Phd Dissertation Prospectus
Applied Mathematics and Scientific Computation
Exploring, Quantitatively Describing, and Normalizing RNA-seq data with Order Statistics

Kwame Okrah
Advisor:
Héctor Corrada Bravo

April 1, 2014

Assays based on massively parallel next-generation sequencing platforms have become the technology of choice for a large variety of transcriptome studies in recent years due to its decreasing cost and measurement advantages over microarray platforms [8, 13]. These technological improvements in measurements have been accompanied by the development of new algorithms and statistical methodologies to analyze the data that they produce. Chief among these are methods 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 distribution, for example DESeq [1] and edgeR [9]; and (2) those that are based on assuming that statistics based on log-transformed counts follow a Gaussian distribution, for example the voom [6] transformation in limma [10]. Some of the pros and cons of these two frameworks have been reported in [11].

Coupled with the efficiency and accuracy in RNA-seq technology is the complexity of experimental designs. In particular host-pathogen experiments. In these experiments an investigator will usually infect a host cell (for example mammalian macrophages) with a parasite (for example the protozoan *T. cruzi*). Samples of the infected host cells are collected at various stages of the infection cycle and sequenced. The goal of these experiments is to study how the host and pathogen transcriptomes interact in-vitro. And perhaps provide leads for credible drug targets that disrupt important pathways in the parasite while minimizing harmful side effects in the host's cell. Host-pathogen RNA-seq data presents many new challenges, in particular with respect to normalization. The natural discrepancies between the size of the host and parasitic transcriptomes lead to data sets that are difficult to normalize. The usual assumption of proportionality of gene expression levels between biological replicates does not necessarily hold [7], and as such one needs a more sophisticated method of normalization.

1 Tukey's g-and-h distribution as an objective framework for describing robustness in RNA-seq data

Tukey's g-and-h distribution [12] is a simple yet flexible distribution whose two shape-parameters can be used to accurately describe the skewness (g parameter) and elongation (h parameter) of a given dataset or theoretical distribution. The key advantage of this framework over the traditional method of using sample moments to describe shape, is in the simplicity of the model and the nice statistical properties of the estimation procedure described by [3]. Tukey's g-and-h distribution is defined as follows: Let $Z \sim N(0,1)$ and consider the transformation:

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

where g and h are any real numbers. The random variable $\mathrm{T}_{g,h}(Z)$ is said to have the standard Tukey's g-and-h distribution. The g parameter controls skewness in both magnitude and direction. When $g \rightarrow 0$ 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 ($h > 0$) or less ($h < 0$) weight is at the tails (kurtosis) of a distribution. Finally, if we scale $\mathrm{T}_{g,h}(Z)$ by B (a positive number) and shift by A (any real number) we obtain the more general g-and-h distribution which we shall denote by $\mathrm{GH}(g,h,A,B)$.

To estimate the g and h parameters we will follow David C. Hoaglin's quantile method [3]. To begin, let X denote a sample assumed to be generated from $\mathrm{GH}(g,h,A,B)$. The estimation procedure will be in two steps, first we will estimate g and A. And then conditioned on the g estimate we will estimate h and B. A is estimated with the median of X, and denoted by $X_{0.5}$. To estimate g, we select the quantiles $\{X_p : p = 0.5^{k+1}, k = 1,2,\ldots,n\}$, where $n = \lfloor \log_2(\text{sample size of X}) \rfloor$. And construct the g_p which exactly gives X_p and X_{1-p}. This is given by:

$$g_p = -\frac{1}{z_p} \log\left(\frac{X_{1-p} - X_{0.5}}{X_{0.5} - X_p}\right). \tag{2}$$

The selection of these quantiles places emphasis on the tail behavior of X. We estimate g as follows:

$$\hat{g} = \operatorname{median}\{g_p : p = 0.5^{k+1}, k = 1,2,\ldots,n\}. \tag{3}$$

This provides us with a single summary of skewness that is outlier resistant. Given an estimated $\hat{g}$, and for some $p < 0.5$ it can be shown that:

$$\log(B) + h\frac{z_p^2}{2} = \log\left(\hat{g}\frac{X_{1-p} - X_p}{\exp(-\hat{g}z_p) - \exp(\hat{g}z_p)}\right). \tag{4}$$

We regress the right hand side of equation (4) on the left hand side. Where $log(B)$ is the intercept and h is the slope. The least squares estimate, $\hat{h}$, is used to estimate h, and B is estimated as $\hat{B} = exp(\widehat{log(B)})$, where $\widehat{log(B)}$ is the least squared estimate of the intercept.

Statistical methods predicated on the strong assumption of normality often produce simple and interpretable statistics with nice properties. Even in cases where the Gaussian assumption fails, these methodologies can provide satisfactory results, even for relatively small samples, after certain adjustments have been made. A prime example is the computation of a more stable variance estimate by borrowing information across genes via Empirical Bayesian methods [10].

Using Tukey's g-and-h we provide an objective framework to characterize the robustness of these methods for analyzing RNA-seq data. In particular we looked at the t-test. We show this in two steps: (1) For a given g and h parameter we compute the power of the t-test at various levels of fold-change. Based on this we can characterize how much power is lost when we deviate from the ideal Gaussian assumptions ($g \to 0, h = 0$). (2) Given an RNA-seq dataset we log transform the data and fit each gene under the same biological condition with a g and h estimate. Depending on where these estimates fall we can get a sense of how powerful t-test based methods will perform. We obsereve that using a log transformation of the counts data and assuming a Gaussian distribution maintains a reasonable amount of power (even without any corrections for the mean-variance trend). We also use the g-and-h distribution to provide evidence in support of the negative binomial distribution assumption of RNA-seq data.

2 Normalizing RNA-seq mixture distributions via L-moments

Host-pathogen RNA-seq data present a significant challenge when it comes to normalization. Currently the methods for normalizing are based on scaling each sample according to a certain criterion [1, 9] or quantile based methods that restrict each sample in the data set to have the same quantiles [2].

The scaling methods are based on the assumption that for biological replicates expression levels for all genes are related by one scalar. However it has been shown that when the library sizes of biological replicates are substantially different this is assumption fails to hold [7]. This is particularly true in the host-pathogen data that we observe. The quantile method is based on the assumption that transcriptomes regardless of the biological conditions should have the same distribution. While this is true for some experiments it does not hold generally [7].

Based on these observations we seek to find a normalization method that is robust to these assumptions. The L-moments of a random variable X are defined as linear combination of order statistics [4]. L-moments fully characterize a random variable [4] and play a similar role as the regular moments. However, L-moments enjoy certain theoretical and practical advantages over regular moments. In particular all L-moments exist if and only if the first moment is finite [4]. Also compared to sample moments, sample L-moments have lower variances and are more robust against outliers [5]. L-moments are defined as follows:

$$\lambda_r = r^{-1}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}\mathrm{E}(X_{r-k:r}); \quad r = 1, 2, \ldots \tag{5}$$

where $X_{k:n}$ is the k^{th}-order statistic from a conceptual sample of size n from X. If the X is continuous then

$$\lambda_r = r^{-1}\int_0^1 Q(u)P^*_{r-1}(u)du; \quad r = 1, 2, \ldots \tag{6}$$

where

$$P^*_r(u) = \sum_{k=0}^{r}(-1)^{r-k}\binom{r}{k}\binom{r+k}{k}u^k, \tag{7}$$

and $Q(u)$ is the quantile function of X. $P^*_r(u)$ is the r^{th} shifted Legendre polynomial. Karvanen [5] defines quantile mixtures, $Q(u) = \sum_{i=1}^{m} a_i Q_i(u)$, where Q_i are the quantile-components of the mixture and the a_is are selected in order to make $Q(u)$ is well-defined. Based on this definition, then the L-moments of $Q(u)$ are obtained as follows:

$$\lambda_r = \sum_{i=1}^{m} a_i\lambda_{ir}; \quad r = 1, 2, \ldots \tag{8}$$

where λ_{ir} are the L-moments of the mixture components, $Q_i(u)$. This setup provides natural estimates for the weight a_i.

We propose to use quantile mixtures to describe the quantile function for each sample in a given RNA-seq dataset. Based on these estimates we can provide statistical tests for checking the equal transtiptome assumption. Also based on these estimates we will propose a normalization method that will constrain certain quantiles of interest, thereby providing method that is experiment dependent. Finally we will look at the relationship between our g and h estimates and the first four L-moments since both statistics describe the same features of a dataset. It will be interesting to see how the interact since both methods are bases on order statistics.

References

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

[2] B. Bolstad, R. Irizarry, M. Astrand, and T. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185–193, 2003.

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

[4] J. Hosking. L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37:105–124, 1990.

[5] J. Karvanen. Estimation of quantile mixtures via l-moments and trimmed l-moments. *Computational Statistics and Data Analysis*, 51:947–959, 2006.

[6] 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.

[7] J. Loven, D. Orlando, A. Sigova, C. Lin, P. Rahl, et al. Revisiting global gene expression analysis. *Cell*, 151:476, 2012.

[8] 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.

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

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

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

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

[13] X. Xu, Y. Zhang, J. Williams, E. Antoniou, W. McCombie, S. Wu, W. Zhu, N. Davidson, P. Denoya, and E. Li. Parallel comparison of illumina rna-seq and affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated ht-29 colon cancer cells and simulated datasets. *BMC Bioinformatics*, 14:S1, 2013.