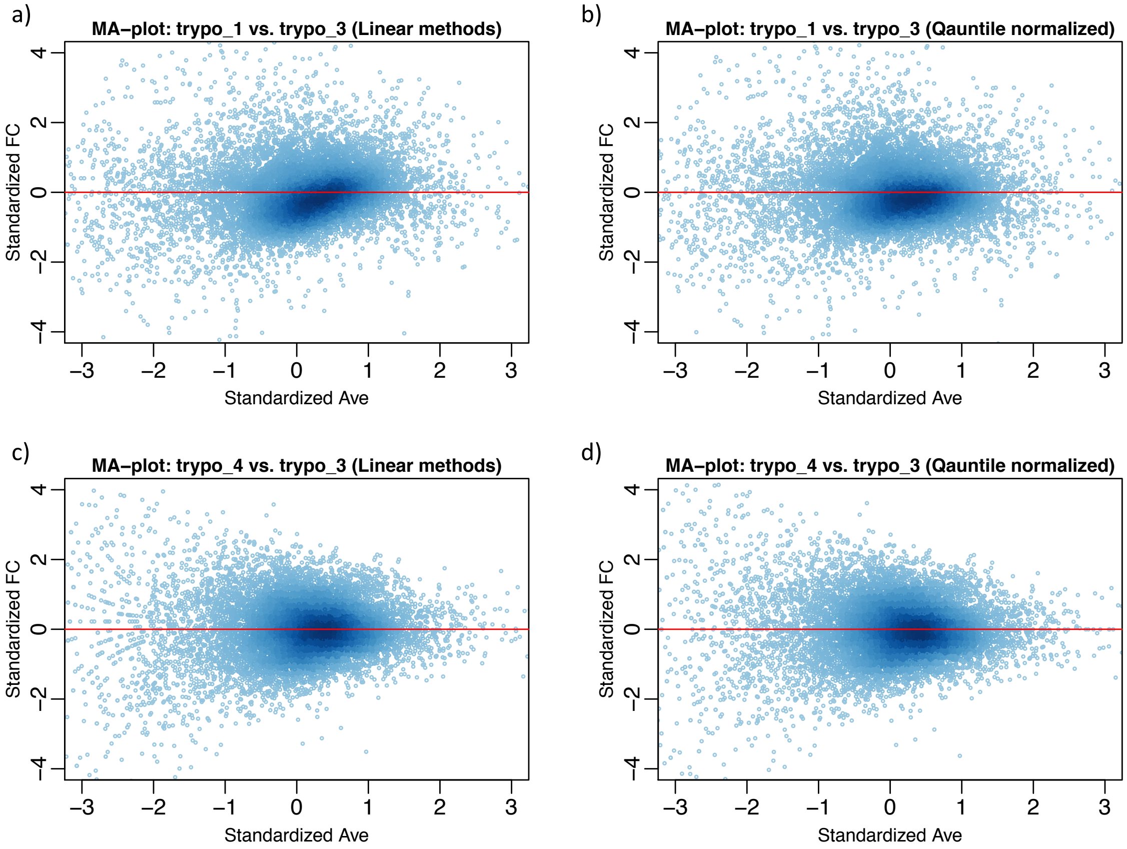


Figure 5.3: **Quantile normalization** MA-plots after quantile normalization shows improvement. (a) Standardized MA-plot. This MA-plot indicares a strong technical bias that cannot be fixed with a scaling method. (b) MA-plot of the quantile normalized samples. (c) MA-plot for scaled samples when shapes are similar. (d) MA-plot for quantile normalized samples when shapes are similar. This is very similar to (c). These plots illustrate the advantages of quantile normalization over scaling methods.

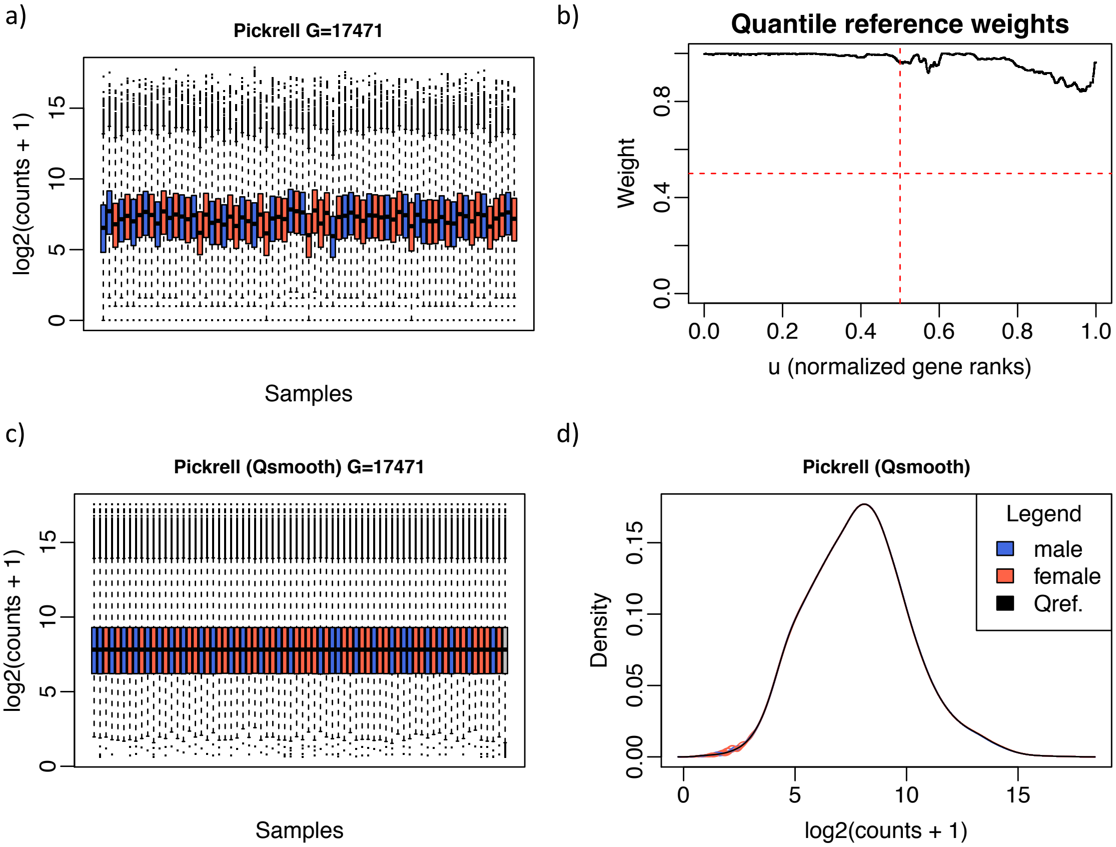


Figure 5.4: **Qsmooth: Pickrell Dataset.** (a) Boxplot of log counts (+1). (b) The weight assigned to the reference quantile. (c) Boxplot after qsmooth normalization. (d) Density plots after qsmooth normalization.

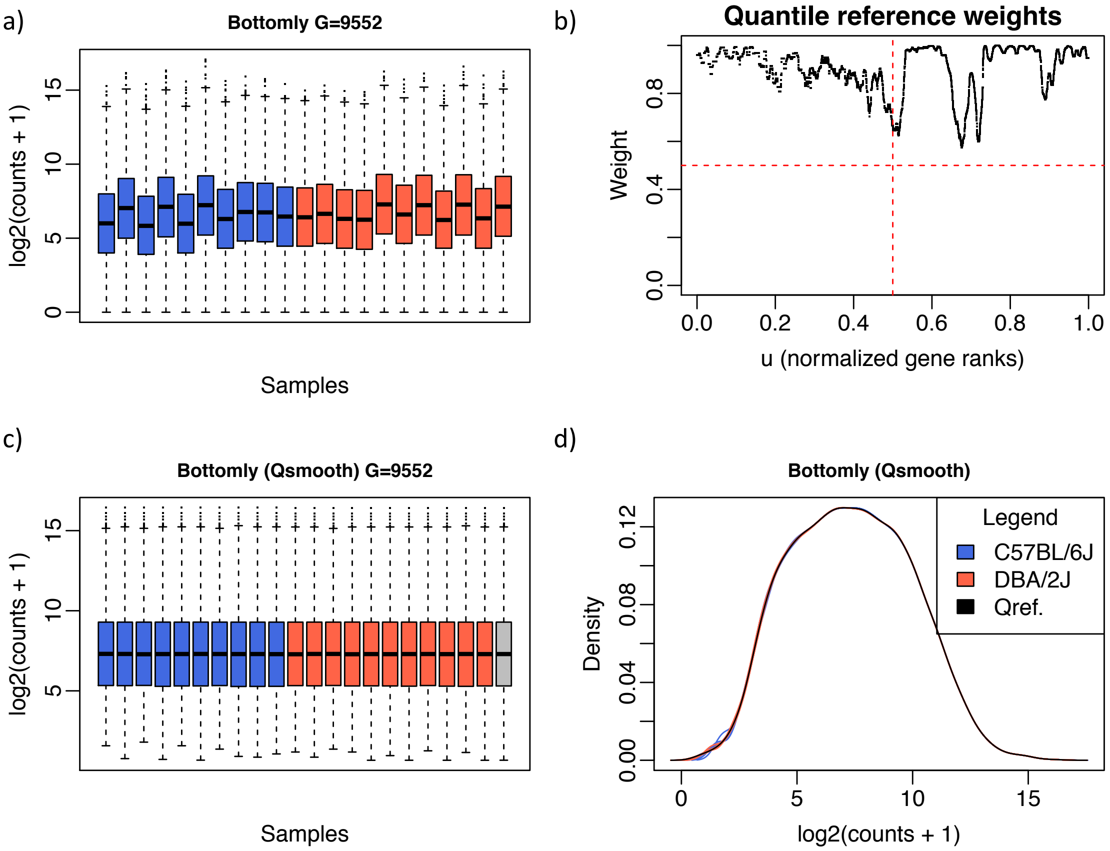


Figure 5.5: **Qsmooth: Bottomly Dataset.** (a) Boxplot of log counts (+1). (b) The weight assigned to the reference quantile. (c) Boxplot after qsmooth normalization. (d) Density plots after qsmooth normalization.

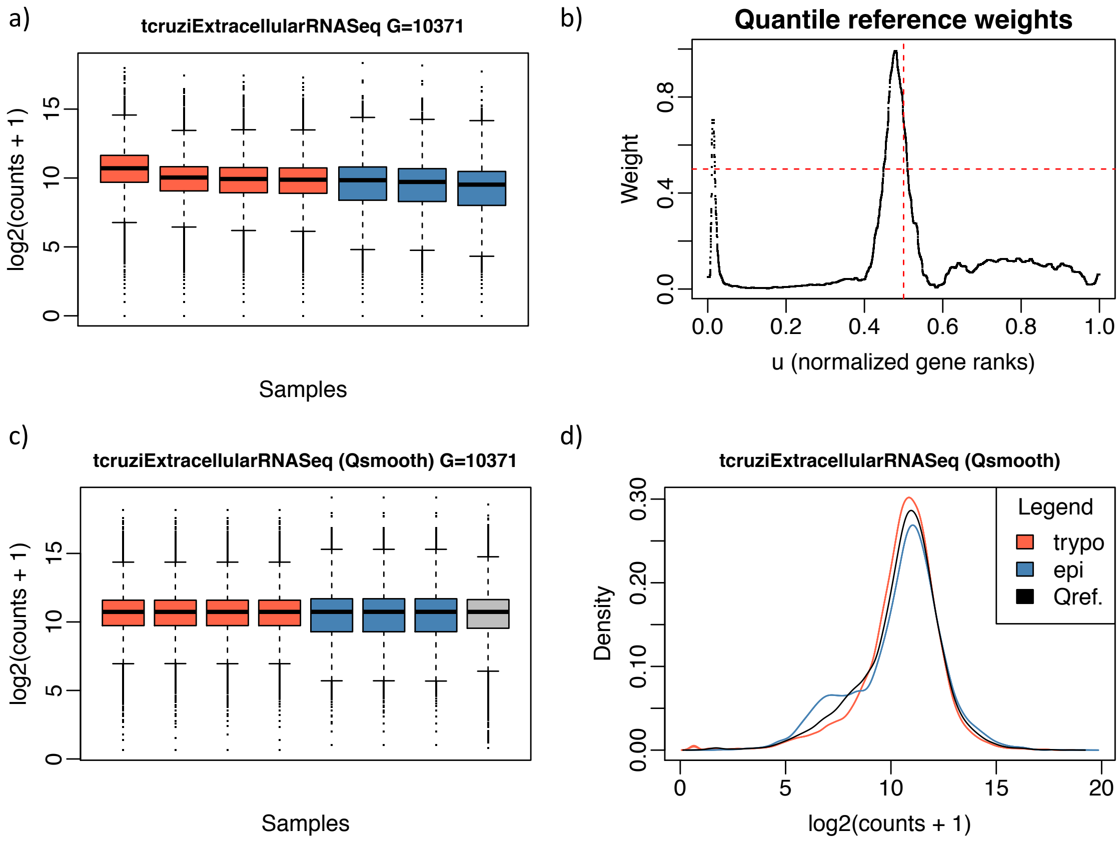


Figure 5.6: **Qsmooth: T.cruzi Dataset.** (a) Boxplot of log counts (+1). (b) The weight assigned to the reference quantile. (c) Boxplot after qsmooth normalization. (d) Density plots after qsmooth normalization.

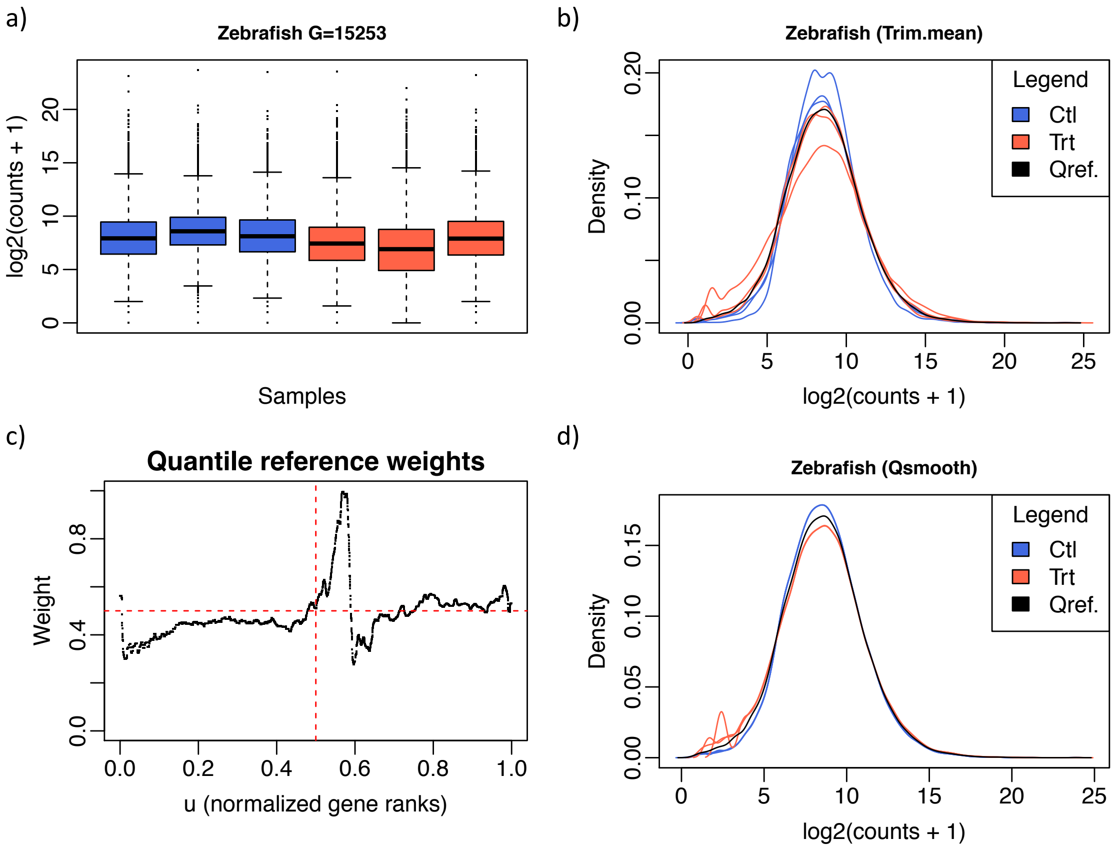


Figure 5.7: **Qsmooth: Zebrafish Dataset.** (a) Boxplot of log counts (+1). (b) The weight assigned to the reference quantile. (c) Boxplot after qsmooth normalization. (d) Density plots after qsmooth normalization.

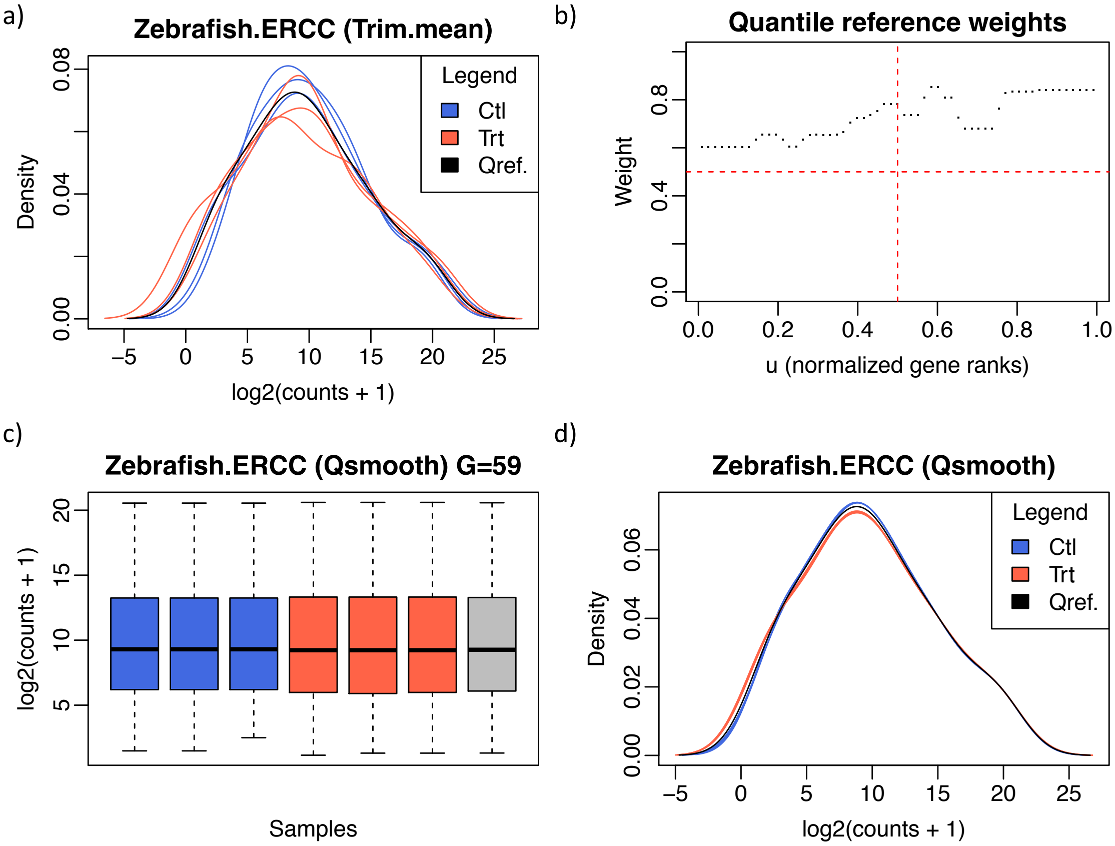


Figure 5.8: **Qsmooth: Zebrafish Spike-in dataset.** (a) Boxplot of log counts (+1). (b) The weight assigned to the reference quantile. (c) Boxplot after qsmooth normalization. (d) Density plots after qsmooth normalization.

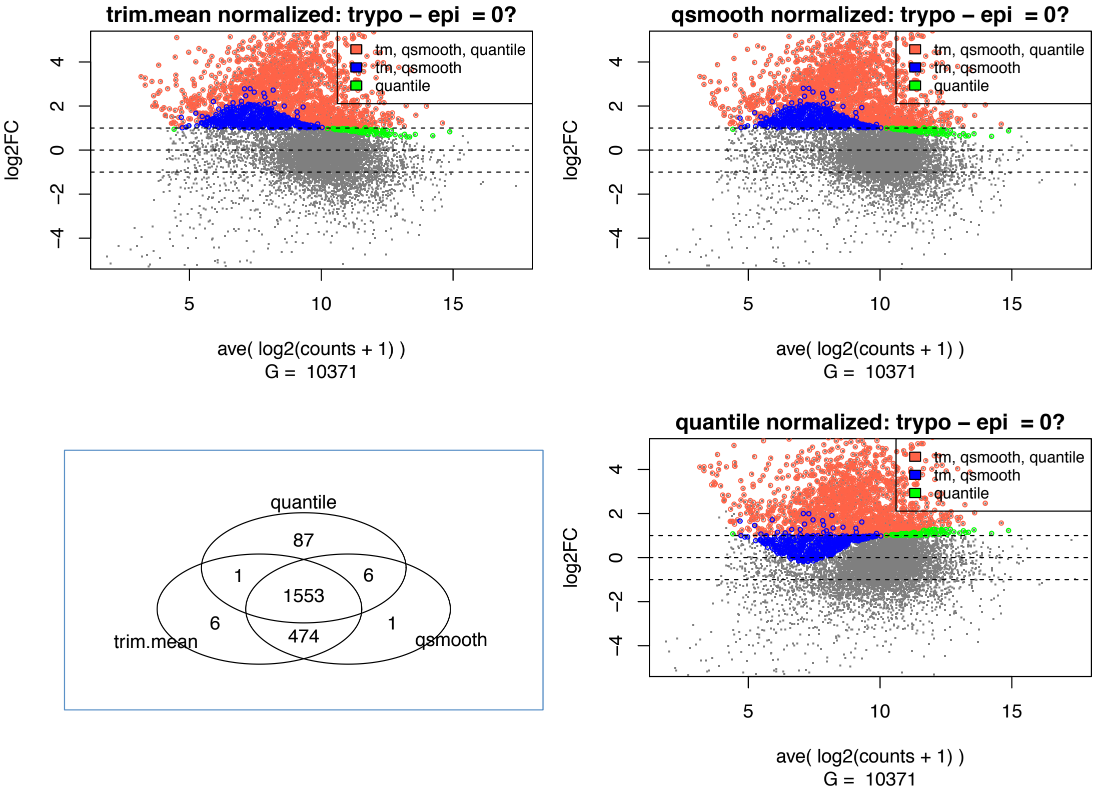


Figure 5.9: *T.cruzi*: **Up-regulated genes.** Differential expression analysis was performed on the *T.cruzi* dataset. Above we have shown the genes that were up-regulated in the trypomastigote (infective) stage as compared to the epimastigote stage. The criteria for selection was a log fold change greater than 1 and an FDR less than 0.05. The differential analysis was perform for 0.25-trim mean scaled data, qsmooth normalized data, and quantile normalized data. In the venn-diagram we have indicated the overlap of differentially expressed genes amongst the three normalization methods used.

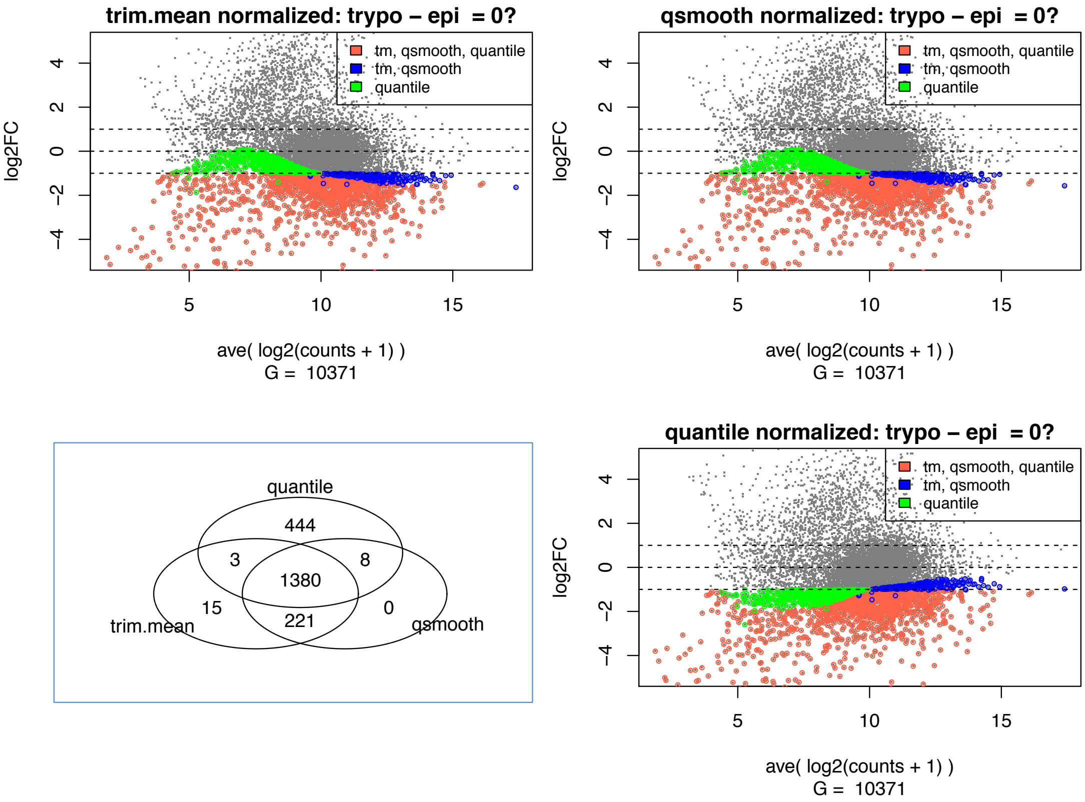


Figure 5.10: *T.cruzi*: **Down-regulated genes.** Differential expression analysis was performed on the *T.cruzi* dataset. Above we have shown the genes that were down-regulated in the trypomastigote (infective) stage as compared to the epimastigote stage. The criteria for selection was a log fold change less than -1 and an FDR less than 0.05. The differential analysis was perform for 0.25-trim mean scaled data, qsmooth normalized data, and quantile normalized data. In the venn-diagram we have indicated the overlap of differentially expressed genes amongst the three normalization methods used.

Chapter 6: Statistical software

6.1 Package: HTShape

We have developed a statistical package in `R` to implement some of the methods discussed in chapter 2 and chapter 3: **HTShape**.

It can be found online at (https://github.com/kokrah/HTShape). The purpose of the **HTShape** package is to compute the shape (i.e. L-skew and L-kurt) statistics of each transcript (e.g. gene) in a high-throughput dataset (e.g. RNA-seq, microarry). Using these statistics we can find genes within a dataset whose sample shape is markedly different from the majority of genes in the same dataset. When put together these shape statistics give an overall description of the entire high-throughput dataset.

The ability to describe the shape of high-throughput genomics data is useful for two reasons: 1. It enriches the exploratory data analysis process, and 2. It provides a means of checking the distributional assumptions of statistical methods.

There are three main functions in this package: (1) **fitShape**, (2) **computeDvals**, and (3) **plotSO**.

Given a dataset such as a high-throughput expression matrix (or just a vector of measurements) the function **fitShape** will compute and return the L-CV, L-skew, and L-kurt estimates for each gene.

Given the shape (i.e. L-skew and L-kurt) estimate of each gene, the function **computeDvals** computes a dissimilarity distance (d-values) between each gene's shape estimate and the typical gene's shape estimate. The d-values range from 0 to 1; where 1 is very close and 0 is very far.

The function **plotSO** shows each gene's shape estimate on a single plot.

6.1.1 Installation

Use devtools (https://github.com/hadley/devtools) to install the latest version of shape from Github:

- require("devtools")
- install_github("kokrah/HTShape")

If all went well you should now be able to load **HTShape**:

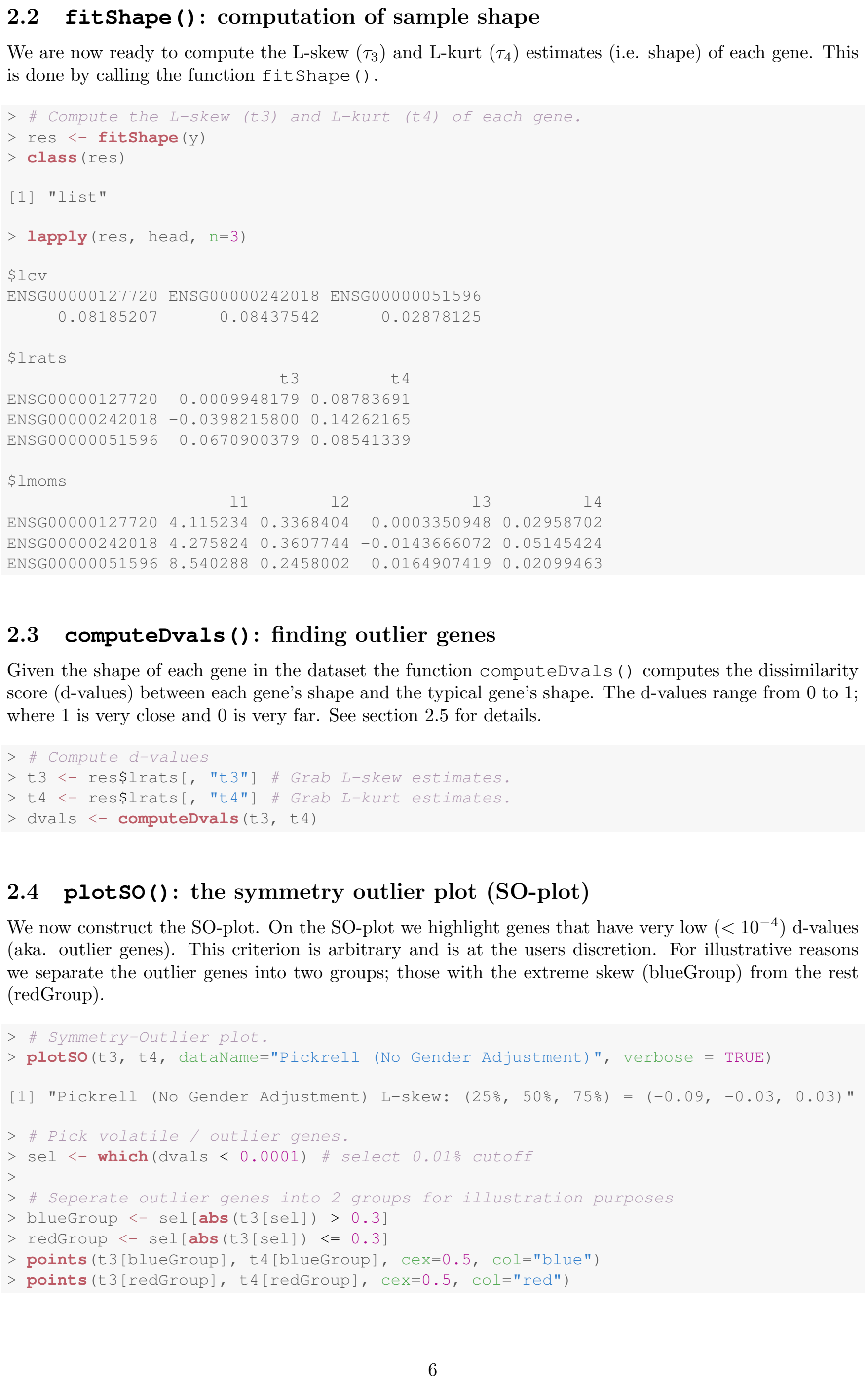
- require("HTShape")
- vignette("HTShape")

6.2 Qsmooth

We have developed a statistical package in `R` to implement some of the methods discussed in chapter 5: **qsmooth**.

It can be found online at (https://github.com/kokrah/qsmooth).

Below is a demonstration of **HTShape** and **qsmooth**.



2.2 `fitShape()`: computation of sample shape

We are now ready to compute the L-skew (τ_3) and L-kurt (τ_4) estimates (i.e. shape) of each gene. This is done by calling the function `fitShape()`.

```
> # Compute the L-skew (t3) and L-kurt (t4) of each gene.
> res <- fitShape(y)
> class(res)

[1] "list"

> lapply(res, head, n=3)

$lcv
ENSG00000127720 ENSG00000242018 ENSG00000051596
     0.08185207      0.08437542      0.02878125

$lrats
                           t3         t4
ENSG00000127720  0.0009948179 0.08783691
ENSG00000242018 -0.0398215800 0.14262165
ENSG00000051596  0.0670900379 0.08541339

$lmoms
                      l1        l2            l3         l4
ENSG00000127720 4.115234 0.3368404  0.0003350948 0.02958702
ENSG00000242018 4.275824 0.3607744 -0.0143666072 0.05145424
ENSG00000051596 8.540288 0.2458002  0.0164907419 0.02099463
```

2.3 `computeDvals()`: finding outlier genes

Given the shape of each gene in the dataset the function `computeDvals()` computes the dissimilarity score (d-values) between each gene's shape and the typical gene's shape. The d-values range from 0 to 1; where 1 is very close and 0 is very far. See section 2.5 for details.

```
> # Compute d-values
> t3 <- res$lrats[, "t3"] # Grab L-skew estimates.
> t4 <- res$lrats[, "t4"] # Grab L-kurt estimates.
> dvals <- computeDvals(t3, t4)
```

2.4 `plotSO()`: the symmetry outlier plot (SO-plot)

We now construct the SO-plot. On the SO-plot we highlight genes that have very low ($< 10^{-4}$) d-values (aka. outlier genes). This criterion is arbitrary and is at the users discretion. For illustrative reasons we separate the outlier genes into two groups; those with the extreme skew (blueGroup) from the rest (redGroup).

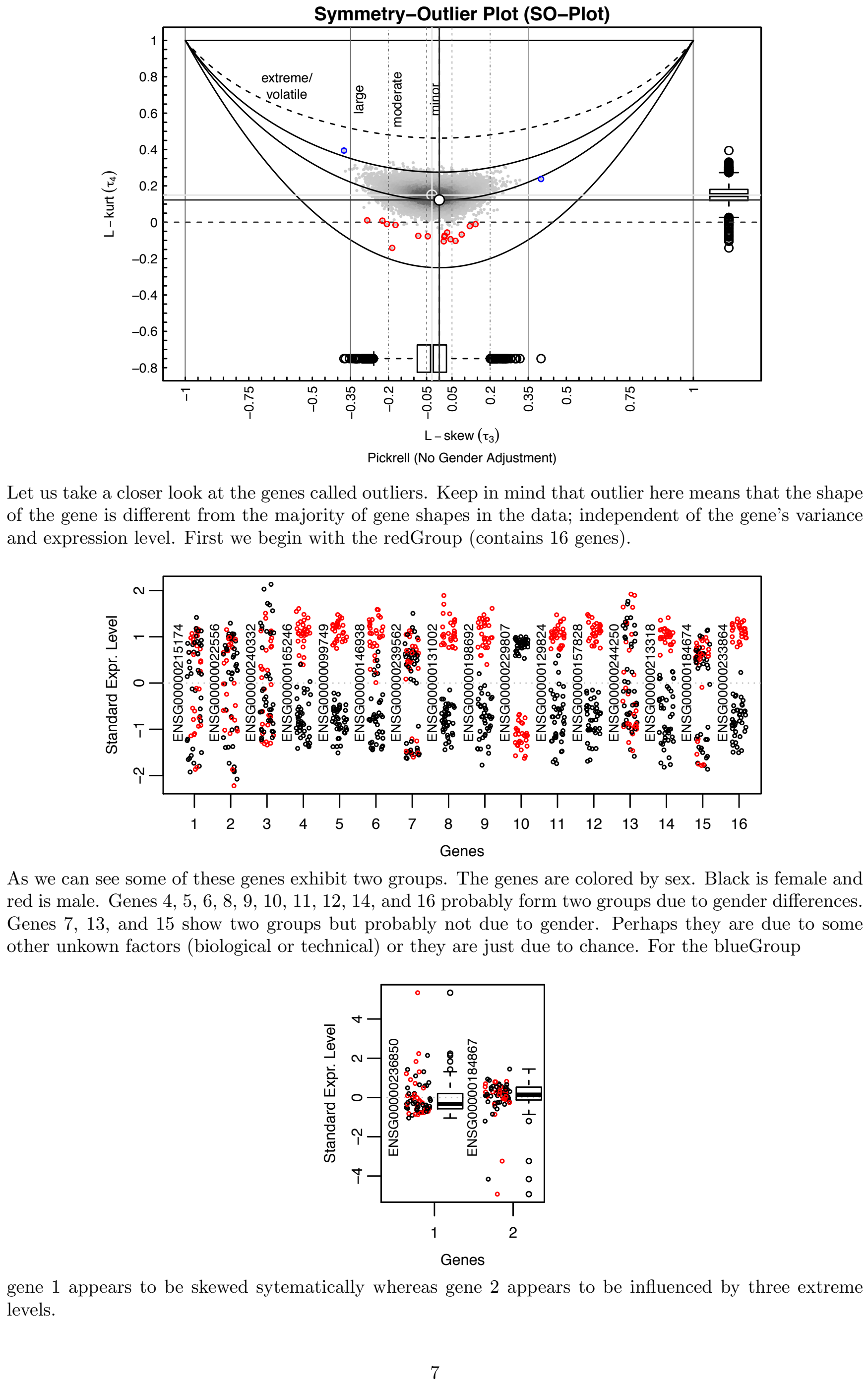
```
> # Symmetry-Outlier plot.
> plotSO(t3, t4, dataName="Pickrell (No Gender Adjustment)", verbose = TRUE)

[1] "Pickrell (No Gender Adjustment) L-skew: (25%, 50%, 75%) = (-0.09, -0.03, 0.03)"

> # Pick volatile / outlier genes.
> sel <- which(dvals < 0.0001) # select 0.01% cutoff
>
> # Seperate outlier genes into 2 groups for illustration purposes
> blueGroup <- sel[abs(t3[sel]) > 0.3]
> redGroup <- sel[abs(t3[sel]) <= 0.3]
> points(t3[blueGroup], t4[blueGroup], cex=0.5, col="blue")
> points(t3[redGroup], t4[redGroup], cex=0.5, col="red")
```

6

Figure 6.1: **Demonstration of HTShape package.** A demonstration of the **fitShape()** function and **plotSO()** function. The input value y above is the Pickrell count matrix.



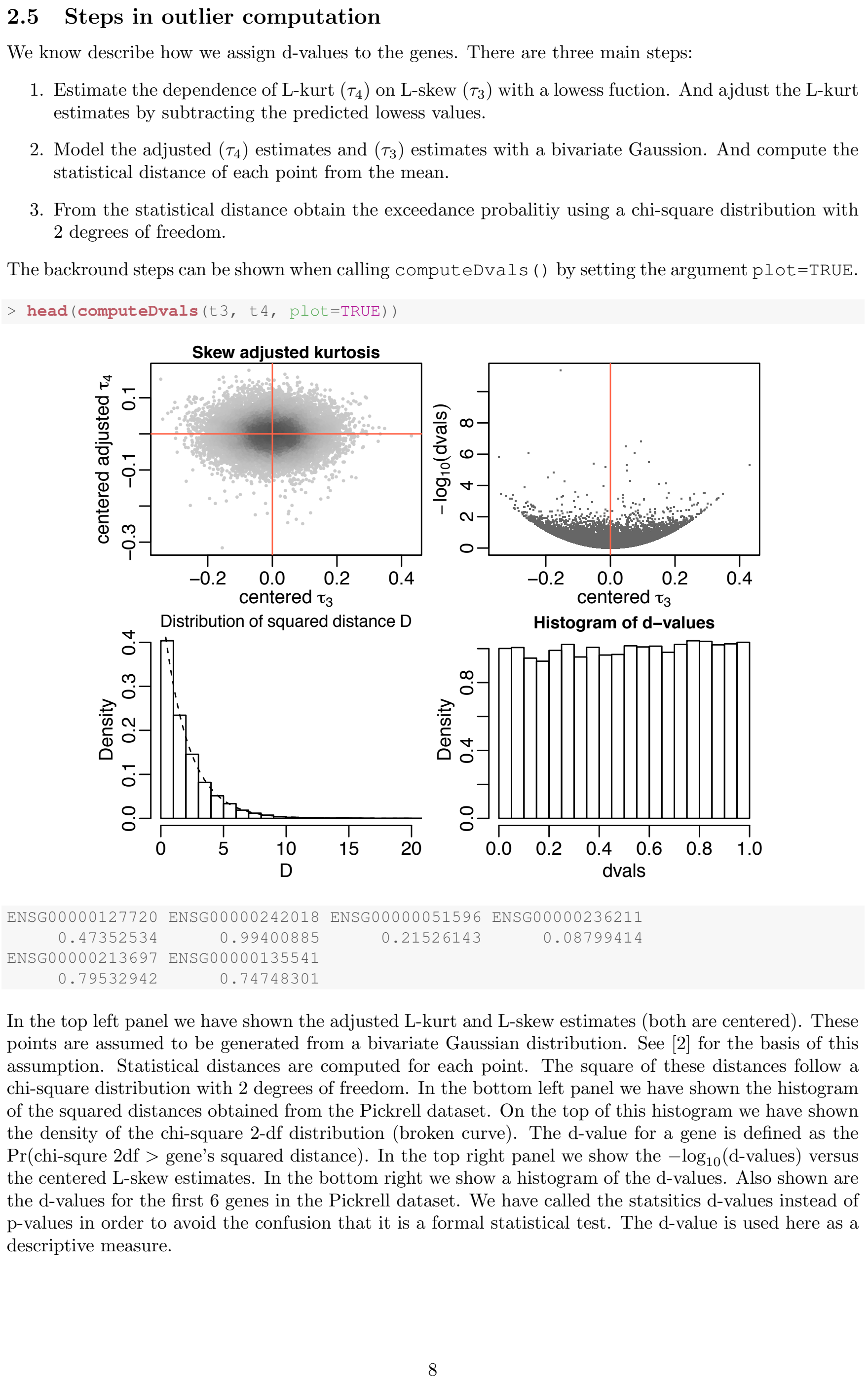
Let us take a closer look at the genes called outliers. Keep in mind that outlier here means that the shape of the gene is different from the majority of gene shapes in the data; independent of the gene's variance and expression level. First we begin with the redGroup (contains 16 genes).

As we can see some of these genes exhibit two groups. The genes are colored by sex. Black is female and red is male. Genes 4, 5, 6, 8, 9, 10, 11, 12, 14, and 16 probably form two groups due to gender differences. Genes 7, 13, and 15 show two groups but probably not due to gender. Perhaps they are due to some other unkown factors (biological or technical) or they are just due to chance. For the blueGroup

gene 1 appears to be skewed sytematically whereas gene 2 appears to be influenced by three extreme levels.

7

Figure 6.2: **Demonstration of HTShape package.** A demonstration of the **plotSO()** function.



2.5 Steps in outlier computation

We know describe how we assign d-values to the genes. There are three main steps:

1. Estimate the dependence of L-kurt (τ_4) on L-skew (τ_3) with a lowess fuction. And ajdust the L-kurt estimates by subtracting the predicted lowess values.
2. Model the adjusted (τ_4) estimates and (τ_3) estimates with a bivariate Gaussion. And compute the statistical distance of each point from the mean.
3. From the statistical distance obtain the exceedance probalitiy using a chi-square distribution with 2 degrees of freedom.

The backround steps can be shown when calling `computeDvals()` by setting the argument `plot=TRUE`.

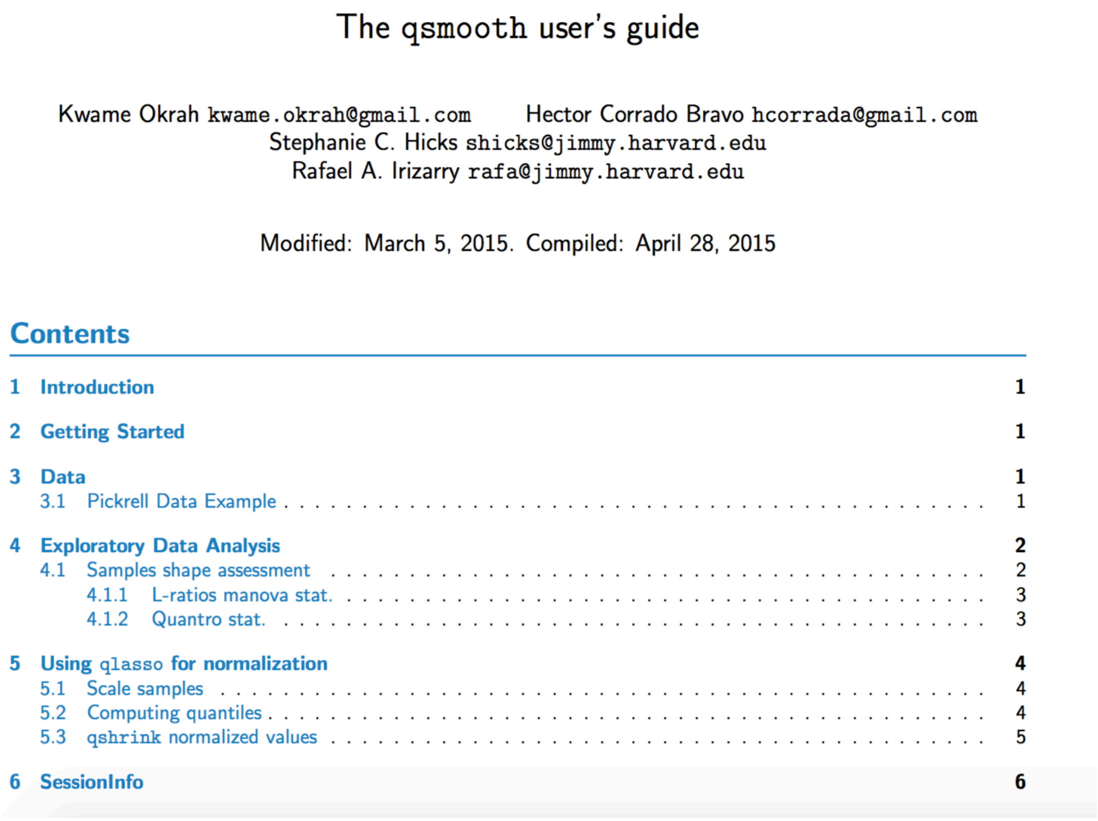
```
> head(computeDvals(t3, t4, plot=TRUE))
```

```
ENSG00000127720 ENSG00000242018 ENSG00000051596 ENSG00000236211
     0.47352534      0.99400885      0.21526143      0.08799414
ENSG00000213697 ENSG00000135541
     0.79532942      0.74748301
```

In the top left panel we have shown the adjusted L-kurt and L-skew estimates (both are centered). These points are assumed to be generated from a bivariate Gaussian distribution. See [2] for the basis of this assumption. Statistical distances are computed for each point. The square of these distances follow a chi-square distribution with 2 degrees of freedom. In the bottom left panel we have shown the histogram of the squared distances obtained from the Pickrell dataset. On the top of this histogram we have shown the density of the chi-square 2-df distribution (broken curve). The d-value for a gene is defined as the Pr(chi-squre 2df > gene's squared distance). In the top right panel we show the $-\log_{10}$(d-values) versus the centered L-skew estimates. In the bottom right we show a histogram of the d-values. Also shown are the d-values for the first 6 genes in the Pickrell dataset. We have called the statsitics d-values instead of p-values in order to avoid the confusion that it is a formal statistical test. The d-value is used here as a descriptive measure.

8

Figure 6.3: **Demonstration of HTShape package.** A demonstration of the **computeDval()** function.



The qsmooth user's guide

Kwame Okrah kwame.okrah@gmail.com Hector Corrado Bravo hcorrada@gmail.com
Stephanie C. Hicks shicks@jimmy.harvard.edu
Rafael A. Irizarry rafa@jimmy.harvard.edu

Modified: March 5, 2015. Compiled: April 28, 2015

Contents

1 Introduction 1

2 Getting Started 1

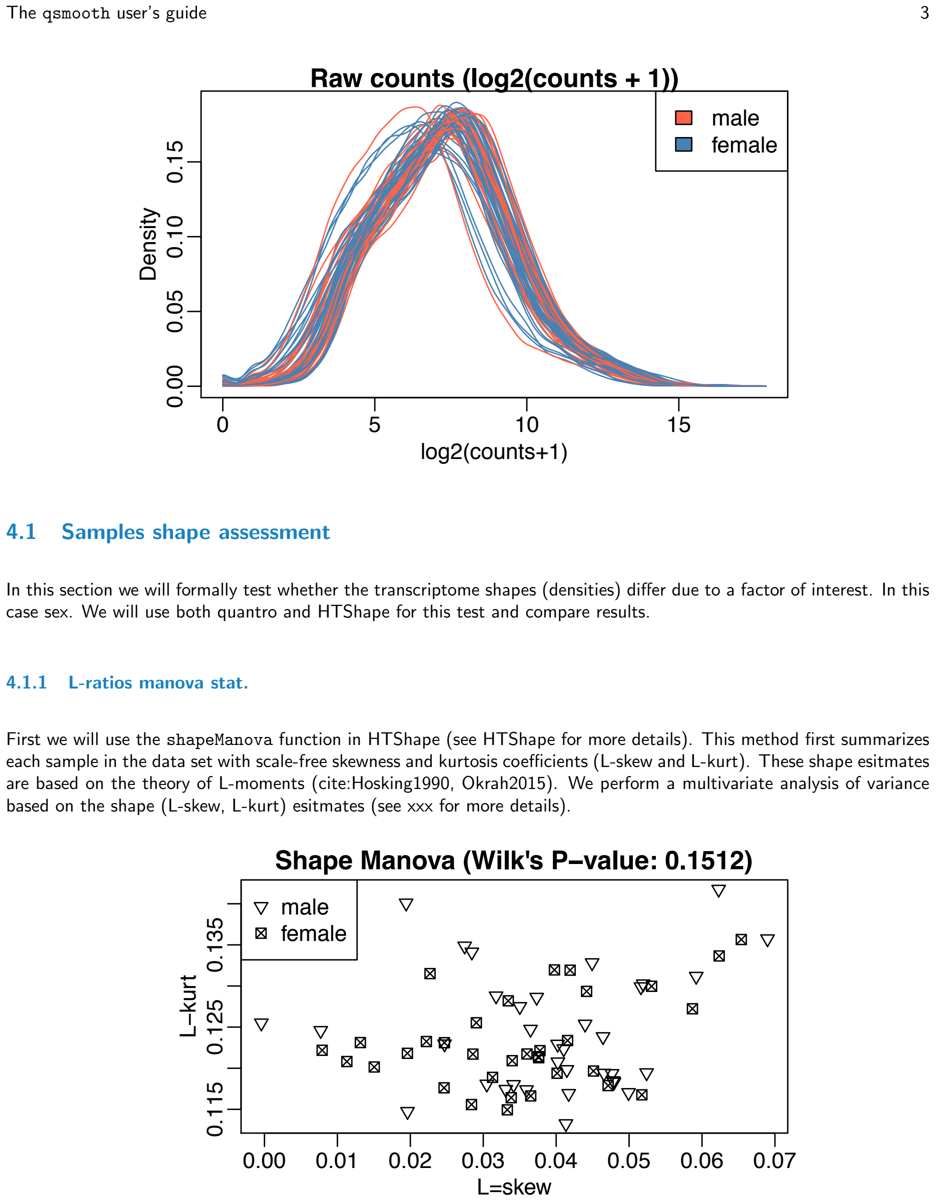
3 Data 1
- 3.1 Pickrell Data Example 1

4 Exploratory Data Analysis 2
- 4.1 Samples shape assessment 2
  - 4.1.1 L-ratios manova stat. 3
  - 4.1.2 Quantro stat. 3

5 Using qlasso for normalization 4
- 5.1 Scale samples 4
- 5.2 Computing quantiles 4
- 5.3 qshrink normalized values 5

6 SessionInfo 6

Figure 6.4: **User guide: qsmooth** A screen capture of the qsmooth user guide. The work is based on a collaboration with Stephanie Hicks and Rafael Irizarry.



The qsmooth user's guide 3

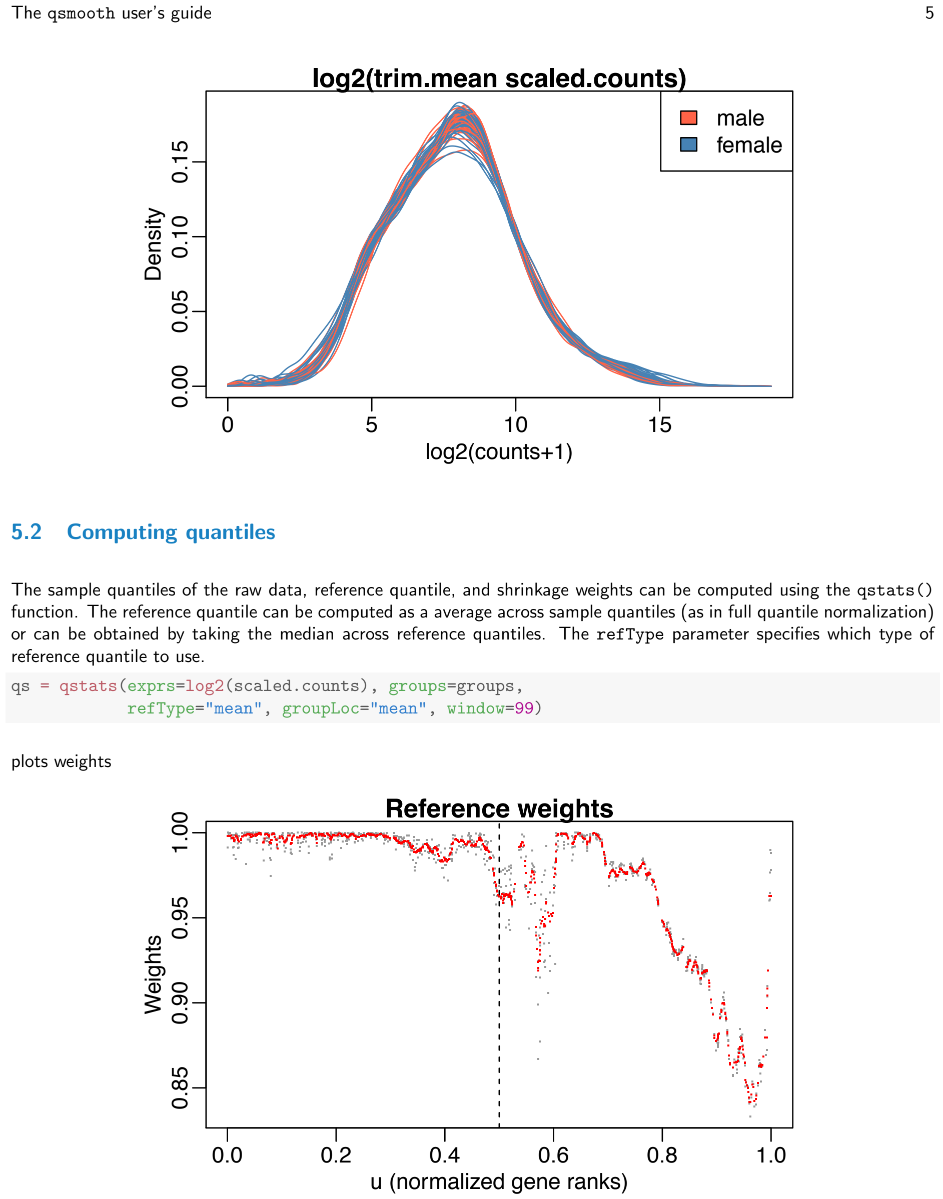
4.1 Samples shape assessment

In this section we will formally test whether the transcriptome shapes (densities) differ due to a factor of interest. In this case sex. We will use both quantro and HTShape for this test and compare results.

4.1.1 L-ratios manova stat.

First we will use the shapeManova function in HTShape (see HTShape for more details). This method first summarizes each sample in the data set with scale-free skewness and kurtosis coefficients (L-skew and L-kurt). These shape esitmates are based on the theory of L-moments (cite:Hosking1990, Okrah2015). We perform a multivariate analysis of variance based on the shape (L-skew, L-kurt) esitmates (see xxx for more details).

Figure 6.5: **Demonstration of qsmooth package.** A demonstration of the **shapeManova()** function. The input value above is the Pickrell count matrix.



The qsmooth user's guide 5

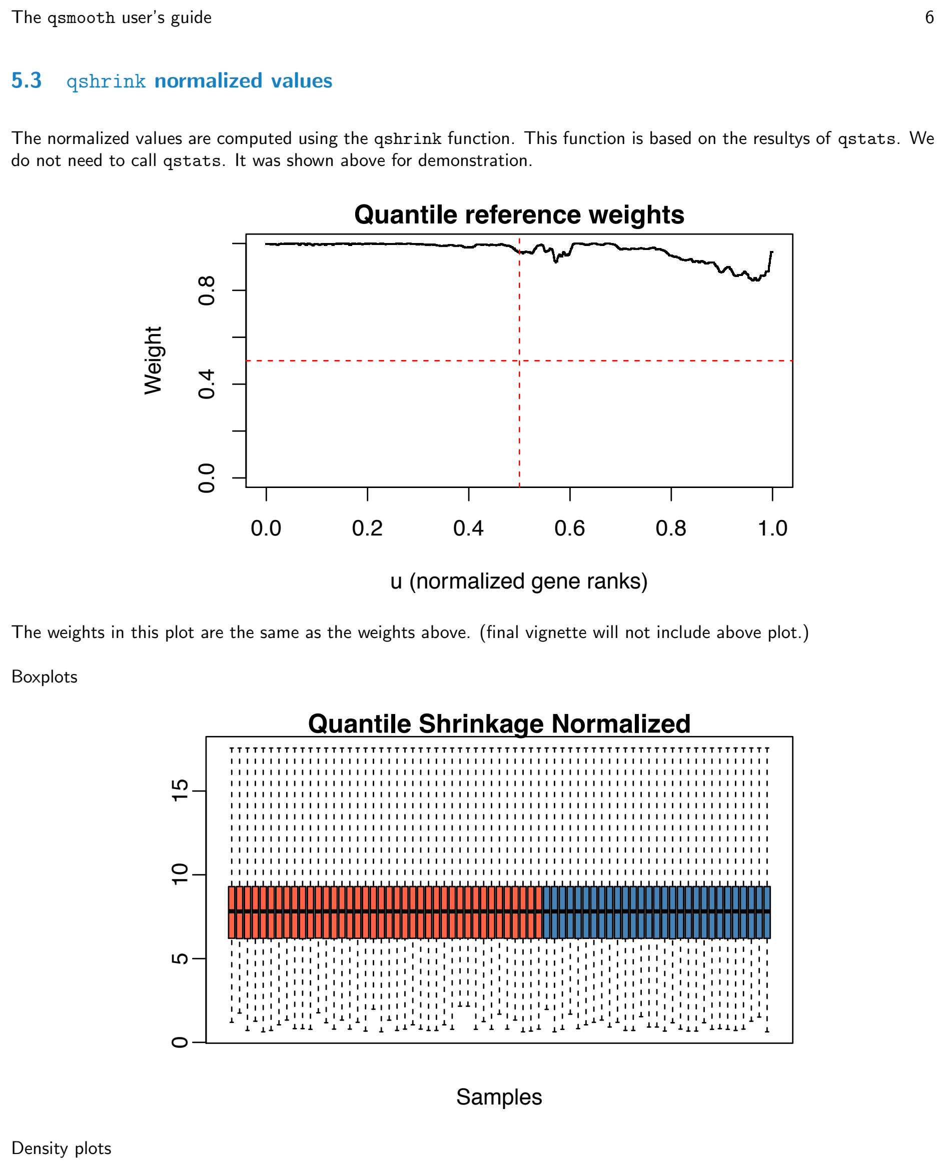
5.2 Computing quantiles

The sample quantiles of the raw data, reference quantile, and shrinkage weights can be computed using the `qstats()` function. The reference quantile can be computed as a average across sample quantiles (as in full quantile normalization) or can be obtained by taking the median across reference quantiles. The `refType` parameter specifies which type of reference quantile to use.

```
qs = qstats(exprs=log2(scaled.counts), groups=groups,
            refType="mean", groupLoc="mean", window=99)
```

plots weights

Figure 6.6: **Demonstration of qsmooth package: qsats.** A demonstration of the **qstats()** function.



The qsmooth user's guide 6

5.3 qshrink normalized values

The normalized values are computed using the qshrink function. This function is based on the resultys of qstats. We do not need to call qstats. It was shown above for demonstration.

Quantile reference weights
Weight
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
u (normalized gene ranks)

The weights in this plot are the same as the weights above. (final vignette will not include above plot.)

Boxplots

Quantile Shrinkage Normalized
0
5
10
15
Samples

Density plots

Figure 6.7: **Demonstration of qsmooth package: qshrink.** A demonstration of the **qshrink()** function.

Chapter 7: Summary

7.1 Summary

In this dissertation we have developed two data exploratory tools (1) SO-plot and (2) Wilk's shape manova. Both of these tools are built on the statistical properties of L-moments statistics. The SO-plot summarizes the shape of individual genes, and when taken together they can give is a global view of the shape of our dataset (gene-wise). We also provided an algorithm for detecting genes (volatile/outlier genes) with shapes that are markedly different from the majority in a given high-throughput dataset. The SO-plot provides a universal plot for assessing the distributional assumptions of high-throughput genomics data. Given a dataset one can construct the SO-plot and determine whether a t-test based method is appropriate.

Although we analyzed RNA-seq and microarray data other types of high-throughput data can benefit from this kind of analysis. For example in methylation analysis where statisticians have defined complex probability models [65, 66] but t-tests are also commonly in use [67]. It would also be worth it to explore the use of these methods for exploration and analysis of differential variability in gene expression [68].

We also introduced the Wilk's shape manova procedure for checking whether the empirical sample distributions of RNA-seq data are the same. This algorithm is general and can be applied to other high-throughput datasets, for example DNA methylation data.

We have presented an algorithm for normalization that is a modification of quantile normalization. This algorithm implicitly tests the global similarity assumption at each quantile and decides how much to shrink towards the quantile reference. This algorithm is new and is currently being tested against other methods.

Appendix A: Datasets

A total of 7 publicly available datasets were used in this dissertation. 6 RNA-seq datastes and one microarray dataset.

A.1 Hammoud

The Hammoud dataset [61] contains 10 samples of mRNA profiles of 8-week old wild type mice (strain: C57BL/6). Of the 10 samples 5 were obtained from spermatids (cells) and other 5 from spermatocytes (cells). Summarized counts in the form of FPKM can be downloaded at GEO:GSE49622. Genes with at least one FPKM in 5 or more samples (the minimum of the 5 spermatid samples and the 5 spermatocyte samples) were kept for analysis. The SO-plot of the Hammoud dataset (log2(FPKM + 1) residuals) is shown in Figure 3.2.

A.2 Pickrell dataset

The Pickrell dataset [55] is part of the International HapMap Project. RNA samples were extracted from the lymphoblastoid cell lines of 69 unrelated Nigerian individuals, 29 males and 40 females. The aligned and summarized count matrix was obtained from the bioconductor (http://www.bioconductor.org) tweeDEseqCountData package. Genes with at least one count per million (cpm) in 29 or more samples were kept and normalized for library size. The SO-plot of the log2(counts + 1) is shown in Figure 2.7.

A.3 Bottomly dataset

The Bottomly dataset [60] was obtained from the ReCount webpage (http://bowtie-bio.sourceforge.net/recount). See [74] for details on alignment and counting. It contains counts summarizing an RNA-seq experiment that includes 21 samples from inbred mouse strains. Eleven of the samples came from the strain DBA/2J and 10 from the strain C57BL/6J. Genes with at least one count per million (cpm) in 10 or more samples were kept and normalized for library size. The SO-plot of the log2(counts + 1) of the 10 C57BL/6J strain samples is shown in Figure 3.2.

A.4 MAQC dataset

The MAQC dataset is part of the sequencing quality control project [62]. It contains 14 technical replicates, 7 from Stratagene's Universal Human Reference RNA (UHRR) and 7 from Ambion's Human Brain Reference RNA (HBRR) and was downloaded from the ReCount website. See [74] for details on alignment and counting. Genes with at least one count per million (cpm) in 7 or more samples were kept and normalized for library size. The SO-plot of the log2(counts + 1) is shown in Figure 3.2.

A.5 Geng (microarray)

The Geng (microarray) dataset contains a time-course experiment using a single-channel Agilent Whole Rat Genome Microarray 4x44K arrays. The dataset can be found at GEO:GSE33005 and was used in the Section 17.4 of the limma user guide [16] as a case study. Female Wister rats were, on a daily basis, given corn oil for 14 days at dosage levels: 2 ml/kg (5 samples), 5 ml/kg (4 samples), 10 ml/kg (5 samples), and saline at the dosage of 10 ml/kg (5 samples). For background correction, normalization and filtration we followed the limma user guide (http://www.bioconductor.org). We use this dataset to demonstrate how a microarray description of the sample shape of genes compares to an RNA-seq description (see Figure 3.2).

A.6 Zebrafish

The Zebrafish dataset consist of samples from Olfactory sensory neurons cells obtained from the embryos of zebrafish. There were 6 samples in total. Gellein was applied to 3 of the samples whereas the other 3 were used as controls. Ambion ERCC RNA spike-in control mix 1 was added to total RNA before mRNA extraction in the library process. The library was sequenced on an Illumina HiSeq2000. For more details on the experiment please see [29]. FASTQ files containing the unmapped reads are available at GEO:GSE53334.

A.7 SEQC

The SEQC dataset is part of the Microarray Quality Control (MAQC) project [62]. The aims of the project is to assess the technical performance of high-throughput genomics technology, including RNA-seq. The dataset contains four tissue types; A, B, C, and D. Tissue A is Stratagene's universal human reference RNA and tissue B is Ambion's human brain reference RNA. Tissues C and D are mixtures of A and

B in the ratios of 3:1 and 1:3 respectively. For each of the total RNA from the four tissue samples, 4 libraries were constructed. Making 4 technical replicates per tissue type. Each of these 4 technical replicates were were divided into 2 parts (to be sequenced on two flow-cells). Each of 2 parts were further separated into 8 parts (to be sequenced in each of the 8 lanes in a flow-cell). The ERCC spike-in mix 1 was spiked into the total RNA of tissue A and mix 2 was spiked into the total RNA of tissue B prior to the creation of tissue C and tissue D.

Bibliography

[1] Robert J Brooker. *Genetics: analysis & principles.* Granite Hill Publishers, 2009.

[2] Isaac S Kohane, Atul J Butte, and Alvin Kho. *Microarrays for an integrative genomics.* MIT press, 2002.

[3] Stephen Fodor, Richard P Rava, Xiaohua C Huang, Ann C Pease, Christopher P Holmes, and Cynthia L Adams. Multiplexed biochemical assays with biological chips. *Nature*, 364:555–556, 1993.

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

[5] Brian T Wilhelm, Samuel Marguerat, Stephen Watt, Falk Schubert, Valerie Wood, Ian Goodhead, Christopher J Penkett, Jane Rogers, and Jürg Bähler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):1239–1243, 2008.

[6] Sam Alsford, Daniel J Turner, Samson O Obado, Alejandro Sanchez-Flores, Lucy Glover, Matthew Berriman, Christiane Hertz-Fowler, and David Horn. High-throughput phenotyping using parallel sequencing of rna interference targets in the african trypanosome. *Genome research*, 21(6):915–924, 2011.

[7] Todd A Minning, Jacqueline Bua, Gabriela A Garcia, RA McGraw, and Rick L Tarleton. Microarray profiling of gene expression during trypomastigote to amastigote transition in trypanosoma cruzi. *Molecular and biochemical parasitology*, 131(1):55–64, 2003.

[8] Emile F Nuwaysir, Michael Bittner, Jeffrey Trent, J Carl Barrett, and Cynthia A Afshari. Microarrays and toxicology: the advent of toxicogenomics. *Molecular carcinogenesis*, 24(3):153–159, 1999.

[9] Qihong Huang, Robert T Dunn, Supriya Jayadev, Olimpia DiSorbo, Franklin D Pack, Spencer B Farr, Raymond E Stoll, and Kerry T Blanchard. Assessment of cisplatin-induced nephrotoxicity by microarray technology. *Toxicological Sciences*, 63(2):196–207, 2001.

[10] Daniela Witten, Robert Tibshirani, Sam G Gu, Andrew Fire, and Weng-Onn Lui. Ultra-high throughput sequencing-based small rna discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC biology*, 8(1):58, 2010.

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

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

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

[14] Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[15] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome biology*, 11:R106, 2010.

[16] G. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:3, 2004.

[17] M. Robinson and G. Smyth. Small sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9:321, 2008.

[18] Fred Sanger and Alan R Coulson. A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of molecular biology*, 94(3):441–448, 1975.

[19] Eija Korpelainen, Jarno Tuimala, Panu Somervuo, Mikael Huss, and Garry Wong. *RNA-seq Data Analysis: A Practical Approach*. CRC Press, 2014.

[20] Jakob Lovén, David A Orlando, Alla A Sigova, Charles Y Lin, Peter B Rahl, Christopher B Burge, David L Levens, Tong Ihn Lee, and Richard A Young. Revisiting global gene expression analysis. *Cell*, 151(3):476–482, 2012.

[21] Yang Liao, Gordon K Smyth, and Wei Shi. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research*, page gkt214, 2013.

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

[23] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[24] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, page btu638, 2014.

[25] Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. *Nature biotechnology*, 32(5):462–464, 2014.

[26] Shawn C Baker, Steven R Bauer, Richard P Beyer, James D Brenton, Bud Bromley, John Burrill, Helen Causton, Michael P Conley, Rosalie Elespuru, Michael Fero, et al. The external rna controls consortium: a progress report. *Nature methods*, 2(10):731–734, 2005.

[27] External RNA Controls Consortium et al. Proposed methods for testing and selecting the ercc external rna controls. *BMC genomics*, 6(1):150, 2005.

[28] Life Technologies. Ercc rna spike-in control mixes (user's guide). publication number: 4455352, catalog number: 4456740, 4456739. *Life Technologies*, 2014.

[29] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902, 2014.

[30] Håvard Aanes, Cecilia Winata, Lars F Moen, Olga Østrup, Sinnakaruppan Mathavan, Philippe Collas, Torbjørn Rognes, and Peter Aleström. Normalization of rna-sequencing data from samples with varying mrna levels. *PloS one*, 9(2):e89158, 2014.

[31] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[32] Eli Eisenberg and Erez Y Levanon. Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574, 2013.

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

[34] Peter W Laird. Principles and challenges of genome-wide dna methylation analysis. *Nature Reviews Genetics*, 11(3):191–203, 2010.

[35] Jessica C Mar, Yasumasa Kimura, Kate Schroder, Katharine M Irvine, Yoshihide Hayashizaki, Harukazu Suzuki, David Hume, and John Quackenbush. Data-driven normalization strategies for high-throughput quantitative rt-pcr. *BMC bioinformatics*, 10(1):110, 2009.

[36] C. Law, Y. Chen, W. Shi, and G. Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15:R29, 2014.

[37] George Casella. An introduction to empirical bayes data analysis. *The American Statistician*, 39(2):83–87, 1985.

[38] J. Tukey. Modern techniques in data analysis. In *NSF-sponsored regional research conference at Southeastern Massachusetts University*, North Dartmouth, MA, 1977.

[39] J. Hosking. L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52:105–124, 1990.

[40] J. Martinez and B. Iglewicz. Some properties of the tukey g and h family of distributions. *Communications in Statistics-Theory and Methods*, 13:353–369, 1984.

[41] Y. Xu, B. Iglewicz, and I. Chervoneva. Robust estimation of the parameters of g-and-h distributions, with applications to outlier detection. *Computational Statistics & Data Analysis*, 75:66–80, 2014.

[42] P. Delicado and M. Goria. A small sample comparison of maximum likelihood, moments and L-moments methods for the asymmetric exponential power distribution. *Computational Statistics & Data Analysis*, 52:1661–1673, 2008.

[43] M. Peel, Q. Wang, R. Vogel, and T. McMAHON. The utility of L-moment ratio diagrams for selecting a regional probability distribution. *Hydrological sciences journal*, 46:147–155, 2001.

[44] W. Kirby. Algebraic boundedness of sample statistics. *Water Resources Research*, 10:220–222, 1974.

[45] R. Vogel and N. Fennessey. L-moment diagrams should replace product moment diagrams. *Water Resources Research*, 29:1745–1752, 1993.

[46] A. Sankarasubramanian and K. Srinivasan. Investigation and comparison of sampling properties of L-moments and conventional moments. *Journal of Hydrology*, 218(1):13–34, 1999.

[47] J. Hosking and J. Wallis. *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press, 2005.

[48] J. Leek, R. Scharpf, H. Bravo, D. Simcha, B. Langmead, W. Johnson, D. Geman, K. Baggerly, and R. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11:733–739, 2010.

[49] J. Wilkins. A note on skewness and kurtosis. *The Annals of mathematical statistics*, 15:333–335, 1944.

[50] D. Hoaglin, F. Mosteller, and J. Tukey. *Exploring Data, Tables, Trends, and Shapes*. Wiley, New York, NY, 1985.

[51] A. Khan and G. Rayner. Robustness to non-normality of common tests for the many-sample location problem. *Journal of Applied Mathematics & Decision Sciences*, 7:187–206, 2003.

[52] W. Asquith. *Distributional analysis with L-moment statistics using the R environment for statistical computing*. CreateSpace Independent Publishing Platform, 2011.

[53] Elsayed AH Elamir and Allan H Seheult. Exact variance structure of sample L-moments. *Journal of Statistical Planning and Inference*, 124:337–359, 2004.

[54] R. McCuen. *Statistical methods for engineers*. Prentice Hall PTR, 1985.

[55] K. Pickrell, J. Marioni, A. Pai, J. Degner, B. Engelhardt, E. Nkadori, J. Veyrieras, M. Stephens, Y. Gilad, and J. Pritchard. Understanding mechanisms underlying human gene expression variation with RNA-sequencing. *Nature*, 464:768–772, 2010.

[56] A. Mortazavi, B.Williams, K. McCue, L. Shchaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5:621–628, 2008.

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

[58] C. Soneson and M. Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics*, 14:91, 2013.

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

[60] D. Bottomly, N. Walter, J. Hunter, P. Darakjian, S. Kawane, K. Buck, R. Searles, M. Mooney, S. McWeeney, and R. Hitzemann. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-seq and microarrays. *PloS one*, 6:e17820, 2011.

[61] S. Hammoud, D. Low, C. Yi, D. Carrell, E. Guccione, and B. Cairns. Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. *Cell stem cell*, 2014.

[62] SEQC/MAQC-III Consortium et al. A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature Biotechnology*, 32(9):903–914, 2014.

[63] 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)*, 52:289–300, 1995.

[64] D. Hoaglin, F. Mosteller, and J. Tukey. *Understanding Robust and Exploratory Data Analysis*. Wiley, New York, NY, 1983.

[65] H. Feng, K. Conneely, and H. Wu. A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42:e69–e69, 2014.

[66] P. Du, X. Zhang, C. Huang, N. Jafari, W. Kibbe, L. Hou, and S. Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11:587, 2010.

[67] M. Aryee, A. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. Feinberg, K. Hansen, and R. Irizarry. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30:1363–1369, 2014.

[68] H. Corrada-Bravo, V. Pihur, M. McCall, R. Irizarry, and J. Leek. Gene expression anti-profiles as a basis for accurate universal cancer signatures. *BMC bioinformatics*, 13:272, 2012.

[69] Alexander Zien, Thomas Aigner, Ralf Zimmer, and Thomas Lengauer. Centralization: a new method for the normalization of gene expression data. *Bioinformatics*, 17(suppl 1):S323–S331, 2001.

[70] Stephanie C. Hicks and Rafael A. Irizarry. When to use quantile normalization? *bioRxiv*, 2014.

[71] Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis*, volume 4. Prentice hall Englewood Cliffs, NJ, 1992.

[72] Julie Clayton. Chagas disease 101. *Nature*, 465(n7301_supp):S4–S5, 2010.

[73] Carlos A Buscaglia, Vanina A Campo, Alberto CC Frasch, and Javier M Di Noia. Trypanosoma cruzi surface mucins: host-dependent coat diversity. *Nature Reviews Microbiology*, 4(3):229–236, 2006.

[74] A. Frazee, B. Langmead, and J. Leek. Recount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC bioinformatics*, 12:449, 2011.