Conference on Applied Statistics in Agriculture | 2012 - 24th Annual Conference Proceedings

THE NUANCES OF STATISTICALLY ANALYZING NEXT-GENERATION SEQUENCING DATA

Sanvesh Srivastava

R. W. Doerge

The Nuances of Statistically Analyzing Next-Generation Sequencing Data

Sanvesh Srivastava and R.W. Doerge*

Department of Statistics, Purdue University, West Lafayette, IN 47907

Abstract

High-throughput sequencing technologies, in particular next-generation sequencing (NGS) technologies, have emerged as the preferred approach for exploring both gene function and pathway organization. Data from NGS technologies pose new computational and statistical challenges because of their massive size, limited replicate information, large number of genes (high-dimensionality), and discrete form. They are more complex than data from previous high-throughput technologies such as microarrays. In this work we focus on the statistical issues in analyzing and modeling NGS data for selecting genes suitable for further exploration and present a brief review of the relevant statistical methods. We discuss visualization methods to assess the suitability of statistical models for these data, statistical methods for modeling differential gene expression, and methods for checking goodness of fit of the models for NGS data. We also outline areas for further research, especially in the computational, statistical, and visualization aspects of such data.

Keywords: Clustering, differential gene expression, dimension reduction, generalized linear models, hierarchical Bayesian modeling, microarrays, next-generation sequencing, negative binomial distribution, Poisson distribution, residual plots.

1. Introduction

1.1 Motivation

Next-generation sequencing (NGS) technologies (Hayden, 2009; Metzker, 2009; Ng et al., 2010; Roach et al., 2010) are increasingly used for exploring the genome, epigenome, and transcriptome. These technologies play an important role in understanding cell organization and functionality. Unlike data from previous technologies (e.g., microarrays), data from NGS technologies are highly replicable with little technical variation (Marioni et al., 2008; Mortazavi et al., 2008; Bullard et al., 2010; Hansen et al., 2012a). It is also possible to measure alternative isoform regulation (Wang et al., 2008) and to discover novel transcription regions in the genome (Mortazavi et al., 2008) using NGS technologies.

Despite their advantages, NGS technologies are still expensive, and similar to other high-throughput data, NGS data are high-dimensional with a limited number of samples (i.e., individuals) compared to the number of predictors (i.e., genes); a problem known as “big p small n” that first gained notice with the analysis of microarray data. Toward this end, and similar to microarray data, the limited number of biological replicates challenges the reliability of the statistical inference. As a remedy for microarray analysis with few biological replicates Efron (2010) recommended taking advantage of the rich genetic information and borrowing information across genes to compensate for the limited information about within group variation. Comparatively, NGS technologies have different sources of bias that are not present in earlier high-throughput technologies such as microarrays, and the issues in NGS data analysis are

magnified simply because of the discrete nature of the data, as well as the complexity and massive size of the data (Bullard et al., 2010). These issues must be addressed by computational algorithms and statistical methods used for the analysis of NGS data if one wishes to draw reliable conclusions from these data. Hansen et al. (2012a) point out that ignoring these issues results in the biological variability not being estimated accurately, which in turn imposes serious restrictions on their reproducibility. And, finally, there is still no standard practice for NGS applications that allows for design of experiments, preprocessing, normalization, or summarization of raw data before their statistical analysis (Auer et al., 2011).

Probably the most common use of NGS technologies to date is for identifying differentially expressed genes, which include genes that may be associated with a response of interest (e.g., disease status). Although this application is referred to as RNA-sequencing (RNA-seq), the computational and statistical issues in analyzing other NGS data are similar to RNA-seq data (Auer et al., 2011). We will use RNA-seq data as an example to illustrate both the statistical inference framework for analyzing NGS data, as well as its limitations. Because the model assumptions in NGS data analysis are not always justified (Auer et al., 2011), we also discuss visualization methods to assess the suitability of statistical models and their goodness of fit.

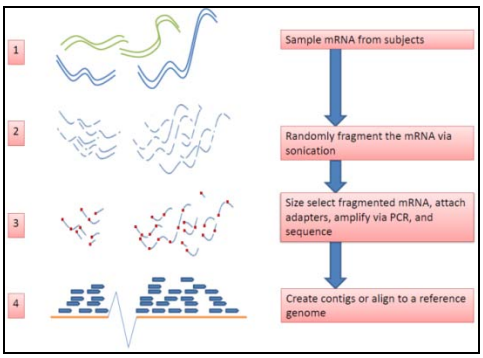
1.2. Experimental Design, Data Preprocessing, Normalization, And Summarization

Microarray data were the first example of high-throughput genomic data. The issues related to their statistical design and analysis are well-understood (Churchill, 2002; Yang and Speed, 2002; Efron, 2010). We will briefly compare microarray and NGS data, which will be useful later in

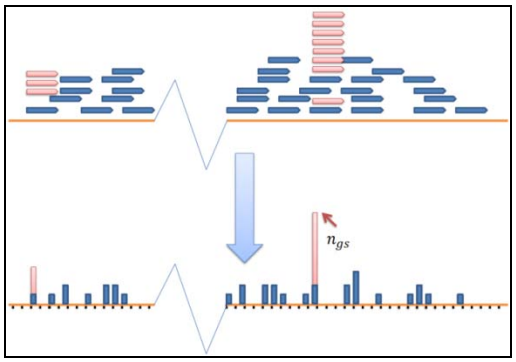
understanding the statistical inference framework and challenges in analyzing NGS data. From a technological perspective, results obtained from NGS technologies and microarrays agree strongly (Cloonan et al., 2008; Mortazavi et al., 2008; Sultan et al., 2008; Fu et al., 2009; Bradford et al., 2010). This said, researchers are switching from microarrays to NGS technologies mainly because of lower technical variation, ability to discover novel genes and epigenetic modifications, and for de novo sequencing (‘t Hoen et al., 2008; Marioni et al., 2008; Liu et al., 2011; Hansen et al., 2012a; Young et al., 2012). From a statistical perspective, the theory of multiple hypotheses testing (Efron et al., 2001), variable selection (Zou and Hastie, 2005; Friedman et al., 2010), and the use of false discovery rates (FDR) for multiple testing problems (Benjamini and Hochberg, 1995; Storey, 2003; Efron, 2010), which were motivated by microarray data, also apply to NGS data analysis with modifications to acknowledge the data-specific features (Oshlack and Wakefield, 2009; Anders and Huber, 2010; Oshlack et al., 2010; Hardcastle and Kelly, 2010; Robinson and Oshlack, 2010; Young et al., 2010; Auer and Doerge, 2011). Similar to microarrays, proper experimental design for NGS experiments is essential if one wishes to reliably answer scientific questions. The principles of blocking, randomization, and replication, proposed by Fisher (1935), are still important for NGS technologies (Auer and Doerge, 2010). Ignoring any of these principles limits the scope and applicability of the results from NGS experiments (Auer et al., 2011; Hansen et al., 2012a). Young et al. (2012) provide an excellent overview of the comparisons between microarray and NGS technologies.

The features that distinguish NGS technologies from microarrays are the preprocessing steps prior to the analysis of the raw data for differential expression. Figure 1 (a) illustrates the workflow for RNA-seq data alignment and summarization. First, RNAs are isolated from the

genomic material obtained from the samples and broken into short fragments. They are then PCR enriched so that the NGS technology can sequence or read them. These short reads are then aligned to a reference genome. The number of reads that map to a gene is its digital gene expression. Similar to microarrays, the raw data are summarized as an expression matrix and its rows represent the genes and columns correspond to samples (for more details see: Auer, 2010; Young et al., 2012). The alignment process is not perfect. Figure 1 (b) shows the observed gene expression values with summarization errors and technological bias due to misalignment and over-enrichment. Recently, another source of dependency for gene expression measurements has been documented: GC content, which is the proportion of G and C nucleotides in a gene (Benjamini and Speed, 2012; Hansen et al. 2012b). In addition to the technology-specific biases, expression measurements are also affected by the differences in PCR enrichment steps prior to sequencing for different experiments. These biases make normalization of the raw expression data essential. Normalization controls for the biases that confound the biological effects, ensures that NGS data from different platforms and experiments are comparable, and minimizes the chances of inferring a gene as differentially expressed simply because of technological artifacts (Hansen et al. 2012b; Young et al., 2012). These issues are similar to the normalization in microarrays (Wu et al., 2004), but more complex because of the different sources of bias mentioned before. Anders and Huber (2010), Bullard et al. (2010), and Robinson and Oshlack (2010) provide a comprehensive overview of the technology related biases and errors and propose methods to control for them. Benjamini and Speed (2012) and Hansen et al. (2012b) show the presence of GC content bias, propose normalization methods, and provide software to control for this bias.



(a)



(b)

Figure 1. Workflow for RNA-seq data alignment and summarization. (a) RNAs from the samples are isolated and broken into small fragments by sonication. These fragments are PCR enriched and then sequenced or read by the NGS technology. The reads are in form of nucleotide sequences and are aligned to the nucleotide sequences of the reference genome. The number of aligned reads is the measurement of the expression of a gene. (b) The alignment process shown in (a) is not perfect. Blue rectangles are the reads that were actually transcribed by the genes and the pink rectangles are the misaligned and over-enriched reads. The total number of reads that mapped to the gene *g* in sample *s*, including the alignment errors, n_{gs} is its digital gene expression (Baumann, 2012).

2. Statistical Approaches For Differential Gene Expression Analysis

Typically, after the alignment, summarization, and normalization steps discussed previously, the expression data from NGS technologies are a matrix of discrete gene counts, with rows representing the genes and columns corresponding to the samples. The expression matrix

represents the relative amount of expression of each gene across different samples. The major theme in NGS experiments is to identify differentially expressed genes for the purpose of understanding complex biological systems. The main issue is to obtain reliable inference with a limited number of samples, that is, accounting for high-dimensionality, which is accomplished by information borrowing among genes to compensate for limited availability of samples (Efron, 2008). This similar issue is accomplished in microarray data analysis using empirical Bayesian (Efron, 2010), fully Bayesian (Baldi and Long, 2001; Ibrahim et al., 2002), and penalized-likelihood based approaches (Tibshirani et al., 2005; Ma and Huang, 2007). Although some of these approaches translate to differential gene expression analysis in NGS data, associated problems remain, and as a result we are motivated to suggest methods to control for some of these problems.

2.1 Gene-Wise Hypotheses Testing For Differential Gene Expression Analysis

For general notation, we allow the observed data to be a matrix of gene counts, $\boldsymbol{N}$. Further, for simplicity and ease of illustration we assume a balanced design with two treatments, and S samples and G genes in each treatment group. The methods discussed here are also applicable to more complicated designs with slight modifications. Let n_{gst} be the observed count of gene g in the sample s with treatment t, and θ_{gt} is the expected value of n_{gst}. The library size of a particular treatment group t is defined as the total gene count in the original population from which we sample the gene counts, and is not known a priori. The gene counts depend on the library size simply because a large library size implies higher gene counts. Therefore, μ_{gt}

denotes the mean expression of gene *g* in treatment *t* normalized for library size, and it is comparable across treatment groups. Our aim is similar to that of microarray analysis, to obtain test statistics or posterior distributions, for testing gene-wise differential expression, $\mu_{g2} - \mu_{g1}$, between the two treatment groups. However, the sampling distribution of the test statistic for NGS data analysis is difficult to obtain without restrictive distributional assumptions.

When analyzing microarray data it is quite common to assume the data follow Gaussian distribution after normalization and log transformation. As such, for each gene, the test statistic for differential expression follows a *t*-distribution with *(2S-2)* degrees of freedom (Efron et al., 2001; Casella and Berger, 2001; Smyth, 2004). Further, to compensate for unreliable variance estimates due to small sample size, many empirical Bayesian approaches make the inference more robust by modifying the gene-wise variances appropriately using a correction obtained by borrowing information across all genes (for more details see: Smyth, 2004; Smyth 2005). The p-values obtained from gene-wise hypothesis tests are then adjusted using the FDR procedure (Benjamini and Hochberg, 1995), and the genes having p-values below a prespecified cutoff (say, 0.05) are declared as differentially expressed. This procedure is very robust in practice (Haury et al., 2011).

Although the above framework is robust and simple for microarray data analysis, there are many statistical issues that need to be addressed for NGS data analysis. First, NGS data are discrete and there is no equivalent theoretically justifiable sampling distribution for $\mu_{g2} - \mu_{g1}$ similar to a *t*-statistic (‘t Hoen et al., 2008; Cloonan et al., 2008; Robinson and Oshlack, 2010). Second, and

related to the first issue, the distribution of the test statistic for testing $\mu_{g2} - \mu_{g1} = 0$ is determined by the asymptotic distribution of the likelihood or quasi-likelihood (Robinson and Smyth, 2007, 2008; Anders and Huber, 2010; Auer and Doerge, 2011). Third, due to overdispersion, small counts, and zero inflation, which are very common in NGS data, the assumption of a Poisson distribution for gene counts is not universally justified (Vêncio et al., 2004; Thygesen and Zwinderman, 2006; Hardcastle and Kelly, 2010; Auer and Doerge, 2011). The last point has led to the use of the negative binomial distribution for modeling NGS data. The methods for determining differentially expressed genes using the exact test and the generalized linear mixed model for negative binomial distribution are available as R/Bioconductor packages (Gentleman et al., 2004; R Development Core Team, 2012), namely, *baySeq* (Hardcastle and Kelly, 2010), *DESeq* (Anders and Huber, 2010), and *edgeR* (Robinson et al., 2010). In this work we use *edgeR* to illustrate the application and limitations of the aforementioned methods. For differential gene expression analysis, two types of tests are available in *edgeR*. These tests represent the most widely used methods: the exact negative binomial test (exact NB test) (Robinson and Smyth, 2007, 2008) and the likelihood ratio test based on negative binomial distribution (NBLRT) (Agresti, 2002; McCarthy et al., 2012).

The exact NB test is limited to simple experimental designs for testing differential expression between two or more treatment groups, but it has the advantage of using less restrictive distributional assumptions. Specifically, the form of negative binomial distribution used by *edgeR* is

$$P[N = n_{gst} \mid \mu_{gt}, \phi_g] = \frac{\Gamma\left(n_{gst} + \phi_g^{-1}\right)}{\Gamma\left(\phi_g^{-1}\right) + \Gamma\left(n_{gst} + 1\right)} \left(\frac{1}{1 + \mu_{gt}\phi_g}\right)^{\phi_g^{-1}} \left(\frac{\mu_{gt}}{\phi_g^{-1} + \mu_{gt}}\right)^{n}, \quad (1)$$

which has the expected value μ_{gt}, variance $\mu_{gt} + \phi_g \mu_{gt}^2$, and dispersion parameter ϕ_g for gene *g* in treatment *t*. Under the Poisson assumption, both mean and variance are μ_{gt}, which is a reasonable model for estimating technical variation in NGS data (Marioni et al., 2008). Therefore, (1) implies that the variation in NGS data can be decomposed into technical variation (μ_{gt}) and biological variation ($\phi_g \mu_{gt}^2$), with a quadratic dependence of the biological variation on the technical variation. This assumption may not be justified in general. *edgeR* extends the Fisher's exact test to the exact NB test using (1) and uses negative binomial probabilities to calculate the p-value for testing $\mu_{g2} - \mu_{g1} = 0$ for each gene (Agresti, 2002). Since this version does not borrow information among genes, Robinson and Smyth (2008) recommend a moderated extension of this test that uses a modified overdispersion parameter estimate (ϕ_{wg}), which is the weighted average of overdispersion parameter estimates under two extreme scenarios. First, each gene *g* has its own overdispersion parameter (ϕ_g). Second, each gene has the same overdispersion parameter (ϕ_0). Although moderation is an appropriate modification to achieve robustness and reliability for small sample size NGS experiments, there is no standard way of choosing the weights to determine (ϕ_{wg}), and it still remains to be shown that its performance is uniformly better than moderated *t*-test of *limma* if we simply assume that NGS data have log-normal distribution (Smyth 2004; Symth, 2005; Bullard et al., 2010; Robinson et al., 2010).

By focusing on the quadratic dependence of the mean and variance that is assumed in (1), we demonstrate via simulation a restriction that can lead to inflated FDR. Specifically, we first randomly generate μ_{gt} and ϕ_g for each gene, and then simulate the gene counts n_{gs} using (1). This procedure ensures that the simulated gene counts have more biological variation than what is assumed in (1). Using this procedure, we simulate count data (RNA-seq data) for 2 treatment groups, each with 1000 genes and 3 biological samples.

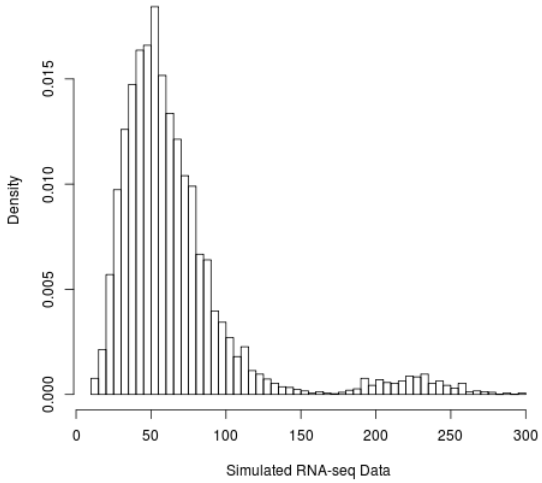


Figure 2. Simulated count data. Count data are simulated from a mixture of negative binomial distributions for 2 treatment groups, each with 1000 genes and 3 biological samples. The first 100 genes in treatment 2 are assumed to be differentially expressed, and they separate visually as the lower peak in the histogram.

We use an alternative parameterization of (1) that defines negative binomial random variable as the number of failures (n_{gst}) which occur in a sequence of Bernoulli trails (with probability p_{gt}) before a target number of successes (m) is reached (R Development Core Team, 2012), with

$p_{gt} = \frac{m}{m + \mu_{gt}}$ and $m = \phi_g^{-1}$. We further assume that 10% of the genes are differentially expressed to make the simulation study close to reality, and set the first 100 genes in treatment 2 to have *p = 0.4* and *m = 150*. The remaining genes have m and p that are generated from Uniform (50, 120) and Beta (30, 20). Using these parameters we generate RNA-seq data from a negative binomial distribution for the 1000 genes in 3 samples and 2 treatment groups, and it matches the simulation settings of Robinson et al. (2010). Figure 2 illustrates the histogram of the simulated RNA-seq data. The counts of differentially and non-differentially expressed genes are separated as two modes. The application of exact NB test and controlling for FDR at 5% results in detection of 113 differentially expressed genes and 20 of these are false positives. This implies a false discovery proportion (FDP) of 17.6%. (In real experiments, it is impossible to calculate the true FDP because the truth is unknown, but the expectation of FDP, FDR, can be calculated. We will assume that the conditions for ensuring that FDP is a reasonable approximation of FDR are justified in our discussions further.) The FDP is much larger than its expected value 5%. This simulation illustrates that in situations where the variation of NGS data are more than that modeled by a negative binomial distribution, the true number of false discoveries could be much higher than expected. Therefore, despite its less-restrictive distributional assumptions, the results of exact NB test can be misleading and can potentially lead to a large number of false positives.

The NBLRT based test in *edgeR* (McCarthy et al., 2012) can be applied to data from complex experimental designs (e.g., time course biological experiments), and can account for technology-specific sources of variations (e.g., lane effects, batch effects, and library preparation methods; Young et al., 2012), but it has restrictive distributional assumptions that rely on the asymptotic

likelihood or quasi-likelihood-based test statistics. Presently, these asymptotic distributions are hard to justify for NGS data because the sample sizes are small (Auer and Doerge, 2010; Auer and Doerge, 2011; Auer et al., 2011; Hansen et al., 2012a). For the simulated data in Figure 2, the NBLRT test identifies 112 differentially expressed genes at FDR control of 5%, and 21 of these genes are false positives. Therefore, the NBLRT test has a FDP of 18.7%, which is much higher than its expected value of 5%. Again, this illustrates that using asymptotic approximations can lead to underestimation of the true variability in the data, which in turn leads to underestimation of false discoveries.

We expect the results from the analyses of real NGS data to be worse, when compared to simulated investigations, for multiple reasons. First, there is a high chance that the variability in NGS data is underrepresented by a negative binomial distribution or by the asymptotic distributions, which leads to overly optimistic results and number of false discoveries. Second, in real data the modes of the differentially and non-differentially expressed gene counts are not well separated as in Figure 2, therefore it would be harder for the exact NB or NBLRT tests to discriminate between differentially and non-differentially expressed genes, and this will lead to lower power. Third, technology-specific biases such as alignment errors, GC content, and library preparation methods further complicate and confound the differential gene expression analysis of NGS data and make the results based on parametric and asymptotic assumptions unreliable, unless these assumptions are verified for the data. These issues, with proposed solutions for controlling biases and errors, are discussed in greater detail in Bullard et al. (2010), Langmead et al. (2010), Oshlack et al. (2010), Robinson and Oshlack (2010), Benjamini and Speed (2012), Hansen et al. (2012b), Young et al. (2012).

2.2 Visualization For Checking Model Assumptions And Goodness Of Fit

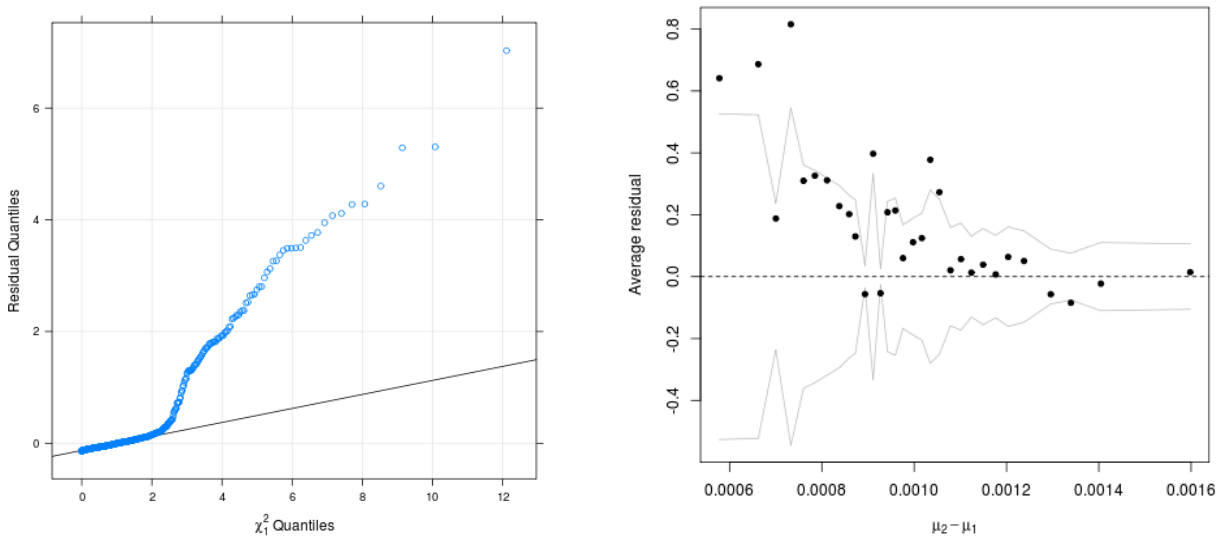
As mentioned earlier, NGS data rarely satisfy the model assumptions because of various sources of bias and variation. Therefore, it is important to check the model assumptions and goodness of fit. Visualization is an effective way to accomplish these two goals (Cleveland, 1993; Gelman, 2004). MA and volcano plots (Yang et al., 2002) were developed for microarray data analysis to visualize any systematic patterns in the data before and after normalization and for visualizing the results of differential gene expression analysis. Because microarray data are continuous and the *t*-test or *F*-test are robust for selecting differentially expressed genes (Haury et al., 2011), these plots are not valid for illustrating lack of fit or checking for validity of distributional assumptions. MA plots have been widely used to visualize the results of NGS data analysis. Although useful for checking any systematic patterns in NGS data, MA plots fail to give any information about the validity of model assumptions and goodness of fit. Further, the discreteness of the data makes it harder to visualize any systematic patterns or biases. Therefore, similar to Gelman (2004), more visual and numerical summary methods need to be developed for model checking.

For NGS data analysis, there are two important assumptions that require verification: the parametric distributional assumption of the data and the goodness of fit of the model. The residuals obtained after fitting a model to NGS data are the best choice for checking these two assumptions (Cleveland, 1993). We propose two methods using QQ plot (Cleveland, 1994) and binned plot Gelman (2004), respectively, for residuals to check these assumptions and illustrate

their application using the fit from *edgeR* in Section 2.1. The residuals obtained after fitting the NBLRT model to the simulated data follow a χ_1^2 distribution (Agresti, 2002; McCarthy et al., 2012). Figure 3 (a) shows the QQ plot of the residuals against quantiles from the χ_1^2 distribution. If the distributional assumptions of the model are justified, then the points on QQ plot would be close the black line (drawn through the 25% and 75% quantiles of the residuals). The plot shows that the higher quantiles (right tail) of the residual distribution are much larger than that predicted by the χ_1^2 distribution, which implies that the variability of the residuals is larger and the tails are heavier than that predicted by the NBLRT model. Therefore, the modeling assumption that the count data follow negative binomial is not justifiable. This agrees with our simulation setting in which the NGS data are simulated from a mixture of negative binomial distribution. The QQ plot also shows lack of fit, but it does not illustrate details about the data that lead to the lack of fit.

Because NGS data are discrete, it is harder to identify systematic patterns and departures from model assumptions on the residuals. A binned residual plot (Gelman, 2004) serves the purpose of showing the goodness of fit in greater detail and solves the visualization problem because of discreteness of the residuals by smoothing. The plot averages the residuals that are grouped together in pre-specified bins and then plots the averaged residuals against the predicted values. The size of the bins is crucial in showing the model fit (for detail see: Gelman and Hill, 2007, page 97). The *arm* package in R (Gelman et al., 2012) provides a function *binnedplot* for drawing binned residual plots and also chooses the optimal number of bins that effectively convey the model fit without over-smoothing. Figure 3 (b) shows the binned residual plot obtained from the *edgeR* NBLRT fit in Section 2.1. It shows that the residuals are large for small

values of $\mu_{g2} - \mu_{g1}$ and are reasonably close to 0 for large values of $\mu_{g2} - \mu_{g1}$. In our earlier simulation setting, small values of $\mu_{g2} - \mu_{g1}$ correspond to the non-differentially expressed genes and large values of $\mu_{g2} - \mu_{g1}$ correspond to the differentially expressed genes. Therefore, the model fit is good for differentially expressed genes and there is lack of fit for the non-differentially expressed genes. Further, this also explains the high power and high FDPs for the exact NB and NBLRT tests. These tests are able to identify more than 90% of the differentially expressed genes because of the good model fit for these genes (with high values of $\mu_{g2} - \mu_{g1}$). The genes with smaller $\mu_{g2} - \mu_{g1}$ are not differentially expressed, but the observed value of the test statistic greatly differs from the null model due to lack of fit, therefore these genes are falsely rejected as discoveries in the exact NB and NBLRT tests (thereby leading to high FDPs). The binned residual plot also shows that the residuals decay exponentially from low to high values of $\mu_{g2} - \mu_{g1}$, which suggests that including an extra term in the NBLRT model for this systematic pattern would resolve this lack of fit. We, however, do not pursue this modification further. Our analysis shows that it is easier to identify the lack of fit and to propose modifications to the model for improving the model fit using binned residual plots compared to the QQ plot or p-value based summary methods. The simulation also shows that if the signal strength is large (here, large values of $\mu_{g2} - \mu_{g1}$), then methods like exact NB and NBLRT tests are good choices for detecting differentially expressed genes. However, in situations where the signal is not strong (here, small values of $\mu_{g2} - \mu_{g1}$), these tests can lead to high FDP (and FDR) and report overly optimistic results for differentially expressed genes. At present, the signal for differential expression in NGS data is low and will likely to remain so until the technology decreases in price and more samples are assessed, therefore model checking is essential in NGS data analysis.



(a) **(b)**

Figure 3. Visualization methods for checking distributional assumptions and goodness of fit for the negative binomial model used in *edgeR* (Section 2.1). (a) QQ plot of the residuals against quantiles of χ^2_1 distribution shows that the higher quantiles (right tail) of the residuals are heavier than the tail of the χ^2_1 distribution, which is the distribution of the residuals under the model assumptions. Therefore, there is a lack of fit, and the model assumptions are not justified for the simulated data. (b) Binned residual plot shows the lack of fit for the model. The grey lines correspond to the 95% confidence interval (CI) for the residuals (Gelman and Hill, 2007, page 97). The residuals for smaller value of $\mu_{g2} - \mu_{g1}$ are much higher than 0 and lie outside the CI. This illustrates that the predictions of $\mu_{g2} - \mu_{g1}$ from negative binomial model are lower than the observed values in those regions. The residuals for higher values of $\mu_{g2} - \mu_{g1}$ are close to 0 showing no systematic departures from model assumptions and fit. The plot also has an exponentially decaying pattern from large to small residuals along the x-axis. These

observations confirm that the variability in the data is underrepresented by a negative binomial distribution, and thus agreeing with the simulation settings.

2.3 Penalized Likelihood Methods For Gene Selection

Penalized likelihood based approaches like lasso and elastic net have been widely used in microarray data analysis for selecting genes suitable for further exploration (Tibshirani, 1996; Tibshirani et al., 2005; Zou and Hastie, 2005; Ma and Huang, 2007; Friedman et al., 2010). Although these methods have been used for performing classification and clustering of NGS data (Witten, 2011), their use for gene selection in NGS data applications is fairly limited. We use the *glmnet* algorithm (Friedman et al., 2010) to select genes that are predictive of the treatment group association based on their digital gene expression from the expression matrix, an objective similar to, but not identical to, differential gene expression analysis. Because there are two treatment groups (1 and 2) in the simulation, we use the *glmnet* algorithm for the *binomial* family (the treatment group 2 is assumed to correspond to the "success" outcome, that is, we will select genes that are predictive of treatment group 2) with the elastic net penalty and apply it to the simulated data of Figure 2. The elastic net penalty has the advantage of selecting all the correlated genes with high probability (Friedman et al., 2010). The algorithm selects 82 genes and 9 are false positive, implying a FDP of 10%, which is lower than the FDPs for the tests in *edgeR*. Although determining FDR in applications of penalized likelihood approaches like *glmnet* still remains an active area of research, the *glmnet* algorithm has been successfully used in microarray and other genomic studies because of its computational efficiency and theoretically justifiable asymptotic variable selection properties of the lasso and its extensions (Friedman et al.

2010; Bühlmann, and van de Geer, 2011). NGS data offer an attractive opportunity for further applications of *glmnet* algorithm because of its computational efficiency and ability to handle massive data. However, more research is needed to explore the theoretical and practical applicability of penalized likelihood approaches for selecting genes in NGS data analysis, and adapting existing methods in microarray data analysis for complex experimental designs such as time course experiments (Meier and Bühlmann, 2007).

3. Hierarchical Bayesian Modeling Of NGS Data

Although methods for determining differential gene expression are important for selecting genes suitable for further exploration, it is useful to develop modeling approaches that adapt to the underlying latent structure of NGS data and that help in the understanding of complex biological systems in an interpretable manner. Approaches that are very helpful for exploring the structure in microarray data, such as the gene-shaving algorithm (Hastie et al., 2000), linear discriminant analysis (Dudoit et al., 2002; Guo et al., 2007), and nearest shrunken centroids algorithm (Tibshirani et al., 2003), have no equivalents for NGS data. Hierarchical Bayesian approaches (Gelman et al., 2003) provide a natural way of modeling the generative mechanism of NGS data and combine the hierarchy in the sampled population, uncertainty about the underlying model and unknown parameters, and prior experimental knowledge in a flexible and interpretable manner; these approaches are also extensible and modular. Because of the complex nature and massive size of NGS data, exact Bayesian inference becomes computationally intractable. Approximate Bayesian inference techniques in machine learning, such as variational inference

and expectation propagation (Bishop, 2006), provide a balance between theoretical and computational ideas to address the challenges in the Bayesian modeling of NGS data.

Recently, Srivastava and Doerge (2011) proposed a novel probabilistic approach for unsupervised modeling of high-dimensional count data, including NGS data. Their approach has two stages, and in the first stage they adapt the Latent Dirichlet Allocation (LDA) framework (Blei et al., 2003) and extend the Latent Process Decomposition (LPD) framework (Rogers et al., 2005) for modeling overall patterns in the high-dimensional count data. Specifically, the first stage of the extended LPD framework is a three-level hierarchical Bayesian model and employs the Poisson variational method from machine learning for parameter estimation, and then uses the estimated parameters in the second stage to extract feature-subsets using a permutation based method, namely the Gap algorithm (Hastie et al., 2000), from classical statistics. In NGS data applications, LPD obtains gene-subsets that explain a large portion of variability in the data, and that have similar expression patterns among the members of any gene-subset. The fundamental concept behind the modeling strategy is that a small fraction of genes, organized into groups, are responsible for a significant amount of biological variation. LPD is a special case of mixed membership models (Airoldi et al., 2005) and is more flexible than classical clustering methods, such as hierarchical clustering. Due to the biological motivation behind the model, LPD provides interpretable parameter estimates that other clustering-based approaches cannot.

As an adaptation of the LDA framework, the real power of the LPD framework lies in its extensibility, flexibility, and modularity. The LDA framework has been extended and studied

extensively in machine learning. These ideas can also be applied to and adapted to the LPD framework and to any other high-dimensional genomic data by simply modifying the distributional assumptions. The three most important and immediate extensions of the LPD framework are: *i.* accounting for any biological annotation of the data such as treatment information or dependence between the genes because of known functional annotations (Blei and McAuliffe, 2007), *ii.* accounting for correlation between the latent group memberships of genes (Blei and Lafferty, 2006), and *iii.* modeling time-course experiments using multiple LPDs, with each LPD modeling NGS data at a particular instant of time during the course of the experiment, which naturally captures the exchangeability of genes and latent group memberships of genes at a particular time-point, but not across multiple time-points (Blei and Lafferty, 2006). LPD is implemented as an R/Bioconductor package (R Development Core Team, 2012) called *themes* (Srivastava, 2011) to facilitate its use by a broad audience.

4. Discussion

NGS technologies have emerged as the preferred approach for exploring both gene function and pathway organization. Although NGS technologies provide highly replicable data, facilitate novel discoveries, and have been used in a variety of applications, statistical design of NGS experiments and statistical methods used to analyze, estimate technical and biological variation, and test scientific hypotheses for NGS data are extremely important to draw reliable conclusions. The discreteness, complexity, and massive size of the data and technology-specific sources of bias, which are absent in other high-throughput data, pose non-trivial computational and

statistical challenges. Issues related to the statistical design of NGS experiments have been addressed, but the statistical methods for NGS data analysis are still under development (Young et al., 2012). Here, we have reviewed the important issues related to analyzing NGS data, to statistical methods for selecting genes suitable for further exploration, to visualization methods to check model assumptions and goodness of fit in the context of differential gene expression analysis, and to penalized likelihood and hierarchical Bayesian modeling approaches that explore the structure of NGS data. The results and modeling approaches are also applicable to any similar data from other sources. We have also outlined areas for further research in the analysis of these data, especially in the computational, statistical, and visualization aspects. The computational efficiency and scalability of the methods developed for NGS data will be extremely important as the sample size increases with decreasing cost of the technology and as projects such as, 1000 Genomes Project (Wang et al. 2008; Stein, 2010) take hold. Because of the discreteness, overdispersion, and zero-inflation of NGS data, it is essential to check the validity of model assumptions and goodness of fit. We also proposed extensions of visualization methods for model checking in the context of NGS data analysis. Further, hierarchical Bayesian methods provide a natural way of modeling the underlying latent structure of NGS data and help in the understanding of complex biological systems in an interpretable manner. It is also easy to extend these models to relax any unjustifiable assumptions. This flexibility comes at the cost of computationally inefficiency to obtain exact inference from NGS data; therefore, we used approximate Bayesian inference techniques in machine learning, such as variational inference and expectation propagation, as novel approaches that address the computational challenges in the analysis and Bayesian modeling of NGS data.

Acknowledgment

We thank Doug Baumann for helpful comments on an earlier version of this manuscript and Figures 1 (a) and (b) are borrowed from his Ph.D. thesis. This work is funded in part by a National Science Foundation (DBI-0733857) grant to RWD and her colleagues. The figures were drawn using the *lattice* package in R (Sarkar, 2008).

References

't Hoen P.A., Ariyurek, Y., Thygesen, H.H., et al. (2008). Deep sequencing-based expression analysis shows major advances in robustness, resolution, and inter-lab portability over five microarray platforms. Nucleic Acids Res. 36:e141.

Agresti, A. (2002). Categorical Data Analysis, Wiley Interscience.

Airoldi, E., Blei, D., Xing, E., and Fienberg, S. (2005). A latent mixed membership model for relational data. In Proceedings of the 3rd international workshop on Link discovery, pp. 82–89. ACM.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data , Genome Biology, Volume 11, 10, BioMed Central Ltd.

Auer, P.L. (2010). Statistical Design And Analysis Of Next-Generation Sequencing Data. Ph. D. Thesis, Department of Statistics, Purdue University.

Auer, P. L. and Doerge, R. W. (2010). Statistical Design and analysis of Next-Generation Sequencing Data. Genetics 185:405-16.

Auer, P. L. and Doerge, R. W. (2011). A Two-Stage Poisson Model for Testing RNA-Seq Data. Statistical Applications in Genetics and Molecular Biology 10(1), 26.

Auer, P.L., Srivastava, S., and Doerge, R.W. (2011). Differential Expression: The next generation and beyond. Briefings in Functional Genomics. doi:10.1093/bfgp/elr041.

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes, Bioinformatics, Volume 17, 6, 509-519.

Baumann, D. D. (2012). Annotation-Informed Integration of 'Omic Data in Next-Generation Sequencing. Ph.D. Proposal Document, Department of Statistics, Purdue University.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society, Series B (Methodological), Volume 57, 1, 289-300.

Benjamini, Y. and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res.May 1;40(10):e72.

Bishop, C. M. (2006). Pattern recognition and machine learning, Volume 4. Springer New York.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. The Journal of Machine Learning Research, Vol. 3, 993-1022.

Blei, D.M. and Lafferty, J. (2006). Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning.

Blei, D.M. and Lafferty, J. (2006). Correlated Topic Models. Neural Information Processing Systems.

Blei, D.M. and McAuliffe, J. (2007). Supervised topic models. Neural Information Processing.

Bradford, J.R., Hey, Y., Yates, T., et al. (2010). A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. BMC Genomics 11:282.

Bullard, J.H., Purdom, E., Hansen, K.D., et al. (2010). Evaluation of statistical methods for normalization and differential expression of mRNA-seq experiments. BMC Bioinformatics 11:94.

Bühlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer.

Churchill, G.A. (2002). Fundamentals of experimental design for cDNA microarrays. Nature Genetics 32 Suppl:490-5.

Casella, G. and Berger, R.L. (2001). Statistical Inference, Duxbury Press.

Cloonan N., Forrest, A.R., Kolle, G., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature Methods 5:613-9.

Cleveland, W.S. (1993). Visualizing Data. Hobart Press.

Cleveland, W.S. (1994). The elements of graphing data. Hobart Press.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association 97(457), 77–87.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment, Journal of the American Statistical Association 96, 1151-1160.

Efron, B. (2010). Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, Cambridge University Press.

Fisher, R.A. (1935). The Design of Experiments, 1935. 3rd. London: Oliver & Boyd Ltd.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1), 1.

Fu, X., Fu, N., Guo, S., et al. (2009). Estimating accuracy of RNA-seq and microarrays with proteomics. BMC Genomics 10:161.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin. D.B. (2003). Bayesian Data Analysis. Chapman and Hall/CRC, Boca Raton, FL, 2nd edition.

Gelman, A. (2004). Exploratory Data Analysis for Complex Models. Journal of Computational and Graphical Statistics, Vol. 13, No. 4, pp. 755-779.

Gelman, A. and Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models, Cambridge University Press, New York, USA.

Gelman, A., Su, Y.-S., Yajima, M., and Hill, J. (2012). Arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.5-05. http://CRAN.R-project.org/package =arm.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S. , Ellis, B. , Gautier, L. , Ge, Y., Gentry, J., and others (2004). Bioconductor: open software development for computational biology and bioinformatics, Genome biology, Volume 5, No. 10.

Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. Biostatistics 8(1), 86.

Hansen, K.D., Wu, Z., Irizarry, R.A., and Leek, J.T. (2012a). Sequencing technology does not eliminate biological variability. Nature Biotechnology 2011, 29:572-573.

Hansen, KD, Irizarry R.A., and Wu, Z. (2012b). Removing technical variability in RNA-Seq data using conditional quantile normalization. Biostatistics 2012, 13(2):204-216.

Hardcastle, T. J. and Kelly, K. A. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data . BMC bioinformatics, Volume 11, 1, 422, BioMed Central Ltd.

Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., and Brown, P. (2000). Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol 1(2), 1–21.

Haury, A-C, Gestraud, P, Vert, J-P (2011). The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. PLoS ONE 6(12): e28210. doi:10.1371/journal.pone.0028210.

Hayden, E. C. (2009). Genome sequencing: the third generation, Nature 457, 769.

Ibrahim, J. G., Chen, M. H., and Gray, R. J. (2002). Bayesian models for gene expression with DNA microarray data, Journal of the American Statistical Association, Volume 97, 457, 88-99.

Langmead B., Hansen, K.D., and Leek, J.T. (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biology 11:R83.

Liu, S., Lin, L., Jiang, P., et al. (2011). A comparison of RNA-seq and high-density exon array for detecting differential gene expression between closely related species. Nucleic Acids Research 39:578-88.

Ma, S. and Huang, J., (2007). Clustering threshold gradient descent regularization: with applications to microarray studies, Bioinformatics , Volume 23, 4, 466, Oxford University Press.

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

McCarthy, DJ, Chen, Y, Smyth, GK (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research 40, 4288-4297.

Meier, L. and Bühlmann, P. (2007). Smoothing L1-penalized estimators for high-dimensional time-course data. Electronic Journal of Statistics 1, 597-615.

Mortazavi, A., Williams, B.A., McCue, K., et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods 5:621-8.

Metzker, M.L. (2009). Emerging Technologies in DNA Sequencing. *Genome Res.* 15(12): 1767-76.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Hu, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., and Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nature Genetics 42, 30-36.

Oshlack, A. and Wakefield M.J. (2009). Transcript length bias in RNA-seq data confounds systems biology. Biol Direct 4:14.

Oshlack, A., M. D. Robinson, and M. D. Young (2010). From RNA-seq reads to differential expression results. Genome Biology 11(12), 220.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Roach, J. C., Glusman, G., Smit, A. F. A., Hu, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L., and Galas, D. J. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328, 636 - 639.

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

Robinson, M. D., and Smyth, G. K., (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9: 321-332.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics}, Volume 26, 1, 139, Oxford Univ Press.

Robinson M.D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 11:R25.

Rogers, S., Girolami, M., Campbell, C., and Breitling, R. (2005). The Latent Process Decomposition of cDNA Microarray Data Sets. IEEE/ACM Transactions on Computational Biology and Bioinformatics Vol.2, No. 2.

Sarkar, D. (2008). Lattice: Multivariate Data Visualization with R. New York: Springer. ISBN 978-0-387-75968-5

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology **3**, No. 1, Article 3.

Smyth, G. K. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pages 397-420.

Srivastava, S. and Doerge, R.W. (2011). Latent Process Decomposition of High-Dimensional Count Data. Technical Report #11-03. Department of Statistics, Purdue University, West Lafayette, IN.

Srivastava, S. (2011). Analysis of High-Dimensional Count Data Using *themes* package. www.stat.purdue.edu/~doerge/lpd.html

Stein, L. D. (2010). The case for cloud computing in genome informatics. Genome Biology 11, 207.

Storey, J. D., (2003). The positive false discovery rate: A Bayesian interpretation and the q-value, The Annals of Statistics, Volume 31, 6, 2013-2035.

Sultan, M., Schulz, M.H., Richard, H., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956-60.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays, Statistical Science, Volume 18, 1, 104-117.

Tibshirani, R., Saunders, M. , Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(1), 91–108.

Thygesen, H. H., and Zwinderman, A. H. (2006). Modeling Sage data with a truncated gamma-Poisson model. BMC Bioinformatics 7: 157.

Vêncio, R. Z., Brentani, H., Patrão, D. F., and Pereira, C. A. (2004). Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). BMC Bioinformatics 5: 119–131.

Wang, E.T., Sandberg, R., Luo, S., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature 456:470-6.

Wang, Z., Gerstein, M., Snyder, M. (2008). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57-63.

Witten, D.M. (2011) Classification and clustering of sequencing data using a Poisson model. Annals of Applied Statistics 5(4): 2493-2518

Wu, Z., Irizarry, R.A., Gentleman, R., Murillo, F.M., and Spencer, F. (2004). A model based background adjustment for oligonucleotide expression arrays. J. Amer. Stat. Assoc. 99, 909-917.

Yang, Y.H. and Speed, T.P. (2002). Design issues for cDNA microarray experiments. Nature Rev. Genetics 3:579-88.

Yang, Y.H., Dudoit, S., Lu, P., Lin, D.M., Peng, V., Ngai, J., Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* vol. 30 (4) pp. e15.

Young, M.D., Wakefield, M.J., Smyth, G.K. et al. (2010). Gene Ontology analysis for RNA-seq: accounting for selection bias. Genome Biology 11:R14.

Young, MD, McCarthy, DJ, Wakefield, MJ, Smyth, GK, Oshlack, A, Robinson, MD (2012). Differential expression for RNA-Seq: mapping, summarization, statistics and experimental design. Chapter 10 of Bioinformatics for High Throughput Sequencing (AM Aransay, ML Hackenberg and N Rodriguez-Ezpeleta, editors), Springer, New York, pages 169--190.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Volume 67, 2, 301 - 320, Wiley Online Library.