Voom! Precision weights unlock linear model analysis tools for RNA-seq read counts

Charity W Law[1,2] Yunshun Chen[1,2] Wei Shi[1,3]
Gordon K Smyth[1,4,5]

1 May 2013
(With corrections 17 May 2013)

(1) Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia. (2) Department of Medical Biology, (3) Department of Computing and Information Systems and (4) Department of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3010, Australia. (5) Corresponding author.

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 analysed 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. The voom method 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. This opens access for RNA-seq analysts to a large body of methodology developed for microarrays, allowing RNA-seq and microarray data to be analysed in closely comparable ways. The performance of voom and related limma-based pipelines is compared to that of edgeR, DESeq, baySeq, TSPM, PoissonSeq, and DSS. Simulation studies show that voom out-performs previous RNA-seq methods even when the data is generated according to the assumptions of the earlier methods. This is especially true when the sequence depths vary between RNA samples. Several data sets are also analysed to demonstrate how voom can handle heterogeneous data and complex experiments as well as facilitating pathway analysis and gene set testing methods.

Keywords

RNA-seq, differential expression, weights, linear modelling.

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 analyse intensity data from microarrays. This includes methods for differential expression analysis [1–3], gene set enrichment [4], gene set testing [5,6] and so on. One popular differential expression pipeline is that provided by the limma software package [7]. The limma pipeline includes linear modeling to analyse 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 [8]. 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 [9–12]. 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 analyse the RNA-seq read counts. For example, the limma package has been used to analyse the log-counts after normalization by sequencing depth [9, 13–15].

Later statistical publications argued that RNA-seq data should be analysed by statistical methods designed specifically for counts. Much interest has focused on the negative binomial 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 negative binomial dispersions [11,16,17], but this does not allow for gene-specific variation. A number of empirical Bayes procedures have been proposed to estimate the genewise dispersions [18–20]. Alternatively, Lund et al [21] proposed that the residual deviances from negative binomial 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 [22–24].

Despite the use of probabilistic distributions, all the statistical methods developed for RNA-seq counts rely on approximations of various kinds. Most rely on the statistical tests that are only asymptotically valid or, worse, are theoretically valid only when the dispersion is small. An exception is the negative binomial exact test [16], but this relies

on normalization and assumes that the dispersions are estimated without error. None of the empirical Bayes methods achieve the same adaptability or error rate control as the corresponding methods for microarrays, due to the mathematical intractability of count distributions as compared to the normal distribution. The quasi-likelihood methods make other assumptions, in particular that the residual deviances are analogous to residual sums of squares from a normal analysis of variance, but this can only be roughly true.

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 or log-counts is that the counts have markedly unequal variabilities. Large counts have much larger standard deviations than small counts, even after a log-transformation. 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 modelling of the mean-variance relationship inherent in a data generating process is the key to designing statistically powerful methods of analysis [25]. Such variance modelling may in fact take precedence over identifying the exact probability law that the data values follow [26–28]. 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 [27] whereby statistical methods based on the normal distribution are applied after estimating an mean-variance function for the data at hand.

Our approach is to estimate the mean-variance relationship robustly and non-parametrically. We work with log-counts normalized for sequence depth and generate a precision weight for each individual normalized observation. The normalized log-counts 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. This opens access for RNA-seq analysts to a large body of methodology developed for microarrays, not just differential expression but data exploration, gene set tests, pathway analysis, visual displays and so on.

This article compares the performance of voom and related limma-based pipelines to edgeR [18, 29], DESeq [11], baySeq [19], TSPM [23], PoissonSeq [24] and DSS [20]. Simulation studies show that voom out-performs previous RNA-seq methods even when the data is generated according to the assumptions of the earlier methods. Several data sets are also analysed to demonstrate how voom can handle heterogeneous data and complex experiments as well as facilitating pathway analysis and gene set testing methods.

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 consists of 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 assumed by a number of previous articles [11, 16, 18, 19, 29].

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 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 [30]. 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 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 [29]. 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 [29]. 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

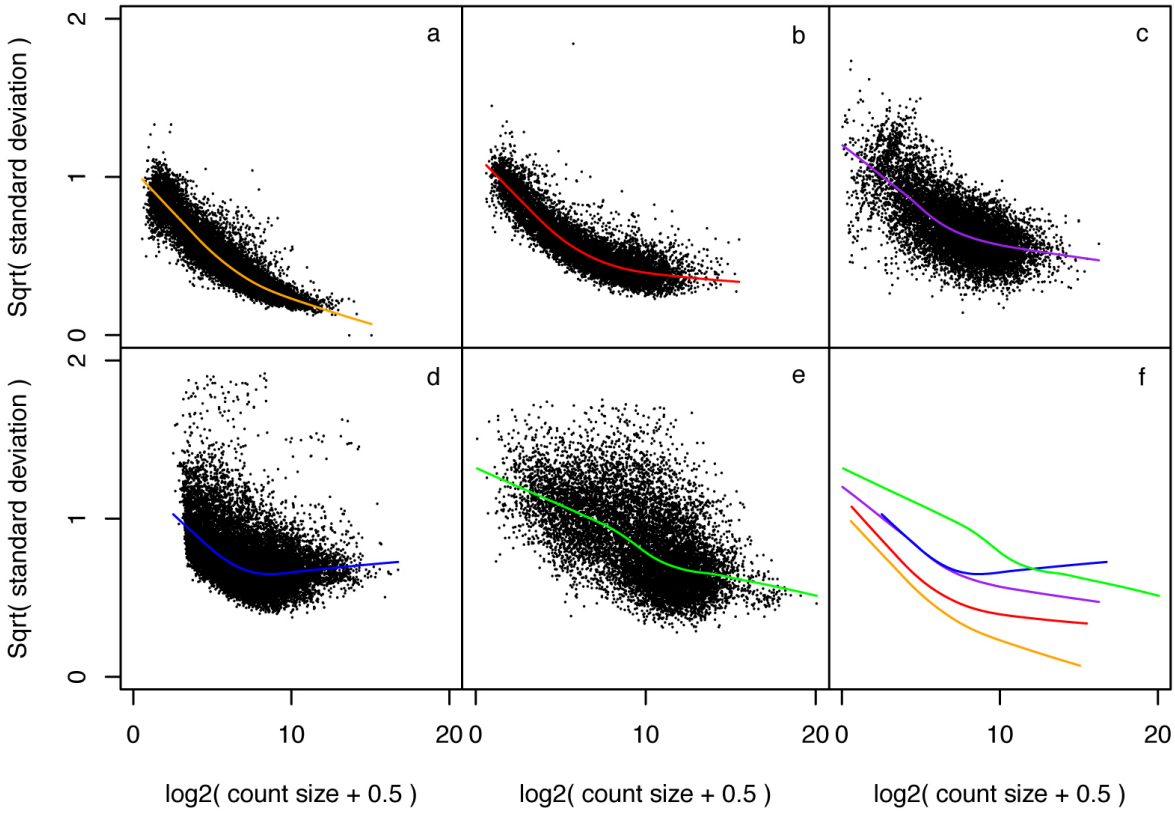

Figure 1: Mean-variance relationships. Gene-wise means and variances of RNA-Seq data represented by black points with a lowess trend. Plots are ordered by increasing levels of biological variation in datasets. Panel (a), voom trend in HBRR and UHRR genes in Sample A, B, C and D of SEQC project; technical 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; no technical or biological replicates. Panel (f), lowess voom trend from datesets above.

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 moderate to large counts.

Using log-cpm in a limma pipeline

A simple approach to analysing 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 package [3, 7]. A standard microarray pipeline can be expected to behave well for log-cpm if the counts are moderately large, but would ignore the mean-variance trend for lower counts. The microarray pipeline should behave much better if modified to include a mean-variance trend as part of the variance modelling [31]. Such a trend modification is readily available as part of the empirical Bayes of the limma package. We will call these two strategies, *limma-notrend* and *limma-trend* respectively.

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.

Voom: variance modelling at the observation-level

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

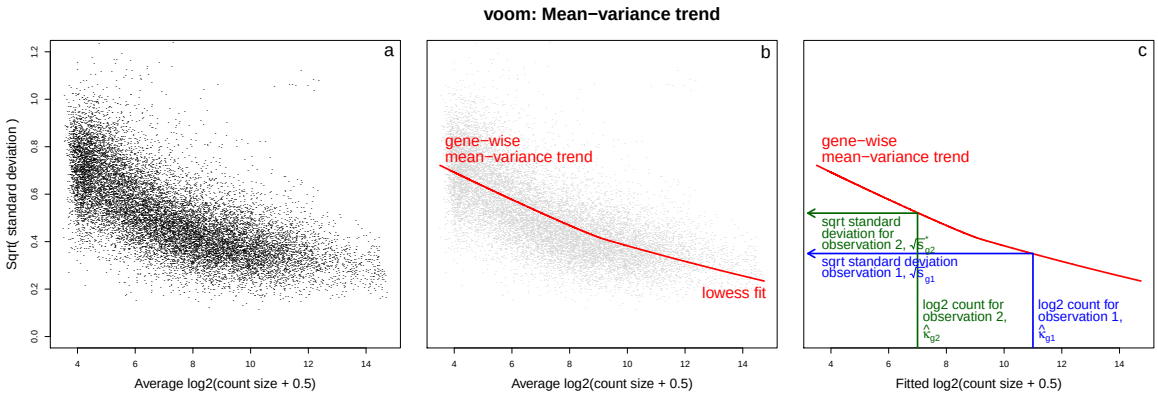

Figure 2: Voom mean-variance modelling. Panel (a), gene-wise square-root residual standard deviations are plotted against average log-count. Panel (b), a functional relationship between gene-wise means and variances is given by a robust lowess fit to the points. Panel (c), the mean-variance trend enables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

(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 controls the type I error rate correctly

Voom has been found 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 negative binomial model as assumed by a number of existing RNA-seq analysis methods.

Six RNA-seq count libraries were simulated with 10,000 genes. The distribution of cpm-values for each library was simulated to match the distribution that we observe for real RNA-seq data sets. To emphasize the effects of differing library sizes, the odd-numbered libraries were simulated to have a sequence depth of 20 million reads while the even-numbered libraries had the smaller sequence depth of 2 million reads. The negative binomial 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 a chisquare distribution on 40 degrees of freedom. The voom mean-variance trend for one such simulated dataset is shown in Figure 1c.

In the first set of simulations, no genes were truly differentially expressed between the groups. Voom was found to be very slightly conservative, returning a type I error rate almost equal to the nominal rate (Figure 3). The other normal or limma based pipelines were more conservative (Figure 3). After voom, edgeR-glm and edgeR-classic were the next closest to the nominal rate and were just slightly anti-conservative. The

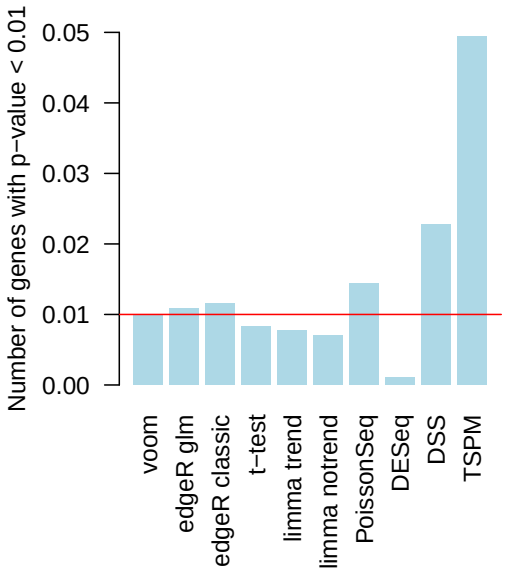

Figure 3: Type I error rates. The proportion of genes with p-value less than 0.01 is displayed for each method when no genes are truly differentially expressed. The red line shows the nominal type I error rate of 0.01. Results are averaged over 100 simulations. Methods are ordered from smallest to largest deviation from the nominal rate. With an empirical error rate of 0.0993, voom is the closest of any method to the nominal rate. The next closest are edgeR-glm and edgeR-classic, with estimated error rates of 1.08% and 1.15% respectively.

other specialist RNA-seq methods give error rates far from the nominal rates, with DESeq very conservative and PoissonSeq, DSS and TSPM increasingly anti-conservative. BaySeq was not included in this comparison because it doesn't return p-values.

To check voom's conservativeness on real data, we used a set of four replicate libraries from the SEQC Project [32]. All four libraries were Illumina HiSeq 2000 RNA-seq profiles of samples of Ambion's Human Brain Reference RNA (HBRR) [33]. 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

In the next set of 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. TSPM declares by far the most DE genes (450) at FDR < 0.2, but these are mostly false discoveries since there are actually only 200 truly DE genes. Of the methods that were found to control the type I error rate correctly, voom detects easily the most DE genes, with limma-trend being the closest competitor (Figure 4). DESeq has almost no power at all in this scenario. As expected, edgeR detects slightly more DE genes than voom, in proportion

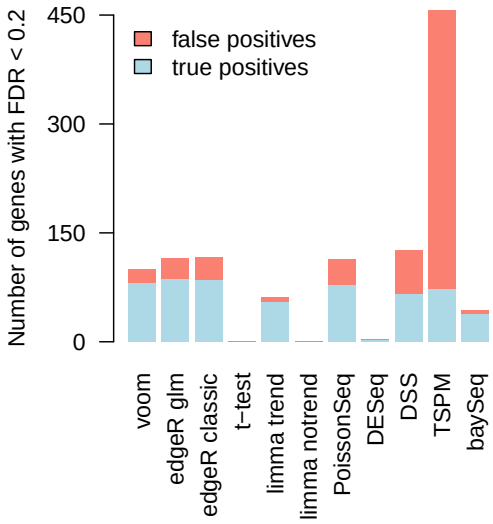

Figure 4: Power comparison with real differential expression. The total number of genes that are detected as DE at a 20% FDR are represented by stacked bars, where the number of true positives are in light blue, false positives in light red. 200 hundred genes are genuinely DE. Results are averaged over 100 simulations. Methods follow the same order as in Figure 2, with the addition of results from baySeq. Voom has the highest power of any method below in the line in Figure 2.

to its slightly under-estimated type I error rate. What is less expected is that voom and edgeR come close to matching PoissonSeq and DSS in numbers of genes detected, even though PoissonSeq and DSS gave distinctly higher type I error rates in Figure 3. Figure 3 and 4 together show that voom has the best power of any of the methods that correctly control the type I error rate.

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 200 truly DE genes in the simulation ahead of non-DE genes. Genes were ranked by posterior log-odds for voom and the limma methods, by posterior likelihoods for baySeq, and by p-value for the other methods. Note that log-odds of differential expression and p-value give identical rankings for limma-trend and limma-notrend but not for voom.

Results from the same power simulation described above show that voom gives the lowest number of false discoveries of any of the methods at any cutoff, with edgeR and voom-trend being the closest competitors (Figure 5). The next group of methods are PoissonSeq and baySeq. After those, the remaining methods give markedly higher FDRs.

Next we compared FDRs using spike-in control transcripts from the SEQC project [34]. 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

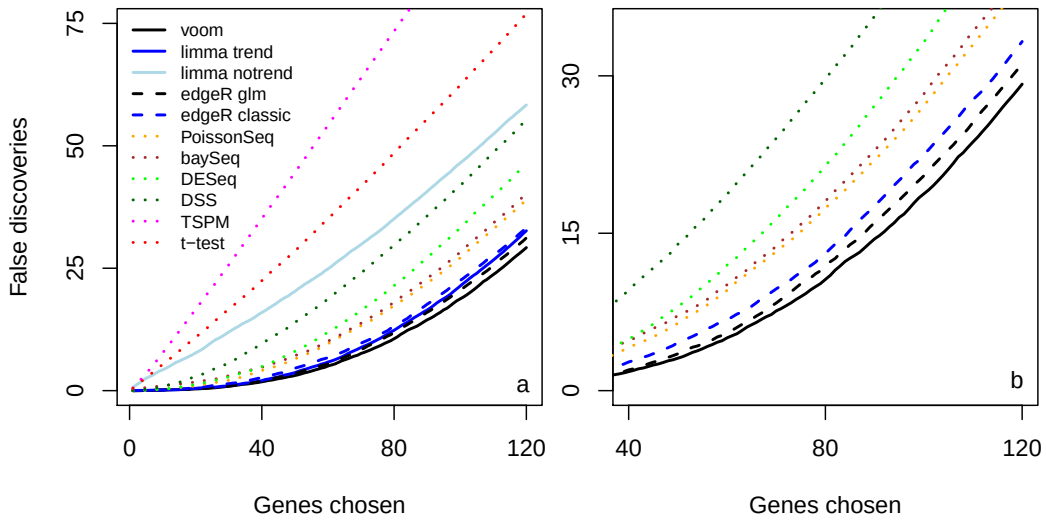

Figure 5: False discovery rate curves using simulated data with real differential expression. For a given number of top ranking genes, we plot the cumulative number of false discoveries averaged over 100 simulations. In panel (a), we represent analysis by limma using voom using a solid black line; limma fitting a gene level trend, a solid blue line; limma with no trends, a solid light-blue line; edgeR with generalize linear modelling, a dashed black line; edgeR using exact tests, a dashed blue line; PoissonSeq, a dotted orange line; baySeq, a dotted brown line; DESeq, a dotted light-green line; DSS, a dotted green line; TSPM, a dotted magenta line; and ordinary t-test, a dotted red line. Panel (b) is cropped and enlarged from panel (a) and displays voom and the most competitive non-normal methods only.

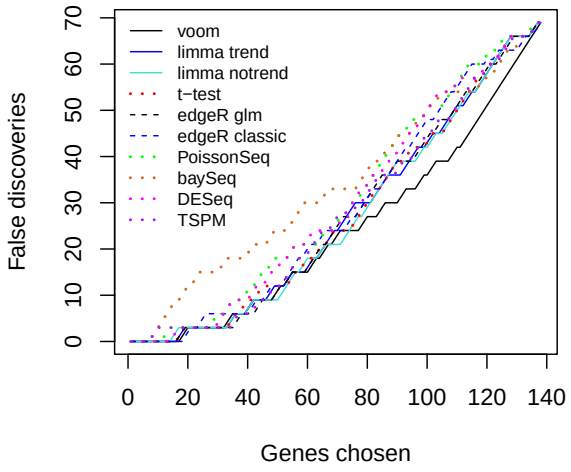

Figure 6: False discovery rate curves using SEQC spike-in data. For a given number of top ranking genes, we plot the cumulative number of false discoveries. Analysis by limma using voom is represented by a solid black line; limma-trend by a solid blue line; limma-notrend by a solid turquoise line; ordinary t-test, a dotted red line; edgeR-glm, a dashed black line; edgeR-classic, a dashed blue line; PoissonSeq, a dotted green line; baySeq, a dotted brownline; DESeq, a dotted magenta line; and TSPM, a dotted purple line.

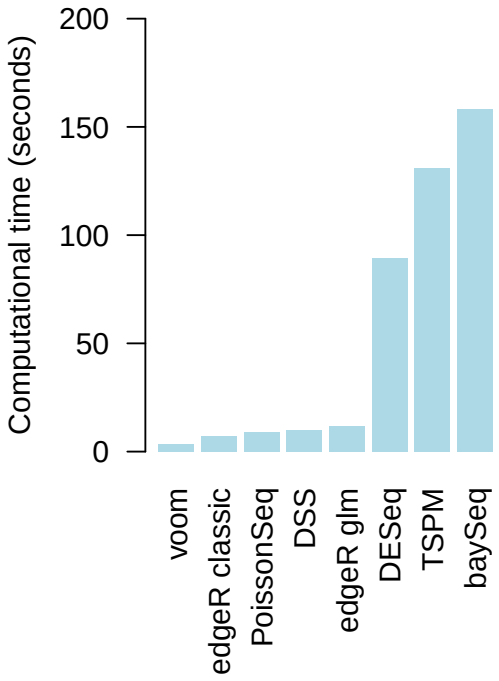

Figure 7: Computing times of RNA-Seq methods. The y-axis shows computing time required for the analysis of one simulated data using various statistical methods. Methods are ordered from quickest to most expensive.

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 other limma methods again being the closest competitors (Figure 6).

Voom is 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 demonstrated the performance of voom on RNA-seq data sets with small numbers of replicate libraries. To demonstrate the performance of voom on a heterogenenous 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 lymphblastoid cell lines from 29 male and 40 female unrelated Nigerian individuals [35]. Summarized read counts and gene annotation are provided by the Bioconductor tweeDEseqCountData package [36].

Table 1: Genes differentially expressed between males and females. Results for the 16 most significantlly differentially expressed genes for the Nigerian data.

	Symbol	Chr	logFC	AveExpr	t	P.Value	adj.P.Val	B
ENSG00000229807	XIST	X	-9.815	3.8084	-36.4	7.03e-48	1.19e-43	74.8
ENSG00000099749	CYorf15A	Y	4.251	0.3146	28.3	1.25e-40	1.05e-36	68.2
ENSG00000157828	RPS4Y2	Y	3.281	3.3081	26.5	9.38e-39	5.27e-35	72.6
ENSG00000233864	TTTY15	Y	4.897	-0.5538	25.9	4.31e-38	1.82e-34	64.0
ENSG00000131002	CYorf15B	Y	5.440	-0.1710	23.2	4.81e-35	1.62e-31	60.0
ENSG00000198692	EIF1AY	Y	2.398	2.6806	20.5	1.09e-31	3.07e-28	58.6
ENSG00000165246	NLGN4Y	Y	5.330	-0.4916	19.7	1.26e-30	3.03e-27	52.4
ENSG00000213318	RP11-331F4.1	16	4.293	2.2654	19.3	4.44e-30	9.34e-27	54.1
ENSG00000129824	RPS4Y1	Y	2.781	4.7118	17.6	9.28e-28	1.74e-24	51.5
ENSG00000183878	UTY	Y	1.878	2.7430	16.6	2.88e-26	4.85e-23	47.7
ENSG00000012817	KDM5D	Y	1.470	4.7046	14.9	1.45e-23	2.22e-20	42.6
ENSG00000146938	NLGN4X	X	4.472	-0.7801	14.8	2.09e-23	2.94e-20	38.9
ENSG00000243209	AC010889.1	Y	2.528	-0.0179	14.5	5.48e-23	7.11e-20	37.9
ENSG00000067048	DDX3Y	Y	1.671	5.3077	13.4	3.05e-21	3.67e-18	37.5
ENSG00000006757	PNPLA4	X	-0.988	2.5341	-10.4	4.78e-16	5.38e-13	25.7
ENSG00000232928	RP13-204A15.4	X	1.434	3.2506	10.3	1.02e-15	1.08e-12	25.2

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 [36,37]. A ROAST gene set test confirmed that these genes collectively 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 ($P = 10^{-18}$).

We also examined 46 X-chromosome genes that have been reported to escape X-inactivation [36, 38]. These genes were significantly up-regulated in females (ROAST $P = 0.0001$, CAMERA $P = 6 \times 10^{-11}$). 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).

Development stages of *D. melanogaster*

The possibilities of linear modeling are so rich that it is impossible to select a representative example. Voom and limma could be used to analyse any gene-level RNA-seq differential expression experiment, including those with multiple experimental factors [29]. Here we give a novel analysis illustrating the use of quadratic regression to analyse a time-course study.

RNA-Seq was used to explore the developmental transcriptome of Drosophila melanogaster [39]. 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 analyse the 12 RNA-seq libraries from these embryonic stages. We seek to iden-

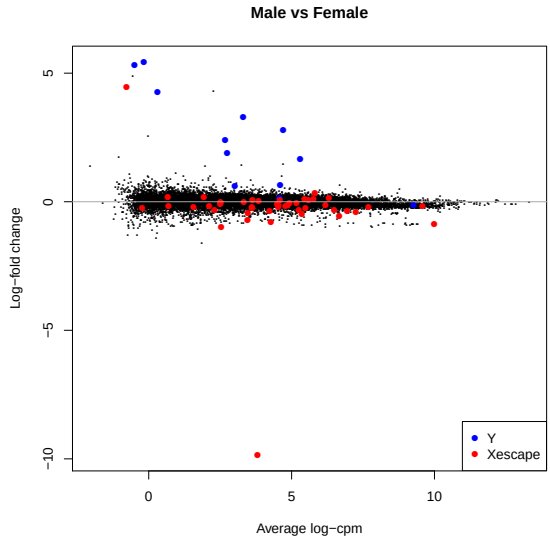

Figure 8: MA-plot with male and female specific genes highlighted. The log-fold change of each gene is plotted against its average log-cpm for the comparison between males and females. Genes on the male-specific region of the Y chromosome genes are highlighted blue and are consistently up-regulated in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

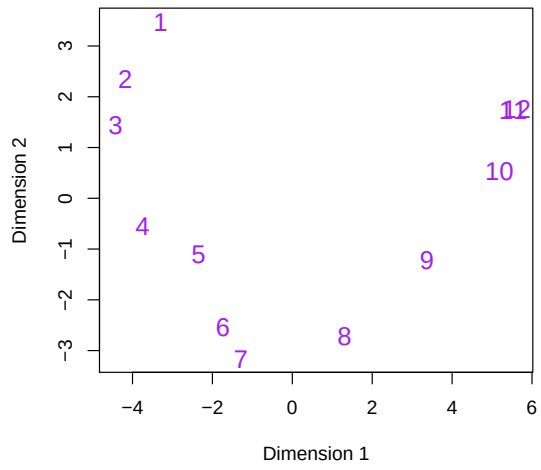

Figure 9: Multidimensional scaling plot of *D. melanogaster* embryonic stages. Distances are computed from the log-cpm values. The 12 successive embryonic developmental stages are labeled 1 to 12, from earliest to latest.

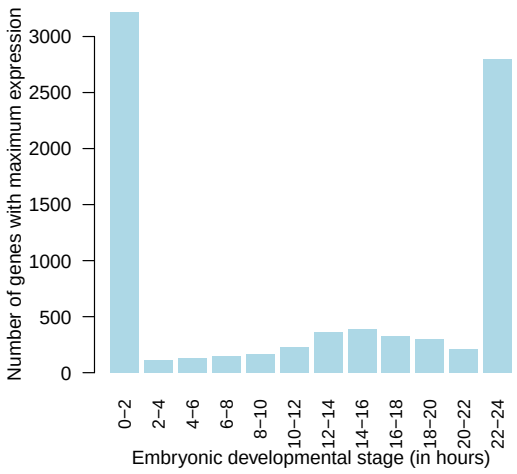

Figure 10: Number of genes associated with each *D. melanogaster* embryonic stage. For a quadratic regression on embryonic developmental stages, the number of genes with maximum expression at each of the stages is recorded.

tify 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 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 substantial set of genes with peak activity between 14–16 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

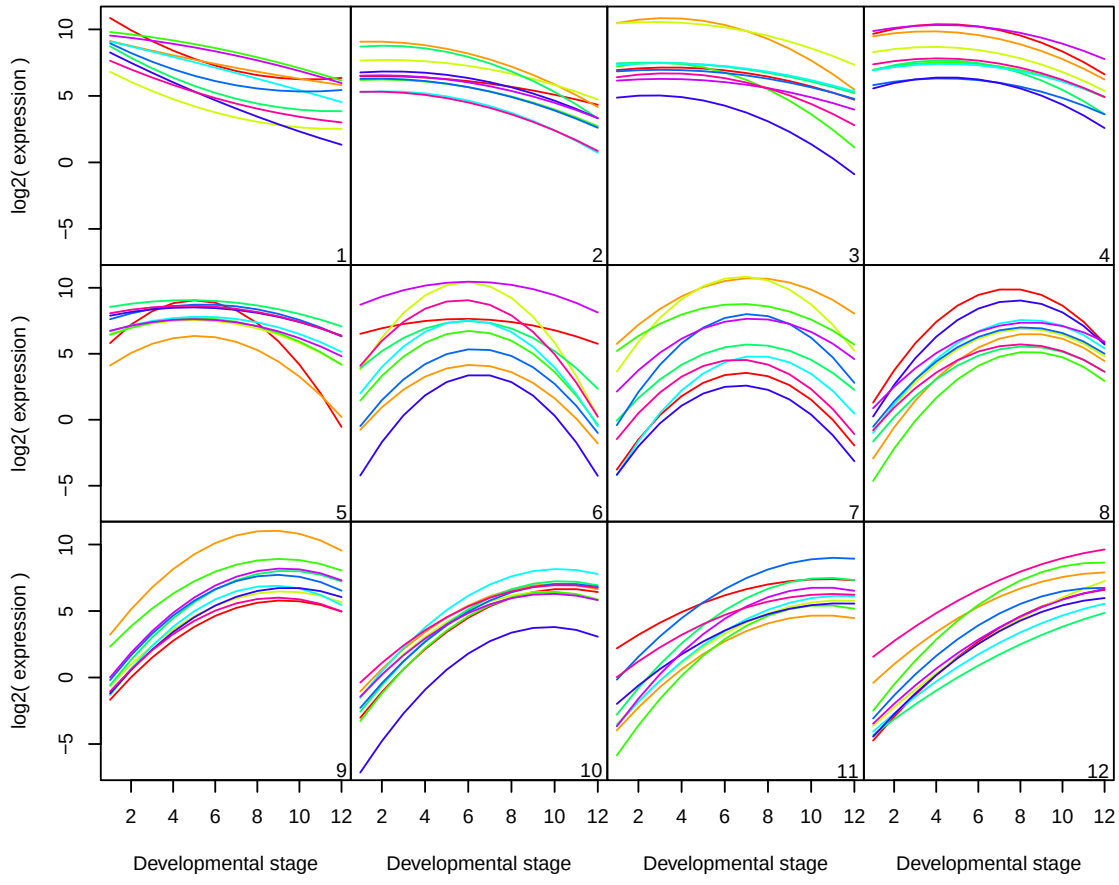

Figure 11: Trends in genes which peak at each embryonic developmental stage. The fitted quadratic regression on embryonic developmental stages is displayed for the top ten ranking genes as determined by FDR, which peak at each stage. Panel (1) displays gene expression for ten genes over the 24 hour period, where maximum expression occurs at stage 1 of embryonic development. Panel (2), maximum expression occurring at stage 2. Panel (3), maximum expression occurring at stage 3; and so on.

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 genes. 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. Moreover, 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 (Figure 4b of [40]). 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 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. Indeed, voom is the clear best performer across our simulations and comparisons. The next best performer is edgeR, with the main disadvantage that it slightly underestimates the true false discovery rate in small samples. The performance advantage of voom over the other methods considered is substantial in our simulations, despite the simulations being conducted under the same distributional assumptions as made by existing methods such as the DESeq package [11].

The worst performer in our simulation was TSPM, presumably because we have simulated from negative binomial distributions, which have quadratic mean-variance relationships, whereas TSPM assumes a linear mean-variance relationship [23]. 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.

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 negative binomial 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 negative binomial

distribution captures most of the mean-variance trend, the negative binomial dispersion still shows a non-ignorable trend with gene abundance [11, 17, 29]. 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 negative binomial dispersions [11, 17, 21, 29], 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 hyperpameters and exact small sample distributions for the test statistics.

Fourth, the use of normal distribution approximations in conjunction with variance modeling is partly supported by generalized linear model theory. Rao's score test [41] 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 [42]. Score tests have similar performance to likelihood ratio tests when the null hypothesis is true or when the changes being detected are relatively small.

When the counts are very small, normal approximations mostly inevitably be quite rough. Our experiments suggest that voom is more conservative than edgeR and related negative binomial methods when the counts are small.

Our simulations were conducted with a ten-fold difference in sequencing depth between different libraries. If the libraries are simulated with equal depths, then limma-trend performs similarly to voom (data not shown). With equal library sizes, the differences in between the methods is not so marked as in the simulations presented. Nevertheless, voom and limma-trend remain the equal best performers (data not shown).

On the other hand, we could have presented simulations more to voom's advantage, for example by simulating data from distributions other than negative binomial. A model in which the true expression levels for each gene follow a log-normal distribution between replicates would lead to a log-normal-Poisson mixture distribution for the counts. This would be at least as scientifically reasonable as the negative binomial distribution, and would improve the performance of voom relative to edgeR, DESeq 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.

Voom allows RNA-seq and microarray data to be analysed in closely comparable ways. It also 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 analysed 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

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

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 has 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. 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 [43], 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, 5, 6, 29], 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 $\mathbf{y}_g$ is the vector of log-cpm values for gene g and X is the design matrix with the $\mathbf{x}_i$ as rows. Interest centers on testing whether one or more of the β_{gj} are equal to zero,

Voom variance modelling

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 $\hat{\beta}^*_{gj}$, 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 loess 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 [44] is statistically robust [45] 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 $\text{lo}(\hat{\lambda}_{ga})$ 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 [5] is a self-containing gene set testing procedure and CAMERA [6] is a competitive gene set testing procedure. Both procedures have considerable flexibility as they have the ability to test the association of a genomic pathway or gene set signature with any contrast defined in 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() in 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 [46]. 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 cpm values in the simulations were based on a real RNA-seq dataset. Using a set of six RNA-seq profiles of mouse T cells, we used the goodTuringProportions function of the edgeR package [10], which implements the Good-Turing algorithm [47], to predict the true proportion of total RNA attributable to each gene. The simulated counts were generated in such a way that the baseline distribution of expected counts for each library, as a proportion of the total library size, matched the empirical distribution of proportions for the mouse data. Counts were simulated from negative binomial distributions, with expected values determined by the baseline proportions and the library sizes.

The negative binomial dispersions were generated as follows. The trend in the dispersions was set to be ψ_{gi} with $\psi_{gi}^{1/2} = 0.2 + \lambda_{gi}^{-0.5}$ 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 $\phi_{gi} = \psi_{gi}\delta_g^2$ where $40/\delta_g \sim \chi^2_{40}$.

For each simulated data set, genes with less than 10 reads across all samples were filtered from the analyse.

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.

SEQC data

The SEQC project, also known as MAQC-III, aims to provide a comprehensive study of next-generation sequencing technologies [32]. Here we analyze a pilot SEQC dataset of 16 RNA-seq libraries in four groups. The groups are labeled A–D and are closely analogous to the similarly labeled RNA samples used in the earlier microarray quality control study [48]. 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 [49]. Fragment counts were summarized by Entrez Gene ID using the featureCounts function of version 1.8.2 of the Bioconductor package Rsubread [50]. 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 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 [51].

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 [46] 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 [35]. Sequencing performed using an Illumina Genome Analyser II. Read counts, summarized by Ensembl gene, and transcript annotation were obtained from the tweeDEseqCountData Bioconductor package [36], 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 [46] using edgeR software [10] prior to the voom analysis.

Development stages of *D. melanogaster*

RNA-seq was used to explore the developmental transcriptome of Drosophila melanogaster [39]. Mapped read counts are available from the ReCount project [52]. Specifically the pooled version of the `modencodefly` dataset from the ReCount website [53] 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 [46] using edgeR software [10] prior to the voom analysis.

Gene ontology analysis used the GOstats [54] and org.Dm.eg.db [55] Bioconductor packages. All GO terms mentioned 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 [56]. 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 [53]. 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 methodology proposed in the article is implemented in the voom function of the limma package. The results presented in this article are from software packages limma 3.14.4, edgeR 3.0.8, baySeq 1.12.0, DESeq 1.10.1, DSS 1.0.0, PoissonSeq 1.1.2 and tweeDEseqCountData 1.0.8. All of the packages mentioned above are part of the Bioconductor project [57], except for PoissonSeq which is part of the Comprehensive R Archive Network [58]. The TSPM function, dated February 2011, was downloaded in March 2013 from the author's webpage [59].

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 as a basis for the simulation studies.

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] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al.: **Gene set enrichment**

analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545–15550.

[5] 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.

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

[7] 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 New York 2005:397–420.

[8] Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat. Rev. Genet.* 2009, **10**:57–63, [http://www.nature.com/nrg/journal/v10/n1/abs/nrg2484.html].

[9] 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.

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

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

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

[13] 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.

[14] 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.

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

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

[17] 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.

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

[19] Hardcastle TJ, Kelly KA: **baySeq: Empirical Bayesian Methods For Identifying Differential Expression In Sequence Count Data**. *BMC Bioinformatics* 2010, **11**:422, [http://www.biomedcentral.com/1471-2105/11/422].

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

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

[27] 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.

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

[29] McCarthy D, Chen Y, Smyth G: **Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation**. *Nucleic acids research* 2012, **40**(10):4288–4297.

[30] 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.

[31] 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.

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

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

[34] Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, Causton H, Conley MP, Elespuru R, Fero M, et al.: **The external RNA controls consortium: a progress report**. *Nature Methods* 2005, **2**(10):731–734.

[35] 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, [http://www.nature.com/nature/journal/v464/n7289/full/nature08872.html].

[36] Gonzalez JR, Esnaola M: *tweeDEseqCountData: RNA-seq count data employed in the vignette of the tweeDEseq package*[http://www.creal.cat/jrgonzalez/software.htm]. [R package version 1.0.9].

[37] Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, et al.: **The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes**. *Nature* 2003, **423**(6942):825–837.

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

[39] 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.

[40] 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.

[41] 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.

[42] Pregibon D: **Score tests in GLIM with applications**. In *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*, Springer 1982:87–97.

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

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

[45] 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.

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

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

[48] Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, et al.: **The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements**. *Nature Biotechnology* 2006, **24**(9):1151–1161.

[49] Liao Y, Smyth GK, Shi W: **The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote**. *Nucleic Acids Research* 2013, **41**:16 pages.

[50] Shi W, Liao Y, Dai JZ: **Rsubread version 1.8.2: a super fast, sensitive and accurate read aligner for mapping next-generation sequencing reads** 2012, [http://www.bioconductor.org].

[51] 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.

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

[53] Frazee A, Langmead B, Leek J: **ReCount** [http://bowtie-bio.sourceforge.net/recount].

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

[55] Carlson M: *org.Dm.eg.db: Genome wide annotation for Fly.* [R package version 2.9.0].

[56] 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.

[57] Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al.: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome biology* 2004, **5**(10):R80.

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

[59] 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].