voom: precision weights unlock linear model analysis tools for RNA-seq read counts*

Charity W Law Yunshun Chen Wei Shi Gordon K Smyth

Walter and Eliza Hall Institute of Medical Research

1 May 2013
(Last revised 16 February 2014)

Abstract

In the past few years, RNA-seq has emerged as a revolutionary new technology for expression profiling. RNA-seq expression data consists of read counts, and many recent publications have argued therefore that RNA-seq data should be analyzed by statistical methods designed specifically for counts. Yet all the statistical methods developed for RNA-seq counts rely on approximations of various kinds. This article revisits the idea of applying normal-based microarray-like statistical methods to RNA-seq read counts, with the idea that it is more important to model the mean-variance relationship correctly than it is to specify the exact probabilistic distribution of the counts. Log-counts per million are used as expression values. Two methods are developed for modeling the mean-variance relationship in the context of a normal-based analysis. One method, called limma-trend, accommodates the mean-variance relationship as part of the empirical Bayes procedure. The second method, called voom, estimates the mean-variance relationship robustly and generates a precision weight for each individual normalized observation. The normalized log-counts per million and associated precision weights are then entered into the limma analysis pipeline, or indeed into any statistical pipeline for microarray data that is precision weight aware. Either method opens access for RNA-seq analysts to a large body of methodology developed for microarrays, allowing RNA-seq and microarray data to be analyzed in closely comparable ways. The performance of voom and limma-trend is compared to that of edgeR, DESeq, baySeq, TSPM, PoissonSeq, and DSS. Simulation studies show that the limma-based pipelines more than hold their own against the count-based RNA-seq methods in terms of power and error rate control even when the data are generated according to the assumptions of the earlier methods. Several data sets are analyzed to demonstrate how voom and limma-trend can handle heterogeneous data and complex experiments as well as facilitating pathway analysis and gene set testing methods.

*Please cite as: Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* 15, R29. http://genomebiology.com/2014/15/2/R29

Background

Gene expression profiling is one of the most commonly used genomic techniques in biological research. For most of the past 16 years or more, DNA microarrays were the premier technology for genome-wide gene expression experiments, and a large body of mature statistical methods and tools has been developed to analyze intensity data from microarrays. This includes methods for differential expression analysis [1–3], random effects [4,5], gene set enrichment [6], gene set testing [7,8] and so on. One popular differential expression pipeline is that provided by the limma software package [9]. The limma pipeline includes linear modeling to analyze complex experiments with multiple treatment factors, quantitative weights to account for variations in precision between different observations, and empirical Bayes statistical methods to borrow strength between genes. Borrowing information between genes is a crucial feature of the genome-wide statistical methods, as it allows for gene-specific variation while still providing reliable inference with small sample sizes. The normal-based empirical Bayes statistical procedures can adapt to different types of data sets and can provide exact type I error rate control even for experiments with a small number of replicate samples [3].

In the past few years, RNA-seq has emerged as a revolutionary new technology for expression profiling [10]. One common approach to summarize RNA-seq data is to count the number of sequence reads mapping to each gene or genomic feature of interest [11–14]. RNA-seq profiles consist therefore of integer counts, unlike microarrays which yield intensities that are essentially continuous numerical measurements. A number of early RNA-seq publications applied statistical methods developed for microarrays to analyze the RNA-seq read counts. For example, the limma package has been used to analyze the log-counts after normalization by sequencing depth [11,15–17].

Later statistical publications argued that RNA-seq data should be analyzed by statistical methods designed specifically for counts. Much interest has focused on the negative binomial (NB) distribution as a model for the read counts, and especially on the problem of estimating biological variability for experiments with small numbers of replicates. One approach is to fit a global value or global trend to the NB dispersions [13,18,19], although this has the limitation of not allowing for gene-specific variation. A number of empirical Bayes procedures have been proposed to estimate the genewise dispersions [20–22]. Alternatively, Lund et al [23] proposed that the residual deviances from NB generalized linear models be entered into the limma empirical Bayes procedure to enable quasi-likelihood testing. Other methods based on over-dispersed Poisson models have also been proposed [24–26].

Unfortunately, the mathematical theory of count distributions is less tractable than that of the normal distribution, and this tends to limit both the performance and the usefulness of the RNA-seq analysis methods. One problem relates to error rate control with small sample sizes. Despite the use of probabilistic distributions, all the statistical methods developed for RNA-seq counts rely on approximations of various kinds. Many rely on the statistical tests that are only asymptotically valid or are theoretically accurate only when the dispersion is small. All the differential expression methods currently available based on the NB distribution treat the estimated dispersions as if they were known parameters, without allowing for the uncertainty of estimation, and this leads to

statistical tests that are overly liberal in some situations [27, 28]. This is true even of the NB exact test [18], which gives exact type I error rate control when the dispersion is known but which becomes liberal when an imprecise dispersion estimator is inserted for the known value. Quasi-likelihood methods [23] account for uncertainty in the dispersion by using an F-test in place of the usual likelihood ratio test, but this relies on other approximations, in particular that the residual deviances are analogous to residual sums of squares from a normal analysis of variance.

A related issue is the ability to adapt to different types of data with high or low dispersion heterogeneity. None of the empirical Bayes methods based on the NB distribution achieve the same adaptability, robustness or small sample properties as the corresponding methods for microarrays, due to the mathematical intractability of count distributions as compared to the normal distribution.

The most serious limitation though is the reduced range of statistical tools associated with count distributions as compared to the normal distribution. This is more fundamental than the other problems because it limits the types of analyses that can be done. Much of the statistical methodology that has been developed for microarray data relies on use of the normal distribution. For example, we often find it useful in our own microarray gene expression studies to estimate empirical quality weights to downweight poor quality RNA samples [29], or to use random effects to allow for repeated measures on the same experimental units [4,5], or to conduct gene set tests for expression signatures while allowing for inter-gene correlations [7,8]. These techniques broaden the range of experimental designs that can be analyzed or offer improved interpretation for differential expression results in terms of higher level molecular processes. None of these techniques are currently available for RNA-seq analysis using count distributions.

For these reasons, the purpose of this article is to revisit the idea of applying normal-based microarray-like statistical methods to RNA-seq read counts. An obstacle to applying normal-based statistical methods to read counts is that the counts have markedly unequal variabilities, even after log-transformation. Large counts have much larger standard deviations than small counts. While a logarithmic transformation counteracts this, it overdoes the adjustment somewhat so that large log-counts now have smaller standard deviations than small log-counts. We explore the idea that it is more important to model the mean-variance relationship correctly than it is to specify the exact probabilistic distribution of the counts. There is a body of theory in the statistical literature showing that correct modeling of the mean-variance relationship inherent in a data generating process is the key to designing statistically powerful methods of analysis [30]. Such variance modeling may in fact take precedence over identifying the exact probability law that the data values follow [31–33]. We therefore take the view that it is crucial to understand the way in which the variability of RNA-Seq read counts depends on the size of the counts. Our work is in the spirit of pseudo-likelihoods [32] whereby statistical methods based on the normal distribution are applied after estimating a mean-variance function for the data at hand.

Our approach is to estimate the mean-variance relationship robustly and non-parametrically from the data. We work with log-counts normalized for sequence depth, specifically with log-counts-per-million (log-cpm). The mean-variance is fitted to the genewise standard deviations of the log-cpm as a function of average log-count. We explore two ways to

incorporate the mean-variance relationship into the differential expression analysis. The first is to modify the limma empirical Bayes procedure to incorporate a mean-variance trend. The second method incorporates the mean-variance trend into a precision weight for each individual normalized observation. The normalized log-counts and associated precision weights can then be entered into the limma analysis pipeline, or indeed into any statistical pipeline for microarray data that is precision weight aware. We call the first method limma-trend and the second method voom, an acronym for "variance modeling at the observational level". Limma-trend applies the mean-variance relationship at the gene level whereas voom applies it at the level of individual observations.

This article compares the performance of the limma-based pipelines to edgeR [20, 34], DESeq [13], baySeq [21], TSPM [25], PoissonSeq [26] and DSS [22], all of which are based on NB or over-dispersed Poisson distributions. Simulation studies show that the limma pipelines perform at least as well in terms of power and error rate control as the NB or Poisson methods even when the data is generated according to the probabilistic assumptions of the earlier methods. A key advantage of the limma pipelines is that they provide accurate the type I error rate control even when the number of RNA-seq samples is small. The NB and Poisson based methods either fail to control the error correctly or are excessively conservative. Limma-trend and voom perform almost equally well when the sequencing depths are the same for each RNA sample. When the sequencing depths are different, voom is the clear best performer.

Either voom or limma-trend give RNA-seq analysts immediate access to many techniques developed for microarrays that are not otherwise available for RNA-seq, including all the quality weighting, random effects and gene set testing techniques mentioned above. This article presents two case studies which demonstrate how voom can handle heterogeneous data and complex experiments as well as facilitating pathway analysis and gene set testing.

Results

Counts per million: a simple interpretable scale for assessing differential expression

We suppose that RNA-seq profiles (or *libraries*) are available for a set of n RNA samples. Each profile records the number of sequence reads from that sample that have been mapped to each one of G genomic features. A genomic feature can be any pre-defined subset of the transcriptome, for example a transcript or an exon or a gene. For simplicity of language, we will assume throughout this article that reads have been summarized by gene, so that the RNA-seq profiles give the number of reads from each sample that have been mapped to each gene. Typically G is large, in the tens of thousands or more, whereas n can be as low as three. The total number of mapped reads (*library size*) for each sample might vary from a few hundred thousand to hundreds of millions. This is the same context as assumed by a number of previous articles [13, 18, 20, 21, 34].

The number of reads observed for a given gene is proportional not just to the expression level of the gene but also to its gene transcript length and to the sequencing depth

of the library. Dividing each read count by the corresponding library size (in millions) yields counts-per-million (cpm), a simple measure of read abundance that can be compared across libraries of different sizes. Standardizing further by transcript length (in kilobases) gives rise to reads per kilobase per million (rpkm), a well-accepted measure of gene expression [35]. In this article we will work with the simpler cpm rather than rpkm, because we are interested in relative changes in expression between conditions rather than absolute expression.

This article treats log-counts per million (log-cpm) as analogous to log-intensity values from a microarray experiment, with the difference that log-cpm values cannot be treated as having constant variances. Differences in log-cpm between samples can be interpreted as log-fold-changes of expression. The counts are augmented by a small positive value (a half of one read) to avoid taking the logarithm of zero. This ensures no missing log-cpm values and reduces the variability at low count values.

Log-cpms have stabilized variances at high counts

Probability distributions for counts are naturally heteroscedastic, with larger variances for larger counts. It has previously been argued that the mean-variance relationship for RNA-seq counts should be approximately quadratic [34]. This leads to the conclusion that the coefficient of variation (CV) of RNA-seq counts should be a decreasing function of count size for small to moderate counts but should asymptote for larger counts to a value that depends on biological variability. Specifically, the squared CV of the counts should be roughly

$$1/\lambda + \phi$$

where λ is the expected size of the count and ϕ is a measure of biological variation [34]. The first term arises from the technical variability associated with sequencing, and gradually decreases with expected count size, while biological variation remains roughly constant. For large counts, the coefficient of variation is determined mainly by biological variation.

A simple linearization calculation suggests that the standard deviation of the log-cpm should be approximately equal to the CV of the counts (Methods). Examination of a wide range of real data sets confirms these expectations. For studies where the replicates are entirely technical in nature, the standard deviation of the log-cpm decreases steadily as a function of the mean (Figure 1a). For studies where the replicates are genetically identical mice, the standard deviation asymptotes at a moderate level corresponding to a biological coefficient of variation of about 10% (Figure 1b). Studies where the replicates are unrelated human individuals show greater biological variation. For these studies, the standard deviation asymptotes early and at a relatively high level (Figure 1d).

We conclude that log-cpm values generally show a smoothly decreasing mean-variance trend with count size, and that the log-cpm transformation roughly de-trends the variance of the RNA-seq counts as a function of count size for genes with larger counts.

Using log-cpm in a limma pipeline

A simple approach to analyzing RNA-seq data would be input the log-cpm values into a well established microarray analysis pipeline such as that provided by the limma software

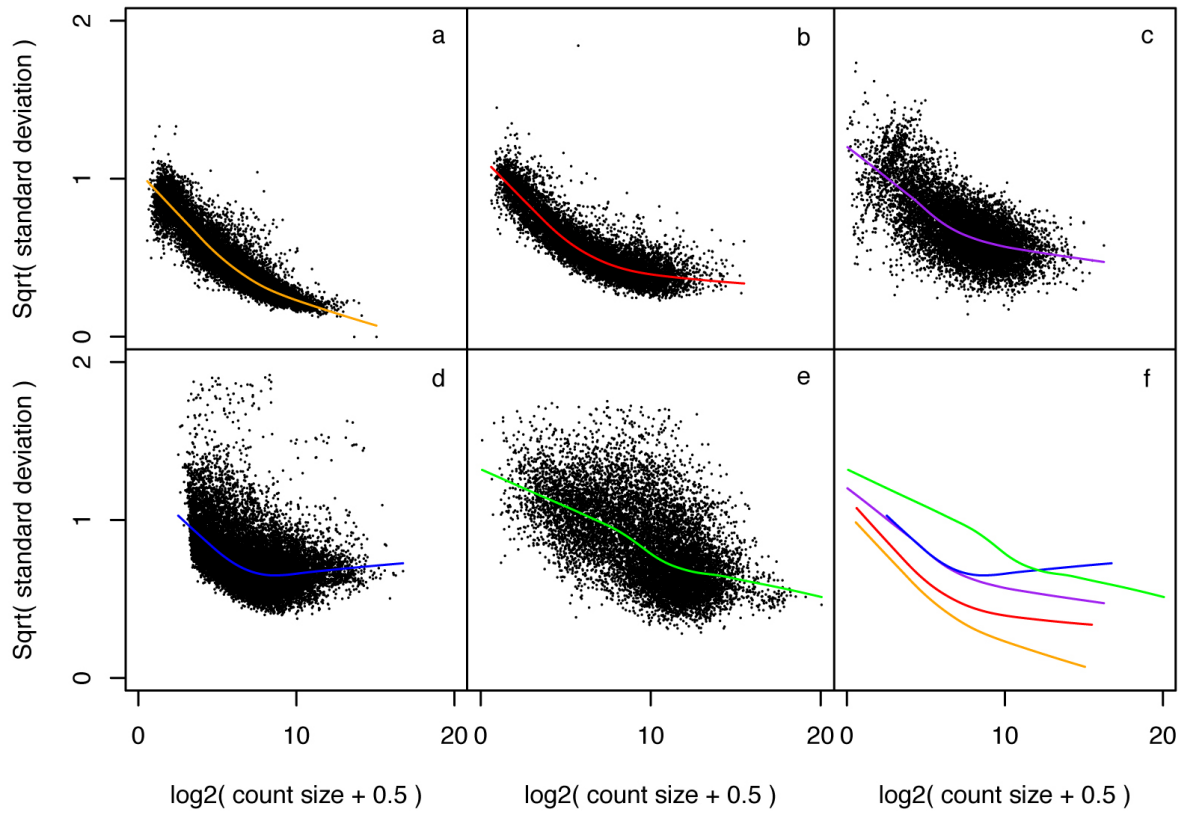

Figure 1: Mean-variance relationships. Gene-wise means and variances of RNA-Seq data represented by black points with a lowess trend. Plots are ordered by increasing levels of biological variation in datasets. Panel (a), voom trend in HBRR and UHRR genes in Sample A, B, C and D of SEQC project; technical variation only. Panel (b), C57BL/6J and DBA mouse experiment; low-level biological variation. Panel (c), simulation study in the presence of 100 up-regulating genes and 100 down-regulating genes; moderate-level biological variation. Panel (d), Nigerian lymphoblastoid cell lines; high-level biological variation. Panel (e), Drosophila melanogaster embryonic developmental stages; very high biological variation due to systematic differences between samples. Panel (f), lowess voom trends for datasets a–e.

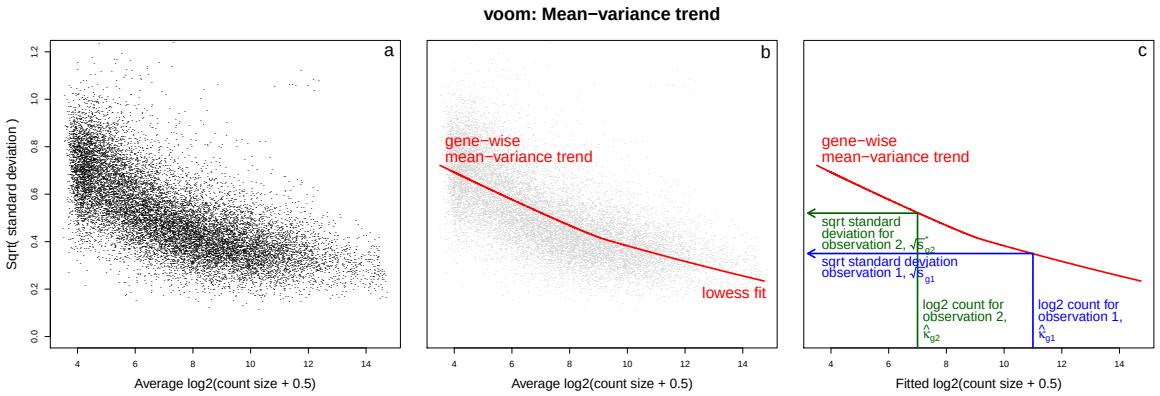

Figure 2: Voom mean-variance modeling. Panel (a), gene-wise square-root residual standard deviations are plotted against average log-count. Panel (b), a functional relationship between gene-wise means and variances is given by a robust lowess fit to the points. Panel (c), the mean-variance trend enables each observation to map to a square-root standard deviation value using its fitted value for log-count.

package [3, 9]. This would be expected to behave well if the counts were all reasonably large, but it ignores the mean-variance trend for lower counts. The microarray pipeline should behave better if modified to include a mean-variance trend as part of the variance modeling. We have therefore modified the empirical Bayes procedure of the limma package so that the genewise variances are squeezed towards a global mean-variance trend curve instead towards a constant pooled variance. This is similar in principle to the procedure proposed by Sartor et al [36] for microarray data, except that we model the trend using a regression spline and our implementation allows for the possibility of missing values or differing residual degrees of freedom between genes. We call this strategy *limma-trend*, whereby the log-cpm values are analyzed as for microarray data but with a trended prior variance. For comparison, the more naive approach without the mean-variance trend will be called *limma-notrend*.

Voom: variance modeling at the observation-level

The limma-trend pipeline models the variance at the gene level. However in RNA-seq applications, the count sizes may vary considerably from sample to sample for the same gene. Different samples may be sequenced to different depths, so different count sizes may be quite different even if the cpm-values are the same. For this reason, we wish to model the mean-variance trend of the log-cpm values at the individual observation level, instead of applying a gene-level variability estimate to all observations from the same gene.

Our strategy is to estimate non-parametrically the mean-variance trend of the logged read counts and to use this mean-variance relationship to predict the variance of each log-cpm value. The predicted variance is then encapsulated as an inverse weight for the log-cpm value. When the weights are incorporated into a linear modeling procedure, the mean-variance relationship in the log-cpm values is effectively eliminated.

A technical difficulty is that we want to predict the variances of individual observations although there is, by definition, no replication at the observational level from which variances could be estimated. We work around this inconvenience by estimating the mean-variance trend at the gene level, then interpolating this trend to predict the

variances of individual observations (Figure 2).

The algorithm proceeds as follows. First, genewise linear models are fitted to the normalized log-cpm values, taking into account the experimental design, treatment conditions, replicates and so on. This generates a residual standard deviation for each gene (Figure 2a). A robust trend is then fitted to the residual standard deviations as a function of the average log-count for each gene (Figure 2b).

Also available from the linear models is a fitted value for each log-cpm observation. Taking the library sizes into account, the fitted log-cpm for each observation is converted into a predicted count. The standard deviation trend is then interpolated to predict the standard deviation of each individual observation based on its predicted count size (Figure 2c). Finally, the inverse squared predicted standard deviation for each observation becomes the weight for that observation.

The log-cpm values and associated weights are then input into the standard limma differential expression pipeline. Most limma functions are designed to accept quantitative weights, providing the ability to perform microarray-like analyses while taking account of the mean-variance relationship of the log-cpm values at the observation level.

Voom and limma-trend control the type I error rate correctly

We have found the voom and limma-trend, especially voom, to perform well and to produce p-values that control error rates correctly over a wide range of simulation scenarios. For illustration we present results from simulations in which read counts were generated under the same NB model as assumed by a number of existing RNA-seq analysis methods. These simulations should represent the ideal for the NB-based methods. If the normal-based methods can give performance comparable or better than count-based methods in these simulations, then this is strong evidence that they will be competitive across a wide range of data types.

Six RNA-seq count libraries were simulated with counts for 10,000 genes. The first three libraries were treated as group 1 and the others as group 2. The distribution of cpm-values for each library was simulated to match the distribution that we observed for a real RNA-seq data set from our own practice. The NB dispersion ϕ was set to decrease on average with expected count size, asymptoting to 0.2-squared for large counts. This degree of biological variation is representative of what we observe for real RNA-seq data, being larger than we typically observe between genetically identical laboratory mice but less than we typically see between unrelated human subjects (Figure 1). An individual dispersion ϕ was generated for each gene around the trend according to an inverse chisquare distribution on 40 degrees of freedom. The voom mean-variance trend for one such simulated dataset is shown in Figure 1c. It can be seen from Figure 1 that the simulated dataset is intermediate between the mouse data (Figure 1b) and the human data (Figure 1d) both in terms of the absolute size of the dispersions and in terms of the heterogeneity of dispersions between genes.

We found that variation in sequencing depth between libraries had a noticeable impact on some RNA-seq analysis methods. Hence all the simulations were repeated under two library size scenarios, one with the same sequencing depth for all six libraries and one with substantial variation in sequencing depth. In the equal size scenario, all libraries were

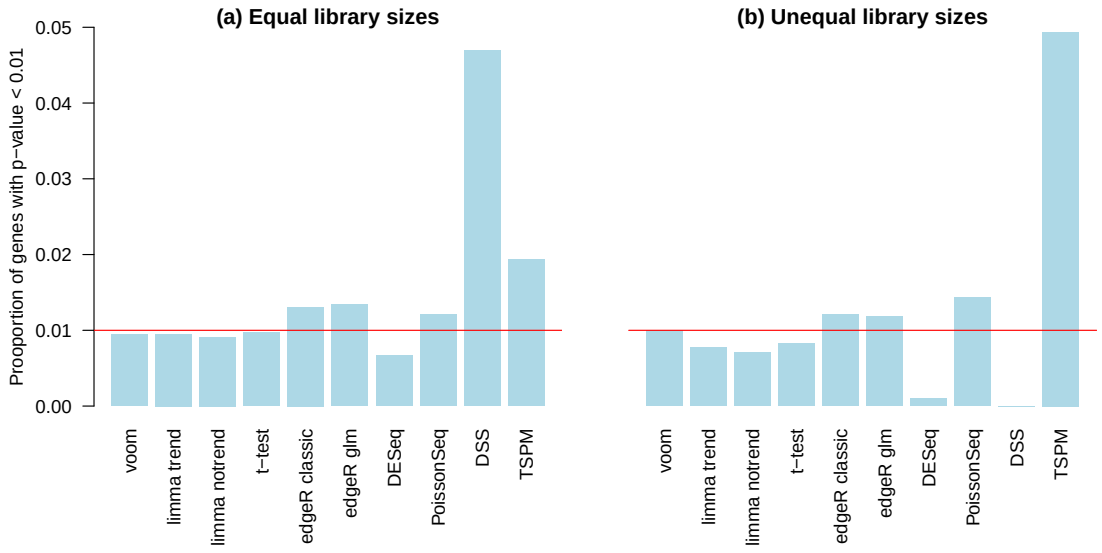

Figure 3: Type I error rates in the absence of true differential expression. The barplots show the proportion of genes with p-value < 0.01 for each method (a) when the library sizes are equal and (b) when the library sizes are unequal. The red line shows the nominal type I error rate of 0.01. Results are averaged over 100 simulations. Methods that control the type I error at or below the nominal level should lie below the red line.

simulated to contain 11 million reads. In the unequal size scenario, the odd-numbered libraries were simulated to have a sequence depth of 20 million reads while the even-numbered libraries had a sequence depth of 2 million reads. Hence the same total number of reads was simulated in this scenario but distributed unevenly between the libraries.

In the first set of simulations, we examined the ability of voom and limma-trend to control the type I error rate correctly in the absence of any genuine differential expression between the groups. When there are no truly differentially expressed genes, the genewise p-values should follow an approximate uniform distribution. If the type I error rate is controlled correctly, then the expected proportion of p-values below any cutoff should be less than or equal to the cutoff value. A number of popular RNA-seq analysis methods based on the negative binomial or Poisson distributions were included for comparison. Figure 3 shows results for a p-value cutoff of 0.01. Results for other cutoffs are qualitatively similar. None of the negative binomial or Poisson-based methods were found to control the type I error rate very accurately. When the library sizes are equal, the negative binomial and Poisson methods were overly liberal, except for DESeq which is very conservative. When the library sizes are unequal, DSS and DESeq became extremely conservative. By contrast, all the normal-based methods were slightly conservative. Voom yields very close to the nominal type I error in both library size scenarios. Limma-trend is similar to voom when the library sizes are equal but somewhat conservative when the library sizes are unequal.

BaySeq was not included in the type I error rate comparison because it doesn't return p-values. However results in the next section will show it to be relatively conservative in terms of false discovery rate (FDR) (Figure 4).

To check voom's conservativeness on real data, we used a set of four replicate libraries

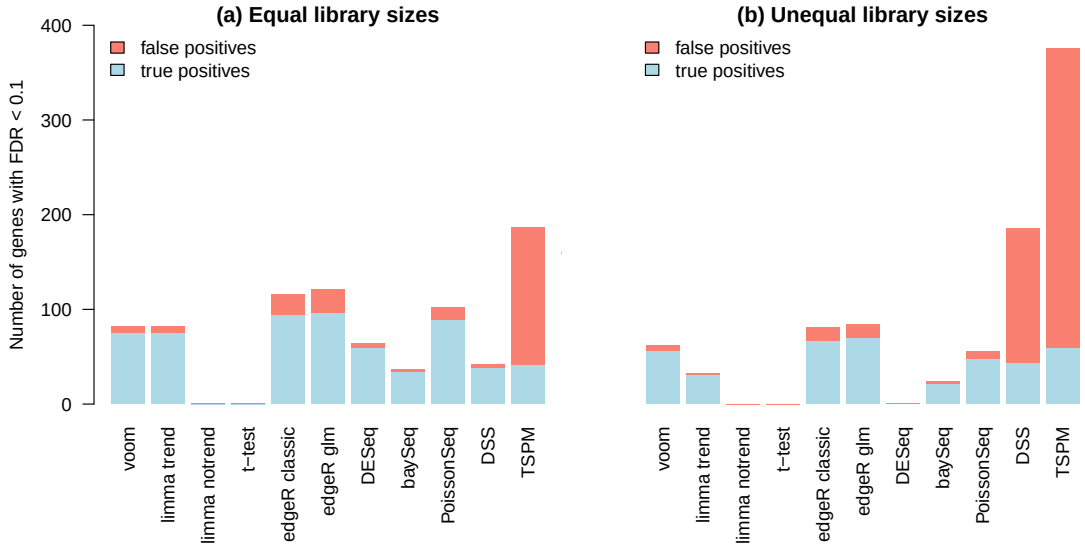

Figure 4: Power to detect true differential expression. Bars show the total number of genes that are detected as statistically significant (FDR < 0.1) (a) with equal library sizes and (b) with unequal library sizes. The blue segments show the number of true positives while the red segments show false positives. The number of genuinely DE genes in this simulation is 200. Results are averaged over 100 simulations. Height of the blue bars shows empirical power. The ratio of the red to blue segments shows empirical FDR.

from the SEQC Project [37]. All four libraries were Illumina HiSeq 2000 RNA-seq profiles of samples of Ambion's Human Brain Reference RNA (HBRR) [38]. We split the four libraries into two groups in all possible ways, and tested for differential expression between the two groups for each partition. Voom returned no DE genes at 5% FDR for six out the seven possible partitions, indicating good error rate control. The voom mean-variance trend for the SEQC data, using all the libraries rather than the HBRR samples only, is shown in Figure 1a.

Voom has the best power of methods that control the type I error rate

Next we examine power to detect to true differential expression. For the following simulations, 100 randomly selected genes were 2-fold up-regulated in the first group and another 100 were 2-fold up-regulated in the second group. This represents a typical scenario for a functional genomics experiment in which the differential expression effects are large enough to be biologically important but nevertheless sufficiently subtle as to challenge many analysis methods. Figure 4 shows the number of true and false discoveries made by various analysis methods at significance cutoff FDR < 0.1. When the library sizes are equal, voom and limma-trend have next best power after edgeR and PoissonSeq. However both edgeR and PoissonSeq give empirical FDRs greater than 0.1, confirming the results of the previous section that these methods are somewhat liberal. Limma-trend gives empirical FDR slightly greater than voom but still less than 0.1. With unequal library sizes, voom has the best power except for edgeR while still maintaining a low FDR. TSPM

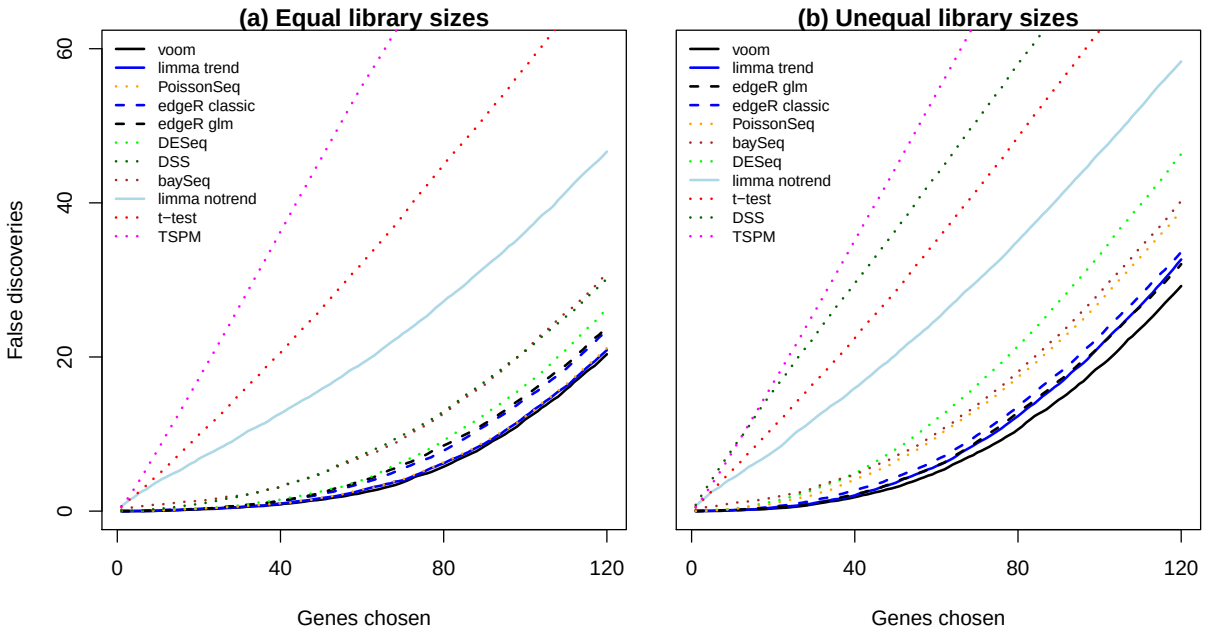

Figure 5: False discovery rates. The number of false discoveries is plotted for each method versus the number of genes selected as DE. Results are averaged over 100 simulations (a) with equal library sizes and (b) with unequal library sizes. Voom has the lowest FDR at any cutoff in either scenario.

declares by far the most DE genes, but these are mostly false discoveries. DSS also gives a worryingly high rate of false discoveries when the library sizes are unequal. Figures 3 and 4 together show that voom has the best power of those methods that correctly control the type I and FDR error rates.

Voom has the lowest false discovery rate

Next we compared methods from a gene ranking point of view, comparing methods in terms of the number of false discoveries for any given number of genes selected as DE. Methods that perform well will rank the truly DE genes in the simulation ahead of non-DE genes. Genes were ranked by posterior likelihoods for baySeq and by p-value for the other methods. The results show that voom has the lowest FDR at any cutoff (Figure 5). When the library sizes are equal, limma-trend and PoissonSeq are very close competitors (Figure 5a). When the library sizes are unequal, limma-trend and edgeR are the closest competitors (Figure 5b).

Next we compared FDRs using spike-in control transcripts from the SEQC project [39]. The data consists of eight RNA-seq libraries, in two groups of four. A total of 92 artificial control transcripts were spiked-in at different concentrations in such as way that three quarters of the transcripts were truly DE and the remaining quarter were not. To make the spike-ins more like a realistic data set, we replicated the counts for each of the 23 non-DE transcripts three times. That is, we treated each non-DE transcript as three different transcripts. This resulted in a dataset of 138 transcripts with half DE and half non-DE. Figure 6 is analogous to Figure 5 but using the spike-in data instead of simulated data. Voom again achieved the lowest FDR, with edgeR and the other limma methods again

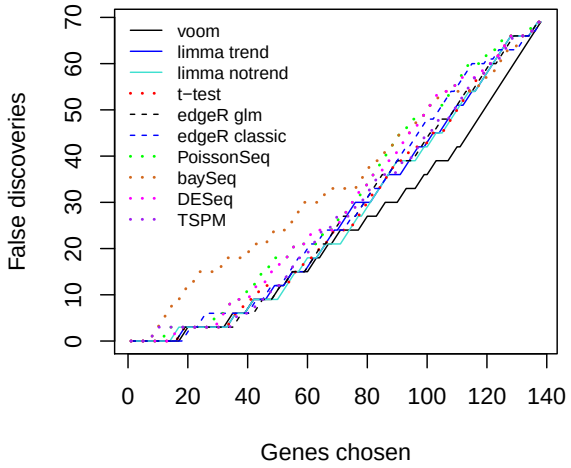

Figure 6: False discovery rates evaluated from SEQC spike-in data. The number of false discoveries is plotted for each method versus the number of genes selected as DE. voom has the lowest FDR overall.

being the closest competitors (Figure 6).

Voom and limma-trend are faster than specialist RNA-seq methods

The different statistical methods compared varied considerably in computational time required, with DESeq, TSPM and baySeq being slow enough to limit the number of simulations that were done. Voom is easily the fastest of the methods compared, with edgeR-classic next fastest (Figure 7).

RNA-seq profiles of male and female Nigerian individuals

So far we have demonstrated the performance of voom on RNA-seq data sets with small numbers of replicate libraries. To demonstrate the performance of voom on a heterogeneous data set with a relatively large number of replicates and a high level of biological variability, we compared males to females using RNA-seq profiles of lymphoblastoid cell lines from 29 male and 40 female unrelated Nigerian individuals [40]. Summarized read counts and gene annotation are provided by the Bioconductor tweeDEseqCountData package [41]. Figure 1d shows the voom mean-variance trend of this dataset.

Voom yielded 16 genes up-regulated in males and 43 up-regulated in females at 5% FDR. As expected, most of the top differentially expressed genes belonged to the X or Y sex chromosomes (Table 1). The top gene is XIST, which is a key player in X-inactivation and is known to be expressed at meaningful levels only in females.

We examined 12 particular genes that are known to belong to the male-specific region of chromosome Y [42, 43]. A ROAST gene set test confirmed that these genes collectively

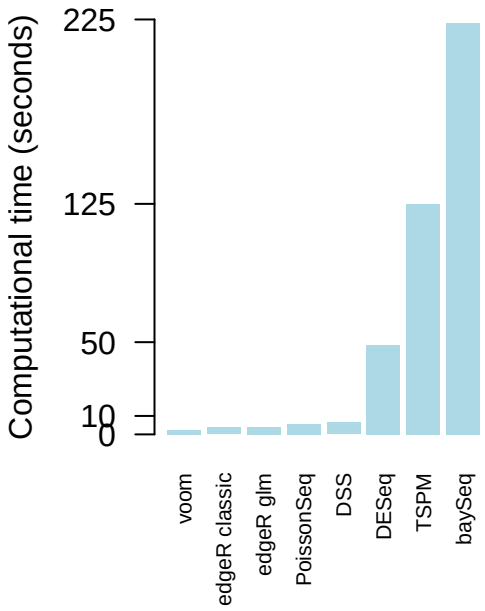

Figure 7: Computing times for RNA-Seq methods. Bars show time in seconds required for the analysis of one simulated dataset on a MacBook laptop. Methods are ordered from quickest to most expensive.

Table 1: Top 16 genes differentially expressed between males and females in the Nigerian data

```
> topTable(fit,coef=2,n=16,sort="p")
                     Symbol Chr  logFC AveExpr     t  P.Value       FDR    B
ENSG00000229807        XIST   X -9.857  3.8076 -32.5 1.21e-44  2.10e-40 66.5
ENSG00000099749      CYorf15A   Y  4.260  0.3139  27.1 2.40e-39  2.08e-35 64.8
ENSG00000157828       RPS4Y2   Y  3.280  3.3074  26.7 5.88e-39  3.39e-35 73.8
ENSG00000233864       TTTY15   Y  4.896 -0.5546  25.4 1.62e-37  7.00e-34 59.6
ENSG00000198692       EIF1AY   Y  2.397  2.6799  20.5 1.04e-31  3.60e-28 59.1
ENSG00000131002      CYorf15B   Y  5.439 -0.1717  20.2 2.46e-31  7.09e-28 51.1
ENSG00000213318  RP11-331F4.1  16  4.295  2.2647  19.4 3.04e-30  7.51e-27 55.0
ENSG00000165246       NLGN4Y   Y  5.330 -0.4924  17.9 3.59e-28  7.78e-25 45.9
ENSG00000129824       RPS4Y1   Y  2.779  4.7110  17.8 6.17e-28  1.19e-24 52.2
ENSG00000183878          UTY   Y  1.877  2.7422  16.6 3.35e-26  5.81e-23 47.9
ENSG00000012817        KDM5D   Y  1.470  4.7039  14.7 2.33e-23  3.67e-20 42.3
ENSG00000243209    AC010889.1   Y  2.536 -0.0186  14.0 2.97e-22  4.28e-19 36.5
ENSG00000146938       NLGN4X   X  4.472 -0.7808  13.8 6.47e-22  8.62e-19 34.8
ENSG00000067048        DDX3Y   Y  1.670  5.3070  13.3 4.54e-21  5.61e-18 37.3
ENSG00000006757       PNPLA4   X -0.989  2.5333 -10.4 5.09e-16  5.87e-13 25.8
ENSG00000232928 RP13-204A15.4   X  1.434  3.2499  10.2 1.43e-15  1.54e-12 25.0
```

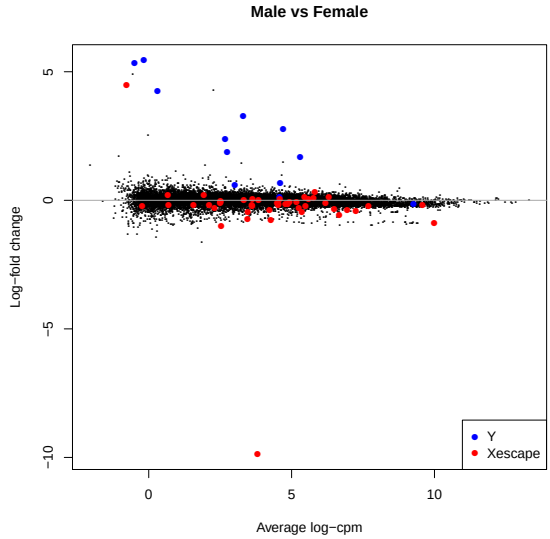

Figure 8: MA-plot with male and female specific genes highlighted. The log-fold change of each gene is plotted against its average log-cpm for the comparison between males and females. Genes on the male-specific region of the Y chromosome genes are highlighted blue and are consistently up-regulated in males, while genes on the X chromosome reported to escape X-inactivation are highlighted red and are generally down in males.

are significantly up-regulated in males ($P = 0.0001$). A CAMERA gene set test was even more convincing, confirming that these genes are significantly more up-regulated in males than are other genes in the genome ($P = 2 \times 10^{-28}$).

We also examined 46 X-chromosome genes that have been reported to escape X-inactivation [43, 44]. These genes were significantly up-regulated in females (ROAST $P = 0.0001$, CAMERA $P = 10^{-10}$). The log-fold-changes for the X and Y chromosome genes involved in the gene set tests are highlighted on an MA-plot (Figure 8).

Note that these gene set testing approaches are not available in conjunction with any of the count-based approach to differential expression. If a count-based method had been used to assess differential expression, we could still have examined whether sex-linked genes were highly ranked among the differentially expressed genes, but we could not have undertaken any formal statistical test for enrichment of this signature while accounting for inter-gene correlation. On the other hand, the voom expression values and weights are suitable for input into the ROAST and CAMERA procedures without any further processing.

Development stages of *D. melanogaster*

Like edgeR-glm, but unlike most other analysis tools, voom and limma-trend offer full-featured linear modeling for RNA-seq data, meaning that they can analyze arbitrary complex experiments. The possibilities of linear modeling are so rich that it is impossible to select a representative example. Voom and limma could be used to analyze any gene-level RNA-seq differential expression experiment, including those with multiple ex-

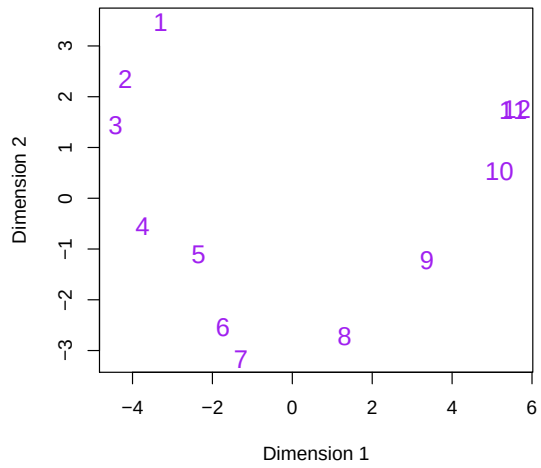

Figure 9: Multidimensional scaling plot of *D. melanogaster* embryonic stages. Distances are computed from the log-cpm values. The 12 successive embryonic developmental stages are labeled 1 to 12, from earliest to latest.

perimental factors [34]. Here we give a novel analysis illustrating the use of quadratic regression to analyze a time-course study.

RNA-Seq was used to explore the developmental transcriptome of Drosophila melanogaster [45]. RNA-Seq libraries were formed from whole-animal samples to represent a large number of distinct developmental stages. In particular, samples were collected from embryonic animals at equi-spaced development stages from 2 hours to 24 hours in 2-hour intervals. Here we analyze the 12 RNA-seq libraries from these embryonic stages. We seek to identify those genes that are characteristic of each embryonic stage. In particular we wish to identify, for each embryonic stage, those genes which achieve their peak expression level during that stage.

As all the samples are from distinct stages, there are no replicate libraries in this study. However we utilize the fact that gene expression should for most genes vary smoothly over time. A multidimensional scaling plot on log-cpm values shows the gradual change in gene expression during embryonic development, with each stage intermediate in expression profile between the stages before and after (Figure 9). We use genewise linear models to fit a quadratic trend with time to the log-cpm values for each gene. These quadratic trends will not match all the intricacies of gene expression changes over time but are sufficient to model the major trends. The voom mean-variance trend for this data is shown in Figure 1e.

Out of 14869 genes that were expressed during embryonic development, 8366 showed a statistically significant trend at 5% FDR using empirical Bayes F-tests. For each differentially expressed gene, we identified the embryonic stage at which the fitted quadratic trend achieved its maximum value. This allowed us to associate each significant gene with a particular development stage (Figure 10). Most genes peaked at the first or last stage (Figure 10), indicating smoothly decreasing or increasing trends over time (Figure 11, panels 1 and 12). Genes peaking at the first embryonic stage tended to be associated with

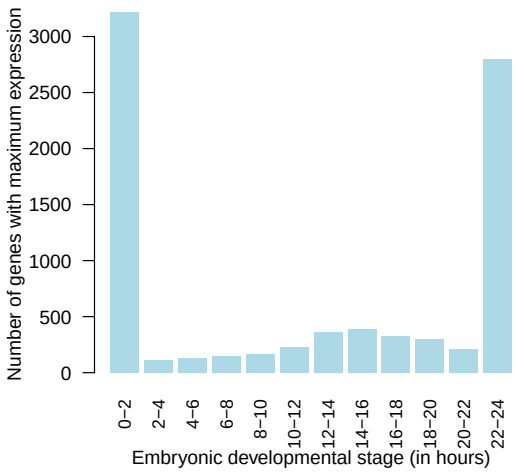

Figure 10: Number of genes associated with each *D. melanogaster* embryonic stage. The number of genes whose peak estimated expression occurs at each of the stages is recorded.

the cell cycle. Genes peaking at the last stage tended to be associated with precursor metabolites and energy, the oxidation-reduction process and metabolic pathways.

Genes peaking at intermediate stages showing expression trends with an inverse-U shape (Figure 11, panels 2–11). There was a substantial set of genes with peak activity between 12–16 hours of embryonic development (Figure 10), suggesting some important developmental change occurring during this period requiring the action of special-purpose genes. Indeed, gene ontology analysis of the genes associated with this period showed that anatomical structure morphogenesis was the most significantly enriched biological process. Other leading terms were organ morphogenesis and neuron differentiation.

This analysis demonstrates a simple but effective means of identifying genes that have a particular role at each developmental stage.

Discussion

This article follows the common practice of examining differential expression on a genewise basis. Our preferred practice is to count the total number of reads overlapping annotated exons for each gene. While this approach does not allow for of the possibility that different isoforms of the same gene may be differentially expressed in different directions, it does provide a statistically robust gene-level analysis even when the sequencing depths are quite modest. The relevance of gene-level analyses is also supported by recent surveys of transcription have shown that each gene tends to have a dominant isoform that accounts for far more of the total expression for that gene that do any of the remaining isoforms [46, 47]. The voom analysis can also be conducted at the exon level instead of at the gene level as an aid to detecting alternative splicing between the treatment groups.

In this article, voom has been applied to log-cpm values. Voom can work however just as easily with logged rpkm values in place of log-cpm, because the precision weights are

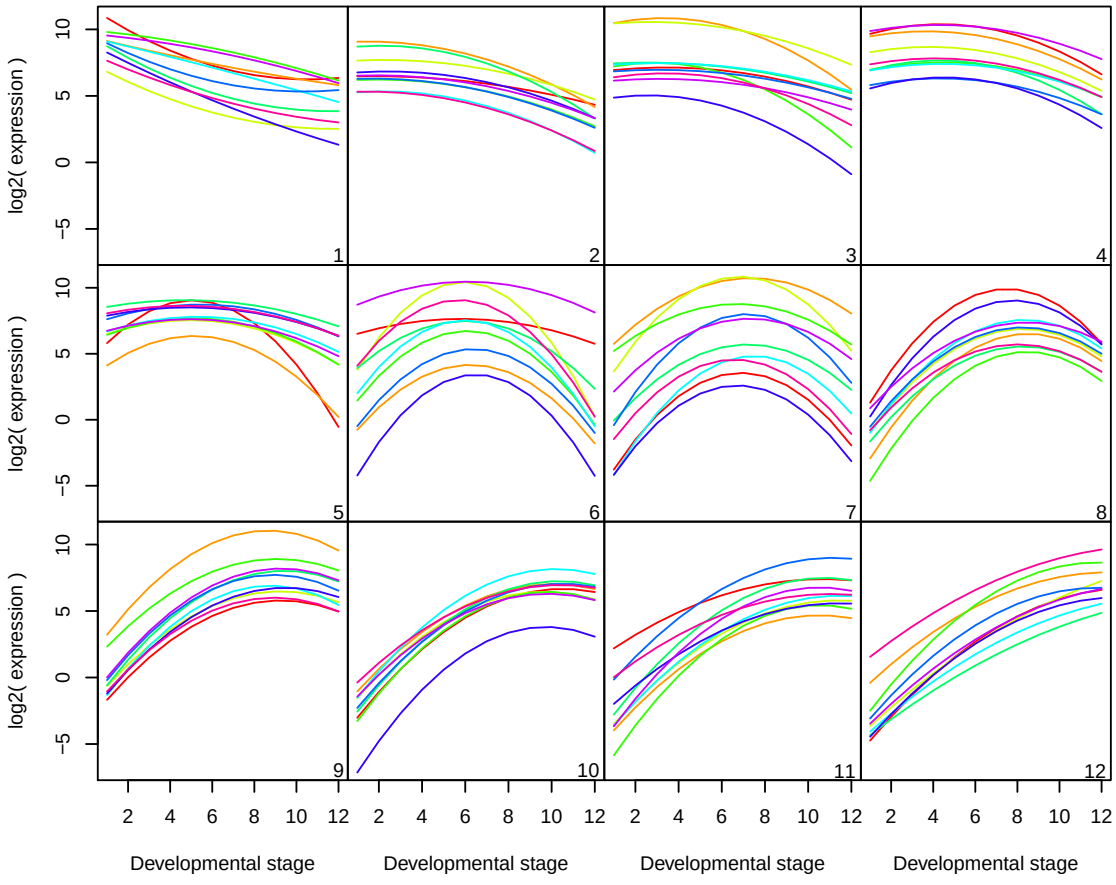

Figure 11: Expression trends for genes that peak at each *D. melanogaster* embryonic stage. The panels (1) to 12) correspond to the twelve successive developmental stages. Each panel displays the fitted expression trends for the top ten genes that achieve their peak expression during that stage. In particular, panel (1) shows genes that are most highly expressed at the first stage and panel (12) shows genes most highly expressed at the last stage. Panels (7) and (8) are notable because they show genes with marked peaks at 12–14 hours and 14–16 hours respectively.

the same for both measures. If the genomic length of each gene is known, then the log-cpm values output by voom can be converted to log-rpkm by subtracting the log-base-2 gene length in kilobases. The downstream analysis is unchanged and will yield identical results in terms of differentially expressed genes and estimated fold changes.

This article has shown that a normal-based analysis of RNA-seq read count data performs surprisingly well relative to methods that use special-purpose count distributions. The motivation for examining normal-based methods was to open up access to a range of microarray-like analysis tools based on the normal distribution. From this point of view, the normal-based methods only need to perform comparably to the count-based methods in terms of power and FDR control in order to be a success. Our comparisons suggest not only that this is so, but that the normal-based methods actually have a performance advantage. We found voom to be the best performer across our simulations and comparisons, and even the simpler limma-trend method performed equal or better than the count-based methods. Voom and limma-trend perform almost equally when the library sizes are equal, but voom has the advantage when the library sizes are unequal. The best performing count-based methods were edgeR and PoissonSeq, although neither of those methods controlled the type I error rate at the nominal level, both being somewhat liberal.

The performance advantage of voom over many of the count-based methods was quite substantial in our simulations, despite the simulations being conducted under the same NB distributional assumptions as made by a number of existing methods. Other simulation scenarios would tend to increase voom's advantage. For example, it would be at least as scientifically reasonable to assume that the true expression levels for each gene follow a log-normal distribution between replicates instead of a gamma distribution, and such an assumption would tend to improve the performance of voom relative to edgeR, DESeq, baySeq and DSS. In general, voom makes fewer distributional assumptions than do competing methods and can therefore be expected to perform robustly across a range of scenarios.

This study presented simulations with equal library sizes between replicates, and also explored the sensitivity of the methods to unequal library sizes. In our experience markedly unequal library sizes can arise in real RNA-seq experiments for a variety of reasons. One scenario is when an experiment is conducted in stages and samples sequenced at a later time have a much higher sequencing depth. Other possible scenarios occur when technical replicates are combined for a subset of samples or when DNA samples are multiplexed onto a sequencing lane in unequal quantities. Some of the negative binomial based analysis methods become very conservative or showed very poor FDR control when the library sizes were unequal. In contrast, voom shows consistent performance in all scenarios.

The worst performer in our simulation was TSPM, presumably because we have simulated from NB distributions, which have quadratic mean-variance relationships, whereas TSPM assumes a linear mean-variance relationship [25]. The second worst performer was the ordinary t-test. This shows that traditional statistical methods cannot be reliably applied to genomic data without borrowing strength between genes. The third worst performer was limma-notrend, showing that the mean-variance trend in the log-cpm values cannot be ignored.

To examine sensitivity of the results to the shape of the dispersion distribution, we repeated all the simulations using a log-normal distribution for the genewise dispersions instead of a inverse-chisquare distribution. The two distributions were chosen to have the same mean and variance on the log-scale. The results were virtually unchanged from those shown in Figures 3–5, showing that the shape of the dispersion distribution is not a major determination of performance. This agrees with a similar conclusion in Wu et al [22].

It requires some explanation why voom, a method that ignores the discrete integer nature of the counts, should perform so well. We think that several issues are important. First, the parametric advantages of the Poisson or NB distributions are mitigated by the fact that the observed mean-variance relationship of RNA-seq data does not perfectly match the theoretical mean-variance relationships inherent in these distributions. While the quadratic mean-variance relationship of the NB distribution captures most of the mean-variance trend, the NB dispersion still shows a non-ignorable trend with gene abundance [13, 19, 34]. This means that the mean-variance relationship still has to be estimated non-parametrically, at least in part.

Second, voom is more precise than previous methods in terms of its treatment of the mean-variance trend. While several previous methods fit a semi-parametric trend to the variances or to the NB dispersions [13, 19, 23, 34], the trend has always been used to estimate gene-level model parameters. This ignores the fact that different counts for the same gene may vary substantially in size, meaning that the trend should be applied differently to different observations. This consideration becomes more critical when different RNA samples are sequenced to different depths.

Third, the use of normal models gives voom access to tractable empirical Bayes distribution theory [3], facilitating reliable estimation of the Bayesian hyperparameters and exact small sample distributions for the test statistics. Amongst other things this facilitates accurate estimate of the prior degrees of freedom determining the optimal amount of squeezing to be applied to the variances.

Fourth, the use of normal distribution approximations in conjunction with variance modeling is partly supported by generalized linear model theory. Rao's score test [48] for a covariate in a generalized linear model is essentially equivalent to the normal theory test statistic, provided that the mean-variance function is correctly estimated and incorporated into appropriate precision weights [49]. Score tests have similar performance to likelihood ratio tests when the null hypothesis is true or when the changes being detected are relatively small.

Some of the count-based methods have been criticized as being sensitive to outlier counts [28]. Voom and limma-trend methods inherit good robustness properties from the normal-based procedures in limma [28]. If necessary, they can be made extremely robust to outliers and hypervariable genes using the robust empirical Bayes options of the limma package [50]

In addition to performance results, voom offers a number of qualitative inducements over the count-based methods. It is fast and convenient. It allows RNA-seq and microarray data to be analyzed in closely comparable ways, which may be an attraction for analysts comparing results from the two technologies. It gives access to a wealth of statistical methods developed for microarrays, including for example the gene set testing

methods demonstrated on the Nigerian dataset.

Conclusions

Voom performs as well or better than existing RNA-seq methods, especially when the library sizes are unequal. It is moreover faster and more convenient, and converts RNA-seq data into a form whereby it can be analyzed using similar tools as for microarrays.

Materials and methods

Log-counts per million

We assume that an experiment has been conducted to generate a set of n RNA samples. Each RNA sample has been sequenced, and the sequence reads have been summarized by recording the number mapping to each gene. The RNA-seq data consist therefore of a matrix of read counts r_{gi}, for RNA samples $i = 1$ to n, and genes $g = 1$ to G. Write R_i for the total number of mapped reads for sample i, $R_i = \sum_{g=1}^{G} r_{gi}$. We define the log-counts per million (log-cpm) value for each count as

$$y_{gi} = \log_2 \left(\frac{r_{gi} + 0.5}{R_i + 1.0} \times 10^6 \right)$$

The counts are offset away from zero by 0.5 to avoid taking the log of zero, and to reduce the variability of log-cpm for low expression genes. The library size is offset by 1 to ensure that $(r_{gi} + 0.5)/(R_i + 1)$ is strictly less than 1 as well as strictly greater than zero.

Delta rule for log-cpm

Write $\lambda = E(r)$ for the expected value of a read count given the experimental conditions, and suppose that $\text{var}(r) = \lambda + \phi\lambda^2$, where ϕ is a dispersion parameter. If r is large, then the log-cpm value of the observation is $y \approx \log_2(r) - \log_2(R) + 6\log_2(10)$, where R is the library size. Note that the analysis is conditional on R, so R is treated as a constant. It follows that $\text{var}(y) \approx \text{var}(\log_2(r)$. If λ also is large, then $\log_2(r) \approx \lambda + (r - \lambda)/\lambda$ by Taylor's theorem [51], so $\text{var}(y) \approx \text{var}(r)/\lambda^2 = 1/\lambda + \phi$.

Linear models

This article develops differential expression methods for RNA-seq experiments of arbitrary complexity, for example experiments with multiple treatment factors, batch effects or numerical covariates. As has been done previously [3, 7, 8, 34], we use linear models to describe how the treatment factors are assigned to the different RNA samples. We assume that

$$E(y_{gi}) = \mu_{gi} = x_i^T \beta_g$$

where x_i is a vector of covariates and β_g is a vector of unknown coefficients representing $\log_2$-fold-changes between experimental conditions. In matrix terms,

$$E(y_g) = X\beta_g$$

where y_g is the vector of log-cpm values for gene g and X is the design matrix with the x_i as rows. Interest centers on testing whether one or more of the β_{gj} are equal to zero,

Voom variance modeling

The above linear model is fitted, by ordinary least squares, to the log-cpm values y_{gi} for each gene. This yields regression coefficient estimates $\hat{\beta}_g$, fitted values $\hat{\mu}_{gi} = x_i^T\hat{\beta}_g$ and residual standard deviations s_g.

Also computed is the average log-cpm $\bar{y}_g$ for each gene. The average log-cpm is converted to an average log-count value by

$$\tilde{r} = \bar{y}_g + \log_2(\tilde{R}) - \log_2(10^6)$$

where $\tilde{R}$ is the geometric mean of the library sizes plus one.

To obtain a smooth mean-variance trend, a lowess curve is fitted to square-root standard deviations $s_g^{1/2}$ as a function of mean log-counts $\tilde{r}$ (Figure 2ab). Square-root standard deviations are used because they are roughly symmetrically distributed. The lowess curve [52] is statistically robust [53] and provides a trend line through the majority of the standard deviations. The lowess curve is used to define a piecewise linear function lo() by interpolating the curve between ordered values of $\tilde{r}$.

Next the fitted log-cpm values $\hat{\mu}_{gi}$ are converted to fitted counts by

$$\hat{\lambda}_{gi} = \hat{\mu}_{gi} + \log_2(R_i + 1) - \log_2(10^6).$$

The function value $\text{lo}(\hat{\lambda}_{gi})$ is then the predicted square-root standard deviation of y_{gi}.

Finally, the voom precision weights are the inverse variances $w_{gi} = \text{lo}(\hat{\lambda}_{gi})^{-4}$ (Figure 2c). The log-cpm values y_{gi} and associated weights w_{gj} are then input into the standard limma linear modeling and empirical Bayes differential expression analysis pipeline.

Gene set testing methods

ROAST [7] and CAMERA [8] are gene set testing procedures which assess changes in the overall expression signature defined by a set of genes. ROAST [7] is a self-contained test that assesses differential expression of the gene set without regard to genes not in the set. CAMERA [8] is a competitive test that assesses differential expression of the gene set relative to all other genes on the array. Both procedures offer considerable flexibility as they have the ability to test the association of a genomic pathway or gene set signature with quite general treatment comparisons or contrasts defined in the context of a microarray linear model. We have adapted both methods to make use of quantitative weights as output by voom. The revised methods are implemented in the functions roast() and camera() of the limma software package.

Normalization

The log-cpm values are by definition normalized for sequencing depth. Other normalization steps can optionally be done. The library sizes R_i can be scale normalized to adjust for compositional differences between the RNA-seq libraries [54]. This produces normalized library sizes R_i^* that can be used in place of R_i in the voom pipeline. Alternatively, between-array normalization methods developed for single channel microarray data, such as quantile or cyclic loess, can be are applied to the log-cpm values.

Simulations

The simulations were designed to generate data with characteristics similar to real data that we analyze in our own practice. First a set of baseline expression values was generated representing the relative proportion of counts expected to arise from each gene. These proportions were translated into expected count sizes by multiplying by library size, and then multiplied by true fold changes as appropriate. Counts were then generated following a NB distribution with the specified mean and dispersion for each observation.

The distribution of baseline values was chosen to match that from RNA-seq experiments conducted at our institution. Specifically we used the goodTuringProportions function of the edgeR package [12], which implements the Good-Turing algorithm [55], to predict the true proportion of total RNA attributable to each gene. We ran this function on a number of different libraries, pooled the predicted proportions and formed a smoothed distribution function. The baseline proportions for the simulations were then generated to follow this distribution.

The NB dispersions were generated as follows. The trend in the dispersions was set to be ψ_{gi} with $\psi_{gi}^{1/2} = 0.2 + \lambda_{gi}^{-1/2}$ where λ_{gi} is the expected count size. A modest amount of genewise biological variation was generated from an inverse chisquare distribution with 40 degrees of freedom. The individual dispersions were set to be $\phi_{gi} = \psi_{gi}\delta_g$ where $40/\delta_g \sim \chi^2_{40}$.

In an alternative simulation, to investigate sensitivity to the distribution of genewise dispersions, the δ_g were simulated as log-normal with mean 0 and standard deviation 0.25 on the log-scale. This produces a distribution with a similar coefficient of variation as for the inverse chisquare simulation.

For each simulated data set, genes with less than 10 reads across all samples were filtered from the analysis. PoissonSeq resets the seed of the random number generator in R, so it was necessary to save and restore the state of the random number generator before and after each call of the main PoissonSeq function.

Complete runnable code that reproduces all the simulation results shown in the article is provided on the voom website [56].

SEQC data

The SEQC project, also known as MAQC-III, aims to provide a comprehensive study of next-generation sequencing technologies [37]. We analyze here a pilot SEQC dataset consisting of 16 RNA-seq libraries in four groups. The full SEQC data including the 16

libraries analyzed here will become available as GEO series GSE47792 when the main SEQC article is published in 2014. In the meantime, the aligned and summarized read counts for the pilot libraries needed to repeat the analyses in this article are available from the voom webpage [56].

The groups are labeled A–D and are closely analogous to the similarly labeled RNA samples used in the earlier microarray quality control study [57]. Libraries in group A are profiles of Stratagene's Universal Human Reference RNA (UHRR) with the addition of RNA from Ambion's ERCC ExFold RNA spike-in mix 1 (Mix 1). Libraries in group B are profiles of Ambion's Human Brain Reference RNA (HBRR) with added RNA from Ambion's ERCC ExFold RNA spike-in mix 2 (Mix 2). RNA samples in group C and D are mixtures of A and B in proportions 75–25 and 25–75 respectively. An Illumina HiSeq 2000 was used to create a FastQ file of paired-end sequence reads for each sample. The library size for each sample varied from 5.4 to 8.0 million read pairs. Fragments were mapped to NCBI Build 37.2 of the human genome using the Subread aligner [58]. Fragment counts were summarized by Entrez Gene ID using the featureCounts function [59] of version 1.8.2 of the Bioconductor package Rsubread [60]. Fragments with both end reads mapped successfully contributed one count if the fragment overlapped any annotated exon for that gene. Fragments for which only one read mapped successfully contributed half a count if that read overlapped an exon. The summarized read count data is available from the voom webpage [56].

The voom mean-variance trend shown in Figure 1a was obtained from all 16 libraries, treated as four groups. Genes were filtered out if they failed to achieve cpm > 1 in at least 4 libraries, and the remaining log-cpm values were quantile normalized between libraries [61].

The comparison between technical replicates to check type I error rate control used only the four group B libraries. Genes were filtered out if they failed to achieve a cpm > 1 in at least two libraries and the log-cpm values for the 16745 remaining genes were quantile normalized. Samples are separated into all possible two-versus-two and three-versus-one combinations and a limma analysis using voom weights are carried out for each partition.

The false discovery rate analysis was conducted on the spike-in transcripts only. The ERCC Mixes 1 and 2 contain 92 transcripts spiked in at different concentrations. For this analysis, fragments were mapped to the known sequences of the spiked-in transcripts using Subread. The experiment is designed so that 23 transcripts have the same concentration in Mix 1 and Mix 2. The remaining transcripts are spiked-in in such a way that 23 transcripts are 4-fold more abundant in Mix 1, 23 are 1.5 higher in Mix 2 and 23 are 2-fold higher in Mix 2. A majority of the spike-in transcripts data are DE. We replicated the counts for each of the 23 non-DE transcripts three times, so that each non-DE transcript was treated as three different transcripts. This resulted in a dataset of 138 transcripts with half DE and half non-DE. Our analysis used read counts for the spike-in transcripts only. TMM-scale normalization [54] was used for all the analysis methods, except for DESeq and PoissonSeq, which have their own built-in normalization methods. No transcripts were filtered, except for PoissonSeq as its standard analysis includes the removal of probes with low counts. The genes that were filtered out by PoissonSeq were re-introduced to the end of the gene ranking ordered from largest mean count to lowest mean count.

Lymphoblastoid cell lines from Nigerian individuals

As part of the International HapMap Project, RNA samples were obtained from lymphoblastoid cell lines derived from 69 unrelated Nigerian individuals including 29 males and 40 females [40]. Sequencing performed using an Illumina Genome Analyzer II. Read counts, summarized by Ensembl gene, and transcript annotation were obtained from version 1.0.9 of the tweeDEseqCountData Bioconductor package [43], specifically from the data objects `pickrell1`, `annotEnsembl63` and `genderGenes`. Genes were filtered if they failed to achieve a cpm value of 1 in at least 20 libraries. Library sizes were scale normalized by the TMM method [54] using edgeR software [12] prior to the voom analysis.

Development stages of *D. melanogaster*

RNA-seq was used to explore the developmental transcriptome of Drosophila melanogaster [45]. Mapped read counts are available from the ReCount project [62]. Specifically the pooled version of the `modencodefly` dataset from the ReCount website [63] provides read counts summarized by Ensembl 61 gene IDs for 30 whole-animal biological samples. We discarded the larval, pupal and adult stages and kept only the 12 embryonic samples. Genes were retained in the analysis if they achieved cpm > 1 for any embryonic stage. Effective library sizes were estimated by TMM scale-normalization [54] using edgeR software [12] prior to the voom analysis.

Gene ontology analysis used the GOstats software package [64] and version 2.9.0 of the org.Dm.eg.db annotation package [65]. All GO terms mentioned in the Results section had Fisher's exact test p-values less than 10^{-10}.

C57BL/6J and DBA/2J inbred mouse strains

An RNA-seq experiment was carried out to detect differential striatal gene expression between the C57BL/6J (B6) and DBA/2J (D2) inbred mouse strains [66]. Profiles were made of 10 B6 and 11 D2 mice. Mapped read counts summarized by Ensembl 61 gene IDs were downloaded as the `bottomly` dataset from the ReCount website [63]. Genes were filtered out if they failed to achieve cpm > 1 in at least 4 libraries and the remaining log-cpm values are quantile normalized. The limma-voom analysis compared the two strains and included a batch effect correction for the Illumina flowcell in which each sample was sequenced. The voom mean-variance trend is shown in Figure 1b.

Software

The results presented in this article are carried out using R version 3.0.0 and software packages limma 3.16.2, edgeR 3.2.3, baySeq 1.14.1, DESeq 1.12.0, DSS 1.4.0, PoissonSeq 1.1.2 and tweeDEseqCountData 1.0.8. All of the packages mentioned above are part of the Bioconductor project [67, 68], except for PoissonSeq which is part of the Comprehensive R Archive Network [69]. The TSPM function, dated February 2011, was downloaded in March 2013 from the author's webpage [70].

The voom methodology proposed in the article is implemented in the voom function of the limma package. The limma-trend method was implemented using the log-cpm values from voom, then running the usual standard limma pipeline using the `eBayes` function with `trend=TRUE`. Hence the limma-trend pipeline was the same as that for voom except that weights were not using in the linear modeling fitting with `lmFit` but trend was turned on with `eBayes`. The limma package can be installed from the Bioconductor project repository [71].

All the count-based packages were used with the default differential expression pipelines as recommended in the software for each package. For edgeR 3.2.3 the default prior degrees of freedom for squeezing the genewise dispersions is 10. Note that this is a change on versions 3.0.X and earlier for which the default had been 20. For DSS the Wald test was used as recommended in the documentation. The DESeq defaults have changed considerably since the original publication. We used the DESeq function `estimateDispersions` with `sharingMode="maximum"` and `fitType="local"` and conducted tests using `nbinomTest`.

The different count-based packages implement different methods of compositional normalization [54]. For our simulations, there are no compositional differences between the libraries so there should be no need to estimate compositional normalization factors. For this reason we did not use `calcNormFactors` with edgeR or `estimateSizeFactors` with DESeq or `estNormFactors` with DSS. This should tend to improve the performance of the packages and to them more comparable, as any differences between the packages can be attributed to the statistical procedures rather than to differences between the normalization strategies.

Acknowledgements

The authors are grateful to Charles Wang and Leming Shi for the preliminary SEQC data and to Stephen Nutt for the mouse RNA-seq mouse data used to motivate aspects of the simulation study.

References

[1] Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proceedings of the National Academy of Sciences* 2001, **98**(9):5116–5121.

[2] Wright GW, Simon RM: **A random variance model for detection of differential gene expression in small microarray experiments**. *Bioinformatics* 2003, **19**(18):2448–2455.

[3] Smyth G: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments**. *Statistical applications in genetics and molecular biology* 2004, **3**:Article 3.

[4] Cui X, Hwang JG, Qiu J, Blades NJ, Churchill GA: **Improved statistical tests for differential gene expression by shrinking variance components estimates**. *Biostatistics* 2005, **6**:59–75.

[5] Smyth G, Michaud J, Scott H: **Use of within-array replicate spots for assessing differential expression in microarray experiments**. *Bioinformatics* 2005, **21**(9):2067–2075.

[6] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545–15550.

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

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

[9] Smyth G: **Limma: linear models for microarray data**. In *Bioinformatics and computational biology solutions using R and Bioconductor*. Edited by Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, New York: Springer 2005:397–420.

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

[11] Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, Mckernan KJ, Grimmond SM: **Stem cell transcriptome profiling via massive-scale mRNA sequencing**. *Nature Methods* 2008, **5**:613–619.

[12] Robinson M, McCarthy D, Smyth G: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**:139–140.

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

[14] Oshlack A, Robinson MD, Young MD: **From RNA-seq reads to differential expression results**. *Genome Biol.* 2010, **11**(12):220.

[15] Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, He M, Croucher NJ, Pickard DJ, Maskell DJ, Parkhill J, Choudhary J, Thomson NR, Dougan G: **A strand-specific RNA-seq analysis of the transcriptome of the typhoid bacillus** *Salmonella* **Typhi**. *PLoS Genetics* 2009, **5**(7):e1000569.

[16] Han X, Wu X, Chung WY, Li T, Nekrutenko A, Altman NS, Chen G, Ma H: **Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing**. *Proceedings of the National Academy of Sciences* 2009, **106**(31):12741–12746.

[17] Parikh A, Miranda ER, Katoh-Kurasawa M, Fuller D, Rot G, Zagar L, Curk T, Sucgang R, Chen R, Zupan B, Loomis WF, Kuspa A, Shaulsky G: **Conserved developmental transcriptomes in evolutionarily divergent species**. *Genome Biology* 2010, **11**(3):R35.

[18] Robinson MD, Smyth GK: **Small-sample estimation of negative binomial dispersion, with applications to SAGE data**. *Biostatistics* 2008, **9**(2):321–332.

[19] Zhou YH, Xia K, Wright FA: **A Powerful and Flexible Approach to the Analysis of RNA Sequence Count Data**. *Bioinformatics* 2011, **27**(19):2672–2678.

[20] Robinson MD, Smyth GK: **Moderated statistical tests for assessing differences in tag abundance**. *Bioinformatics* 2007, **23**(21):2881–2887.

[21] Hardcastle TJ, Kelly KA: **baySeq: Empirical Bayesian Methods For Identifying Differential Expression In Sequence Count Data**. *BMC Bioinformatics* 2010, **11**:422.

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

[23] Lund S, Nettleton D, McCarthy D, Smyth G: **Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates**. *Statistical Applications in Genetics and Molecular Biology* 2012, **11**(5):Article 8.

[24] Srivastava S, Chen L: **A two-parameter generalized Poisson model to improve the analysis of RNA-seq data**. *Nucleic Acids Res.* 2010, **38**(17):e170.

[25] Auer PL, Doerge RW: **A Two-Stage Poisson Model for Testing RNA-Seq Data**. *Statistical Applications in Genetics and Molecular Biology* 2011, **10**:Article 26.

[26] Li J, Witten D, Johnstone I, Tibshirani R: **Normalization, testing, and false discovery rate estimation for RNA-sequencing data**. *Biostatistics* 2012, **13**(3):523–538.

[27] Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM: **Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing**. *BMC Genomics* 2012, **13**:484.

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

[29] Ritchie M, Diyagama D, Neilson J, Van Laar R, Dobrovic A, Holloway A, Smyth G: **Empirical array quality weights in the analysis of microarray data**. *BMC bioinformatics* 2006, **7**:261.

[30] McCullagh P, Nelder JA: *Generalized Linear Models*. Boca Raton, Florida: Chapman & Hall/CRC, 2nd edition edition 1989.

[31] Wedderburn RWM: **Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method**. *Biometrika* 1974, **61**:439–447.

[32] Carroll RJ, Ruppert D: **A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model**. *Journal of the American Statistical Association* 1982, **77**:878–882.

[33] Nelder JA, Pregibon D: **An extended quasi-likelihood function**. *Biometrika* 1987, **74**:221–232.

[34] McCarthy DJ, Chen Y, Smyth GK: **Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation**. *Nucleic Acids Research* 2012, **40**(10):4288–4297.

[35] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nature Methods* 2008, **5**(7):621–628.

[36] Sartor MA, Tomlinson CR, Wesselkamper SC, Sivaganesan S, Leikauf GD, Medvedovic M: **Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments**. *BMC bioinformatics* 2006, **7**:538.

[37] **Sequencing Quality Control (SEQC) Project** [http://www.fda.gov/MicroArrayQC].

[38] **Ambion FirstChoice Human Brain Reference RNA** [http://products.invitrogen.com/ivgn/product/AM6050].

[39] Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, Causton H, Conley MP, Elespuru R, Fero M, Foy C, Fuscoe J, Gao X, Gerhold DL, Gilles P, Goodsaid F, Guo X, Hackett J, Hockett RD, Ikonomi P, Irizarry RA, Kawasaki ES, Kaysser-Kranich T, Kerr K, Kiser G, Koch WH, Lee KY, Liu C, Liu ZL, Lucas A, Manohar CF, Miyada G, Modrusan Z, Parkes H, Puri RK, Reid L, Ryder TB, Salit M, Samaha RR, Scherf U, Sendera TJ, Setterquist RA, Shi L, Shippy R, Soriano JV, Wagar EA, Williams M, Wilmer F, Wilson M, Wolber PK, Wu X, Zadro R: **The external RNA controls consortium: a progress report**. *Nature Methods* 2005, **2**(10):731–734.

[40] Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing**. *Nature* 2010, **464**(7289):768–72.

[41] Esnaola M, Puig P, Gonzalez D, Castelo R, Gonzalez JR: **A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments.** *BMC Bioinformatics* 2013, **14**:254.

[42] Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou SF, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang SP, Waterston RH, Wilson RK, Rozen S, Page DC: **The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes**. *Nature* 2003, **423**(6942):825–837.

[43] Gonzalez JR, Esnaola M: **tweeDEseqCountData: RNA-seq count data employed in the vignette of the tweeDEseq package** [http://www.bioconductor.org]. [Experiment data package].

[44] Carrel L, Willard HF: **X-inactivation profile reveals extensive variability in X-linked gene expression in females**. *Nature* 2005, **434**(7031):400–404.

[45] Graveley BR, Brooks N, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri G, van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S: **The developmental transcriptome of Drosophila melanogaster**. *Nature* 2011, **471**:473–479.

[46] Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al.: **Landscape of transcription in human cells**. *Nature* 2012, **489**(7414):101–108.

[47] Gonzalez-Porta M, Frankish A, Rung J, Harrow J, Brazma A: **Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene**. *Genome biology* 2013, **14**(7):R70.

[48] Bera AK, Bilias Y: **Rao's score, Neyman's $C\alpha$ and Silvey's LM tests: an essay on historical developments and some new results**. *Journal of Statistical Planning and Inference* 2001, **97**:9–44.

[49] Pregibon D: **Score tests in GLIM with applications**. In *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*. Edited by Gilchrist R, New York: Springer 1982:87–97.

[50] Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK: **Empirical Bayes in the presence of exceptional cases, with application to microarray data** 2013, [http://www.statsci.org/smyth/pubs/RobustEBayesPreprint.pdf].

[51] Oehlert GW: **A note on the delta method**. *The American Statistician* 1992, **46**:27–29.

[52] Cleveland WS: **Robust Locally Weighted Regression and Smoothing Scatterplots**. *Journal of the American Statistical Association* 1979, **74**:829–836.

[53] Oshlack A, Emslie D, Corcoran L, Smyth G: **Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes**. *Genome Biology* 2007, **8**:R2.

[54] Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome Biology* 2010, **11**(3):R25.

[55] Gale WA, Sampson G: **Good-Turing frequency estimation without tears**. *Journal of Quantitative Linguistics* 1995, **2**(3):217–237.

[56] Law CW, Chen Y, Shi W, Smyth GK: **Supplementary information for "Voom: precision weights unlock linear model analysis tools for RNA-seq read counts"** [http://bioinf.wehi.edu.au/voom].

[57] Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, Herman D, Jensen RV, Johnson CD, Lobenhofer EK, Puri RK, Scherf U, Thierry-Mieg J, Wang C, Wilson M, Wolber PK, Zhang L, Amur S, Bao W, Barbacioru CC, Lucas AB, Bertholet V, Boysen C, Bromley B, Brown D, Brunner A, Canales R, Cao XM, Cebula TA, Chen JJ, Cheng J, Chu TM, Chudin E, Corson J, Corton JC, Croner LJ, Davies C, Davison TS, Delenstarr G, Deng X, Dorris D, Eklund AC, hui Fan X, Fang H, Fulmer-Smentek S, Fuscoe JC, Gallagher K, Ge W, Guo L, Guo X, Hager J, Haje PK, Han J, Han T, Harbottle HC, Harris SC, Hatchwell E, Hauser CA, Hester S, Hong H, Hurban P, Jackson SA, Ji H, Knight CR, Kuo WP, LeClerc JE, Levy S, Li QZ, Liu C, Liu Y, Lombardi MJ, Ma Y, Magnuson SR, Maqsodi B, McDaniel T, Mei N, Myklebost O, Ning B, Novoradovskaya N, Orr MS, Osborn TW, Papallo A, Patterson TA, Perkins RG, Peters EH, Peterson R, Philips KL, Pine PS, Pusztai L, Qian F, Ren H, Rosen M, Rosenzweig BA, Samaha RR, Schena M, Schroth GP, Shchegrova S, Smith DD, Staedtler F, Su Z, Sun H, Szallasi Z, Tezak Z, Thierry-Mieg D, Thompson KL, Tikhonova I, Turpaz Y, Vallanat B, Van C, Walker SJ, Wang SJ, Wang Y, Wolfinger R, Wong A, Wu J, Xiao C, Xie Q, Xu J, Yang W, Zhang L, Zhong S, Zong Y, Slikker W: **The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements**. *Nature Biotechnology* 2006, **24**(9):1151–1161.

[58] Liao Y, Smyth GK, Shi W: **The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote**. *Nucleic Acids Research* 2013, **41**(10):e108.

[59] Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general-purpose read summarization program**. *Bioinformatics* 2013, [http://bioinformatics.oxfordjournals.org/content/early/2013/11/30/bioinformatics.btt656].

[60] Shi W, Liao Y: **Rsubread: a super fast, sensitive and accurate read aligner for mapping next-generation sequencing reads** [http://www.bioconductor.org]. [Software package].

[61] Bolstad BM, Irizarry RA, Åstrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185–193.

[62] Frazee AC, Langmead B, Leek JT: **ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets**. *BMC Bioinformatics* 2011, **12**:449.

[63] Frazee A, Langmead B, Leek J: **ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets** [http://bowtie-bio.sourceforge.net/recount].

[64] Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association.** *Bioinformatics* 2007, **23**(2):257–8.

[65] Carlson M: **org.Dm.eg.db: Genome wide annotation for Fly** [http://www.bioconductor.org]. [Annotation package].

[66] Bottomly D, Walter NA, Hunter JE, Darakjian P, Kawane S, Buck KJ, Searles RP, Mooney M, McWeeney SK, Hitzemann R: **Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays**. *PLoS One* 2011, **6**(3):e17820.

[67] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth GK, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome biology* 2004, **5**(10):R80.

[68] **Bioconductor: open source software for bioinformatics** [http://www.bioconductor.org].

[69] **Comprehensive R Archive Network** [http://www.r-project.org].

[70] Auer P, Doerge RW: **TSPM.R: R code for a two-stage Poisson model for testing RNA-seq data** 2011, [http://www.stat.purdue.edu/~doerge/software/TSPM.R].

[71] Smyth GK: **limma: Linear Models for Microarray Data** [http://www.bioconductor.org/packages/release/bioc/html/limma.html].