Zhou and Robinson *Genome Biology* (2015) 16:222
DOI 10.1186/s13059-015-0781-3

Genome Biology

CORRESPONDENCE Open Access

Do count-based differential expression methods perform poorly when genes are expressed in only one condition?

Xiaobei Zhou[1,2] and Mark D. Robinson[1,2*]

Abstract

A response to 'Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data' by Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND and Betel D in Genome Biology, 2013, 14:R95

Background

Statistical methods for determining transcriptional changes between (replicated) groups of cell populations using RNA sequencing (RNA-seq) data are now quite mature. Several themes that emerged from the past decade of modeling microarray data apply analogously to RNA-seq data: parameter moderation is critical, multiple testing corrections are necessary and flexible frameworks (e.g., linear models) to account for the effect of covariates are essential. For RNA-seq data, popular packages such as edgeR, DESeq and DESeq2 [1–3] perform detailed modeling of the dispersion–mean relationship, with variations on fitting a dispersion by mean trend and moderating estimates toward the trend. Likewise, careful modeling of the mean–variance relationship of transformed data has been proven effective, essentially 'unlocking' the world of heteroskedastic linear regression [4].

A recent report in Genome Biology from Rapaport and co-authors claimed that some methods, namely PoissonSeq [5] and limma [6], 'have improved modeling of genes expressed in one condition', where they showed a striking difference in the ability to separate differential expression (DE) [7]. From a methodological perspective, this result caught our interest and prompted us to understand how aspects of the all-zero-in-one-condition manifest undesirable properties in count-based models. Briefly, (i) we found a coding error in the calculation of edgeR's signal-to-noise (S/N) metric and (ii) our re-analysis suggests that count-based methods perform as well or better than other methods, counter to the original conclusion.

The Rapaport manuscript is an excellent model of modern bioinformatics research, in terms of making processed data and code available that reproduce figures from their manuscript. In many cases, the small details can be important and this open-source model facilitates quick access in understanding precisely what settings were used. We fully support this model and by default, also make our code available. In this correspondence, we investigate the genesis of differences in method performance that Rapaport and co-authors observed and provide our view of how performance results can be sensitive to decisions made.

Genes expressed in only one condition

We first briefly summarize the analysis that Rapaport and colleagues reported, with respect to the all-zero-in-one-condition case.

Using gene-level read counts, they isolated genes that exhibit zero-counts across all replicates of a single condition; in general, the number of such genes is related to the depth of sequencing dedicated to each sample, with deeper sequencing resulting in fewer such cases. The dataset in question, comparing GM12892 cells to H1-hESC cells [8], with three and four replicates, respectively, had typical read depths for such experiments (16–39 million mapped reads). They used the following pipeline: (i) from the count table, generate DE P values for several

*Correspondence: mark.robinson@imls.uzh.ch
[1] SIB Swiss Institute of Bioinformatics, University of Zurich, CH-8057 Zurich, Switzerland
[2] Institute of Molecular Life Sciences, University of Zurich, CH-8057 Zurich, Switzerland

methods; (ii) calculate S/N using 'normalized' data; (iii) plot negative log P value versus S/N, where they expect a monotonic positive dependency (correlation); and, (iv) generate receiver operating characteristic (ROC) curves with thresholds on the S/N to illustrate the ability to separate low S/N ($<$3) from high S/N ($>$3).

They highlighted that count-based methods such as DESeq and edgeR, which infer changes in expression via the negative binomial (NB) model, do not perform very well in this case. It is worth noting that this is a non-standard use of ROC curves: here, all genes are strictly DE, but they vary in their magnitude of change. So, the ROC curve represents the ability to separate low S/N from high S/N. Rapaport and colleagues postulated that the NB model reduces to Poisson (dispersion ≈ 0) and lacks the ability to handle the 'wide variations' in gene counts among replicate libraries. Our aim with this report is to understand the origins of this result, whether it is a shortcoming of the dispersion estimation strategy or in the inference machinery, since parameter estimates are on the boundary of the parameter space.

Signal-to-noise has some potential limitations

We became interested in the suitability and robustness of the S/N metric itself, since it forms the basis for the 'truth' in Rapaport's ROC result. In theory, the S/N of the non-zero observations should accurately reflect the significance of model-based P values for the expressed-in-one-condition versus zero differences. In practice, however, there are some potential difficulties: the sample sizes are small and therefore, the S/N itself is subject to considerable estimation uncertainty; it is well known that for count data the variance is intimately tied to the mean, so it is not clear whether S/N should be calculated on a linear scale. In addition, a notable aspect of the Rapaport ROC comparison is that while the same S/N cutoff (= 3) is used across all methods, different sets of true and false DE labels are used; this makes the curves difficult to compare, since both the truth and score change by method. We explore these issues here.

Table 1 and Fig. 1 give illustrative examples of the differences in the originally calculated S/N between edgeR and voom. Figure 1 gives a scatter plot of S/N calculated on each method's normalized data, highlighting in some cases large differences. Table 1 shows the top ten genes for both edgeR's (estimated) false discovery rate (FDR) and calculated S/N. (The full table of zero-counts, differential statistics and S/N is given in Additional file 1.) Here, it is evident that several genes that show little evidence for DE, have very high S/N for edgeR but not for voom (e.g., C17orf66, TM4SF19 and NPY1R). However, the P values seem to reflect appropriately the magnitude of evidence for DE, although they are on drastically different scales between edgeR and voom (see 'Discussion' for further commentary on this). In addition, several genes that show the largest evidence against the null hypothesis (e.g., PLEK, MS4A1, etc.) show relatively low S/N for edgeR and would be counted as false discoveries (according to a S/N = 3 cutoff), while voom's higher S/N would result in these counted as true positives. Therefore, it is not clear whether the ROC curve reflects the accuracy of the S/N calculation itself or of the statistical method's capabilities. Upon investigation, the differences in S/N exhibited in Fig. 1 resulted from a code error in the original report (see Additional file 2: Fig. S1).

Another aspect to understand is the scale on which the S/N is calculated. As is well known with count data, the variance is related to the mean. In particular, using the NB parameterization with mean μ and variance $\mu(1 + \mu\phi)$, the theoretical S/N is then:

$$\mathrm{S/N} = \frac{1}{\sqrt{1/\mu + \phi}},$$

which implies S/N $\to \phi^{-1/2}$ with sufficiently large μ. Thus, depending on the mean, the S/N calculation is capturing the (inverse square root of) dispersion. For the ENCODE data, this relationship is shown in Additional file 2: Fig. S2. Since the S/N calculations are most relevant when the variance is independent of the mean, we explored how transforming the data, which alters the mean–variance relationship, affects the results of the ROC comparisons that Rapaport and co-authors performed. Figure 2a–c show mean–variance relationships for S/N calculated on different scales and Fig. 2d–f highlight their corresponding ROC performances. In all cases, the true/false labels for the ROC curves are the same across methods; specifically, counts-per-million from edgeR are used to base the S/N calculation. Since the scale of data changes the scale of S/N, true genes are selected according to S/N $>$ 40th percentile and false as the lowest 20 % of S/N to give a gray zone of uncertainty in the middle. (Additional file 2: Fig. S3 gives alternative settings for these cutoffs, but the results are unaffected.) Figure 2d shows similar results to the original Rapaport study, whereas Fig. 2e, f show a remarkable reversal in performance, giving clear evidence for our earlier concern regarding the S/N calculation.

Count-based methods perform well on zero-in-one-condition simulation

Given recent efforts in simulating RNA-seq count tables [9–11], we tried to create a representative simulation for the zero-in-one-condition situation. The simulation was designed as follows: (i) generate a dataset with no DE and (ii) randomly select genes across the spectrum of expression levels and set counts for one condition (chosen at random) equal to zero to represent 'true' DE genes.

Table 1 Top ten genes originally calculated using S/N (for edgeR-normalized data; first ten rows) and top ten genes calculated using FDR for DE (edgeR *P* values; second ten rows). The table includes the counts-per-million table (A = GM12892 and B = H1-hESC), S/N and estimated false discovery rate (FDR) for edgeR and limma-voom for the ENCODE dataset comparing three replicates of GM12892 to four replicates of H1-hESC

Id	A1	A2	A3	B1	B2	B3	B4	edgeR S/N	edgeR FDR	voom S/N	voom FDR
MIPOL1	0.0	0.0	0.0	237.1	232.5	226.0	227.5	45.75	7.47e−31	3.79	3.32e−07
AQP4	0.0	0.0	0.0	46.1	45.0	46.7	44.4	43.87	1.44e−14	3.90	1.99e−06
FAM19A4	0.0	0.0	0.0	142.1	131.1	143.8	131.4	20.21	1.91e−24	4.06	4.72e−07
C17orf66	2.3	2.1	2.1	0.0	0.0	0.0	0.0	16.96	1.89e−02	2.72	7.40e−04
TM4SF19	3.5	3.2	3.2	0.0	0.0	0.0	0.0	16.96	3.00e−03	2.72	2.39e−04
SOX1	0.0	0.0	0.0	7.5	6.7	6.5	6.3	13.02	8.37e−05	4.33	1.27e−04
HPGD	0.0	0.0	0.0	22.6	21.0	19.6	19.0	12.97	5.08e−09	4.56	7.48e−06
LOC100131176	0.0	0.0	0.0	17.9	15.3	18.7	17.2	12.02	5.14e−08	3.51	1.49e−05
ZNF385D	0.0	0.0	0.0	135.5	155.0	155.0	132.3	11.80	2.60e−25	4.67	3.70e−07
NPY1R	0.0	0.0	0.0	209.8	179.9	179.3	208.5	11.38	1.33e−27	2.95	6.77e−07
PLEK	25 082.8	12 622.5	11 394.8	0.0	0.0	0.0	0.0	2.16	1.79e−216	16.11	9.36e−09
MS4A1	25 455.1	14 937.7	12 886.8	0.0	0.0	0.0	0.0	2.63	2.62e−215	7.26	1.60e−08
SLAMF1	7 407.2	4 859.3	4 283.2	0.0	0.0	0.0	0.0	3.32	2.98e−165	5.11	2.53e−08
CCL3	11 057.5	3 413.1	3 544.1	0.0	0.0	0.0	0.0	1.37	4.62e−165	5.13	2.15e−08
FCRLA	7 742.0	2 979.1	3 879.8	0.0	0.0	0.0	0.0	1.92	1.01e−161	6.08	1.84e−08
RGS1	9 939.5	9 967.3	7 741.6	0.0	0.0	0.0	0.0	7.22	4.53e−159	2.66	1.44e−07
DPPA4	0.0	0.0	0.0	14 580.2	15 215.1	14 745.3	10 617.3	6.47	2.37e−158	12.02	1.84e−08
TDGF1	0.0	0.0	0.0	15 699.8	15 481.1	13 374.5	8 522.5	3.98	6.37e−157	15.48	1.84e−08
SFRP2	0.0	0.0	0.0	14 673.3	15 229.5	13 067.2	7 234.4	3.43	1.84e−153	9.87	2.15e−08
BLK	9 943.0	2 954.7	2 351.8	0.0	0.0	0.0	0.0	1.20	2.98e−147	3.28	5.17e−08

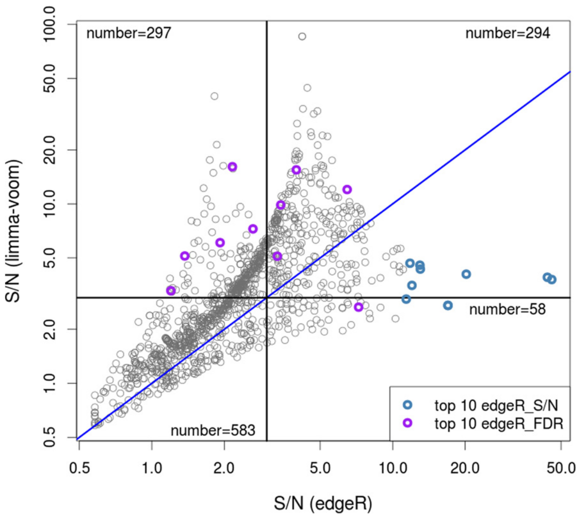

Fig. 1 Scatter plot of S/N (signal-to-noise) for limma-voom and edgeR f or the ENCODE dataset. S/N is calculated from the non-all-zero condition. More information about the colored points is given in Table 1

As previously, we sampled NB mean and dispersion estimates from the joint distribution of estimates using a large dataset (here, from [12]) and filtered out extreme dispersion values. Altogether, 30,000 features were generated in a 5 versus 5 two-group comparison and zero-counts were introduced to 5 % of the features. To reflect that zeros occur somewhat more often at lower expression across various datasets (see Additional file 2: Fig. S4), we increased the frequency of zero-counts at low expression strength.

Based on the results of this simulation (Fig. 3), ROC curves with the method's 5 % FDR highlighted (panel a) and plots of true positive rate versus achieved FDR (panel b), we again see that count-based models perform well in the zero-in-one-condition situation. In addition, we explored the postulation that the NB model is reduced to a Poisson in these zero-count situations. By comparing the dispersion estimates calculated from the single non-zero condition to the original non-zero-in-both-conditions data, it does not appear that the dispersion estimates are drastically reduced (see Additional file 2: Fig. S5).

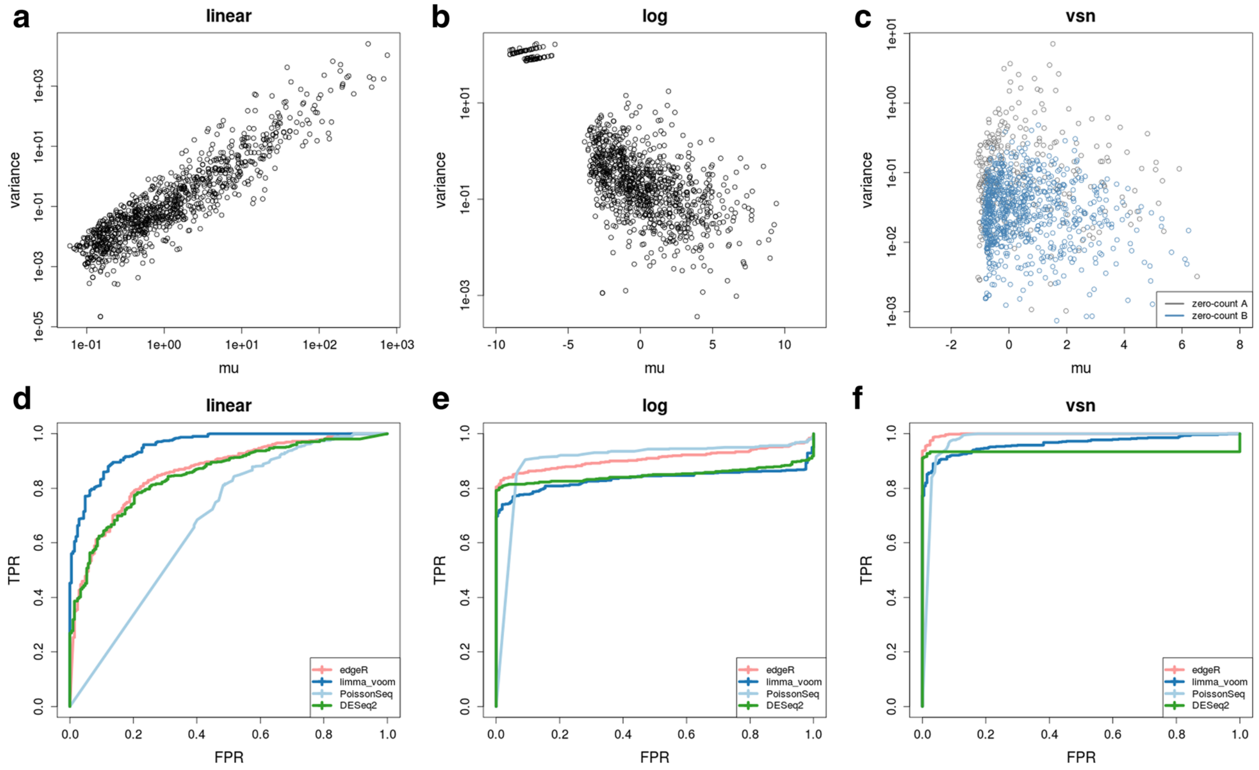

Fig. 2 The effect of scale that signal-to-noise is calculated on. **a**–**c** Mean–variance relationships for different scales of the original all-zero-in-one-condition data. **d**–**f** Corresponding ROC curves for the ENCODE dataset (GM12892 cells to H1-hESC), using S/N to set the true labels. Here, the signal-to-noise (S/N) is calculated from (trimmed mean of M-values-normalized) counts-per-million and used for all methods. Linear is equivalent to Rapaport's method, where S/N is calculated on the counts-per-millions. Log represents S/N calculated on log-transformed counts-per-million. vsn represents S/N calculated on variance-stabilized data [14]. ROC curves employ the same labels across all methods: the top 40 % of S/N are used as true DE genes whereas the lowest 20 % are false. Each method's *P* value is used for ranking the genes. *FPR* false positive rate, *TPR* true positive rate

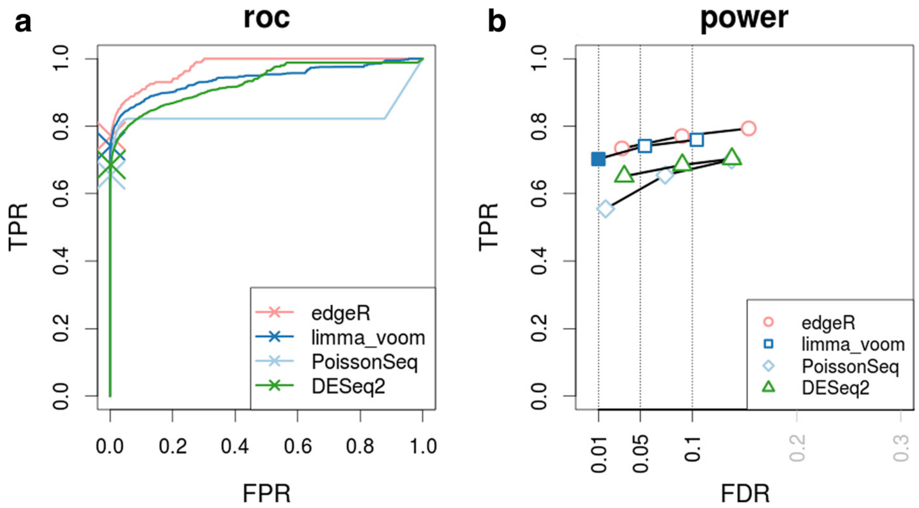

Fig. 3 Performance for zero-in-one-condition simulation. **a** ROC curves and **b** true positive rate (TPR) versus 'achieved' FDR curves of DE methods for the simulation dataset with zero-counts introduced as the true DE genes (overall performance of three simulations). The achieved FDR is the actual rate of false discoveries at the corresponding cutoff and this rate should ideally be controlled at the desired level. For the ROC curves, the *cross* on each curve represents the method's TPR at the (estimated) 5 % FDR cutoff. For the TPR versus achieved FDR curves, *points* are plotted at the following cutoffs: 0.01, 0.05 and 0.1. *Filled-in points* represent that the method has correctly controlled the error rate at the cutoff. *FDR* false discovery rate, *FPR* false positive rate, *roc* receiver operating characteristic, *TPR* true positive rate

Discussion

As developers and users of bioinformatics strategies, we are particularly interested in the metrics and methods that differentiate performance between the available tools. In this paper, we claim that count-based methods perform well when genes are only expressed in one condition, in contrast to an earlier report. We showed that a code error and the chosen scale of S/N resulted in the earlier conclusion that count-based methods suffer performance in this situation. By calculating the S/N on a different scale and using the same set of labels across methods, a reversal of method performance was observed. This highlights a sensitivity to decisions made in constructing the benchmark.

Using a customized simulation that introduces zero-counts in one experimental condition, we demonstrated that the performance of the count-based method is actually on a par with or better than other methods. We also debunked the postulation that poor performance is related to dispersion estimation in count models.

In the process of seeking the origins of this statistical performance difference, we discovered another potentially interesting phenomenon that may affect the interpretation of results. Looking at Table 1 and Additional file 1, it is evident that the scale of *P* values is drastically different between edgeR and voom. Although this observation appears rather unrelated to the ability to separate true from false DE genes, it is an indication that the scale of observations modeled affects the magnitude of statistical evidence derived. Not surprisingly, method performance is ultimately dependent on the scales, parameters and datasets used for the evaluation.

Software

R code and data that can be used to reproduce the figures in the main manuscript and in the supplement are available online [13].

Additional files

Additional file 1: Table of statistics for zero-count genes. Table of zero-counts, differential statistics and S/N for the ENCODE dataset. (CSV 112 kb)

Additional file 2: Supplementary figures. This file contains the mentioned supplementary figures. (PDF 712 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Both MDR and XZ designed and conducted analyses and wrote the manuscript. Both authors read and approved the final manuscript.

Acknowledgments

XZ is supported by an Swiss National Science Foundation project grant (143883). MDR acknowledges funding from the European Commission through the 7th Framework Collaborative Project RADIANT (grant agreement number 305626). We acknowledge data processing help from Malgorzata Nowicka at the outset of the project and to her and Charity Law for helpful feedback on an earlier version of the manuscript.

Received: 11 December 2014 Accepted: 15 September 2015
Published online: 08 October 2015

References

1. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
2. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):106. doi:10.1186/gb-2010-11-10-r106.
3. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12):550.
4. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):29. doi:10.1186/gb-2014-15-2-r29.
5. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics. 2012;13(3):523–38.
6. Smyth GK. Limma: linear models for microarray data. Chap. 23. In: Bioinformatics and computational biology solutions using R and Bioconductor. New York: Springer; 2005. p. 397–420. doi:10.1007/0-387-29362-0_23.
7. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013;14(9):95.
8. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500(7463):477–81. doi:10.1038/nature12433.
9. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14(1):91. doi:10.1186/1471-2105-14-91.
10. Soneson C. compcodeR – an R package for benchmarking differential expression methods for RNA-seq data. Bioinformatics. 2014;30(17):2517–18.
11. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res. 2014. doi:10.1093/nar/gku310.
12. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010;464(7289):773–7.
13. Additional material. http://imlspenticton.uzh.ch/robinson_lab/zero_count/.
14. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 2002;18 Suppl 1:96–104. doi:10.1093/bioinformatics/18.suppl_1.S96.