Authors' reply letter to the Academic Editor

From Jaume Bacardit
Academic Editor for PeerJ:

Thank you for your submission to PeerJ. I am writing to inform you that in my opinion as the Academic Editor for your article, your manuscript "Getting the most out of RNA-seq data analysis" (#2015:06:5530:0:0:REVIEW) requires a number of major revisions before we could accept it for publication.

The comments supplied by the reviewers on this revision are pasted below. My comments are as follows:

Editor's comments

Please address carefully all issues raised by the reviewers, especially about the definition of your 'gold standard'. I would strongly suggest to follow reviewer 2's advice and use a dataset for which qPCR validation data exists

Please be aware that we consider these revisions to be major, and your revised manuscript will probably have to be re-reviewed.

If you are willing to undertake these changes, please submit your revised manuscript (with any rebuttal information*) to the journal within 60 days.

Authors' reply: We thank you and the reviewers for the careful and very constructive responses. We have implemented most of the suggestions from the reviewers, particularly, on the important issue of justifying the construction of the reference differentially expressed gene (DEG) set using voom. This led to additional analyses involving a very recently published RNA-seq data set from Rajkumar et al. (published 25 July 2015; first PeerJ decision on 28 July, 2015), which contained sufficient numbers of qPCR-validated results (115 genes) for us to justify voom's usage.

For ease of reference, each reviewer's comment is given a prefix R1,R2,or R3 (Reviewers 1,2,3 respectively), followed by a number. Thus, R1.1 means comment 1 from Reviewer 1. Our replies are always in italics.

In the "Tracked Changes" version of the resubmitted work, we have color-coded paragraphs / statements (gray for Reviewer 1 – Jose M.G. Izarzugaza; green for Reviewer 2 – anonymous; yellow for Reviewer 3 – Hao Zheng) to indicate to whom the addition/changes answered to. Sticky notes are associated with the highlighted paragraphs/statements to provide some details to which specific comment they addressed to. New additions / changes from us are indicated in red. We have included an acknowledgement section to reflect the contributions of the reviewers in improving the presentation and coherence of the present work.

Yours sincerely,

T.F.Khang & C.Y.Lau

24 September 2015

Authors' reply letter to the individual reviewer's comments

Reviewer 1 (Jose MG Izarzugaza)

Basic reporting

The submitted manuscript adheres to current standards in the field and I do not see any main reason that should prohibit the review of this article. The text is written clearly in proper English.

The quantitative analysis of the relationship between biological effect, number of duplicates and DEG calling performance is sound and of interest to the scientific community.

Experimental design

R1.1. My main concern regards the validity of the gold standard for comparison. Currently, the authors use a gold standard that is based on predictions where "at least 70% are expected to be real". The presented results, then, do not evaluate how much the predictors perform on predicting DEG. Contrarily, they assess the degree of correlation between their predictions and voom's. The discussion of this aspect is very vague in the manuscript. The author should provide a solid argumentation of the validity of the voom gold-standard. If experimental validation is included, it might turn out that voom itself is the methodology of choice for the prediction of DEG.

Authors' reply: *To justify the use of voom for setting the gold standard (changed to "reference") DEG set, we performed additional analysis using the RNA-seq data set from Rajkumar et al. (2015; Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. BMC Genomics, 16:548.) which contained 115 genes qPCR-validated results. Details of the comparison procedure and the results for this analysis are given in the new sections: "Method for constructing reliable reference DEG set", "Characteristics of constructed reference DEG sets", and "Validity of voom for setting reference DEG set". To summarize, for the Rajkumar data set, we showed that only voom and edgeR called DEG sets with sizes that were not too small or too large, and voom was preferred on account of its relatively higher expected PPV compared to edgeR (88.9% vs. 72.6%) and also less variable DEG set size (±43 vs. ±694).*

Validity of the findings

R1.2. Following the previous comment, the authors should justify the validity of this assumption "Assuming that at most half of them were false positives, at least 70% of the voom-called DEG were expected to be real." Otherwise, they might consider providing the results as a range that considers (all correct, all false positives).

Authors' reply: *In light of the results from the analysis of the Rajkumar data set supporting voom as a reasonable method for setting the reference DEG, we have removed the statements:*

"The remainder of the DEG that were unique to RNA-seq may either be false positives, or DEG that could not be detected using microarrays. Assuming that at most half of them were false positives, at least 70% of the voom-called DEG were expected to be real."

as they potentially confuse readers.

R1.3. Figure 2 is difficult to follow. I would suggest that the information regarding the number of replicates and the calling method is incorporated in the scatterplots as a legend. The number of DEGs predicted might be encoded as a color whereas the predictor could be a shape, the diameter is difficult to interpret. Also, it might be benefitial for the reader to include n=10 currently provided as Supplementary. I hope this would improve the interpretation of the results.

Authors' reply: We feel that Figure 2 is reasonably clear. It is an example of a bubble plot, popularized by Hans Rosling in his Gapminder presentation. Here, the diameter of the bubbles represents one variable of interest (DEG set size in our case), and the color of the bubbles indicates a specific DEG method. Using shape of the plotting symbol to represent the latter would be rather confusing in our opinion, because of the overlapping of (seven) different symbols. Also, if color were used to represent DEG set size, one would have to use a continuous color tone, filling up an entire plotting symbol. This makes the plot quite complex to read at regions of substantial overlap between data points. For these reasons, we decided to keep Figure 2 as it is (with updates from current revision). We think that retaining the result for n=10 in as Supplemental Material Fig. S3 is fine, as the result does not differ much from the case of n=6 and this also keeps the number of figures in the manuscript low.

Comments for the author

R1.4. What I would consider minor comments:
- Closing parenthesis in "Supplemental Material Table 1). The remainder …"

Authors' reply: *Corrected.*

R1.5.- To facilitate the interpretation of the results I would refer to the datasets as "mild response" and "strong response" instead of the study names as these terms relate directly to the conclusions drawn.

Authors' reply: *We decided to keep the original reference to the two data sets using the name of the studies' first author, since there are extensive references to these data sets throughout the text and using a special name for these data sets is a more effective signposting strategy. We think it is not difficult to remember the data set name – effect size association as there are clear signposting in the text (e.g.in the "Characteristics of constructed reference DEG sets" section: "For the Bottomly data set (strong effect size) …; for the Cheung data set (mild effect size)").*

R1.6.- In the abstract, the authors should specify that PPV refers to positive predictive value before using the acronym.

Authors' reply: *We have added PPV in parentheses in the sentence where the term was first used in the Results section of the Abstract: "the differentially expressed genes should be called using methods with good positive predictive value (PPV)".*

Reviewer 2 (Anonymous)

Basic reporting

R2.1. Introduction, 2nd paragraph, 3rd line: “thorugh” should be changed to “through”.

Authors’ reply: *Corrected.*

R2.2. - When considering replicates, Z-test is also applied. How are the proportions for a given gene and phenotype class computed? As a mean of proportions? Please, be more specific on this.

Authors’ reply: *We have added more technical details in the revision (see section “Statistical methods for calling differentially expressed genes”) to clarify this issue.*

R2.3.- Criteria for differential expression, 1st paragraph: I understand that DE calls are selected according to p-value (adjusted p-value I guess) and fold-change but it is not clear for me the reason to add the explanation on x and y (from “The product of y>2” until the end of the paragraph, and therefore the definition of x and y). I think that it is enough if the threshold for p-value and fold-change are indicated.

Authors’ reply: *When only the p-value and fold-change thresholds are given, their combination results in a double-filtering criterion, whereby genes falling in a rectangular region defined by the two cut-offs are selected. It has been shown that regions defined by a hyperbolic curve (i.e. $y > 2/|x|$) give better DEG candidates in terms of p-value and fold-change trade-off (Xiao et al., 2014).*

R2.4.- Criteria for differential expression, last paragraph: The first sentence is incomplete (“a fold change of” ??). What is then the GFOLD value?

Authors’ reply: This was a mistake in editing and has been corrected. The correct sentence should be:

“For GFOLD, we used the default significant cut-off of 0.01. ~~for fold change of~~”

R2.5- I missed a short description of the differential expression methods to be compared. Most readers may know about edgeR or DESeq and DESeq2, but perhaps they do not know the procedure the rest of methods use to compute differential expression. A brief explanation would be enough so they do not have to go to the original papers.

Authors’ reply: *This is a fine suggestion and we have provided a new table (Table 2) that summarizes the modelling strategy behind each of the DEG call methods considered.*

R2.6. - Benchmarking, 4th paragraph: "The exception is DESeq2, which specifically requires raw counts instead of normalized count for analysis". Is this only true for DESeq2? As far as I know, edgeR also requires raw counts, except the authors mean that edgeR accepts other normalization or scaling factors.

Authors' reply: DESeq2 insists on raw counts as input data since its algorithm does not work well when normalized data is given as input. We have rephrased the statement for clarity.

R2.7 - Transcriptome coverage effect: Why are the authors discussing about similarities of transcriptomes between human and rat? The Bottomly data is from mouse, not rat. Furthermore, the number of transcripts in human is approximately twice the number of mouse transcripts. Could you elaborate more on this?

Authors' reply: *We have corrected the organism from rat to mouse for the Bottomly data description. Transcriptome coverage affects the statistical power for detecting differentially expressed genes. Conceptually, given two similar studies with the same sample size but different transcriptome coverage, we can expect differential expression call methods to make more calls in the study with higher transcriptome coverage, rather than the one with lower coverage.*

The true count of the transcripts of a gene is better approximated by the observed count when transcriptome coverage is high, thus all differential expression call methods experience increased statistical power to detect DEG in such a situation, compared to low coverage situation (Tarazona et al., 2011). We think the assumption that transcriptome size is approximately equal between human and mouse is reasonable, since estimates of number of genes are approximately the same in both species (~25,000) and about 99% of mouse genes are homologous with human genes (Guénet,J.L. 2005.The mouse genome. Genome Research, 15:1729-1740.). Once the transcriptome size issue is settled, transcriptome coverage becomes easier to interpret based on the reads per sample for both the mouse and human data set. We wanted to accentuate the fact that although the transcriptome coverage in the human data (mild biological effect size) was about three times larger than that of the mouse data (strong biological effect size), it did not result in more DEG detected due to its weak effect size (sex differences in immune-related B-cells). This highlights the relative importance of biological effect size in affecting the number of detected DEG.

R2.8. - In general, I think that the authors should state more clearly that to perform a real statistical inference analysis, biological replicates are needed. Otherwise the results obtained cannot be generalized to the population and validation is essential (as they already mention). They could also include references that sustain the need of triplicates as many studies are recently claiming.

Authors' reply: *We think the original paragraph in the "Conclusion" section already makes these points sufficiently clear. The work of Liu et al. (2014) highlighting the need for biological replicate is now cited: "A standard triplicate design (Liu et al., 2014) …"*

Experimental design

R2.9. The manuscript provides a comparison of six differential expression methods for RNA-seq data on two publicly available datasets. The authors use two performance parameters: sensitivity and positive predictive value. To compute them, they define a gold-standard set of DEGs by applying voom transformation and limma method. In my opinion, it is arguable that the set of DEGs considered as a gold-standard can be defined from the results of a particular DE method. All methods are subject to declare false positives or negatives, and using this as a gold standard may produce misleading conclusions. The results of the study would have more impact if at least an additional RNA-seq dataset with qPCR validation were included to corroborate at least some of these conclusions.

Authors' reply: *Setting a reference set is a challenge in differential expression analysis if one does not adopt a simulation-based approach, such as in our case. There is no easy way out of this problem, since even in simulation-based approaches, the validity of the reference set also depends on whether assumptions of the simulation mimic reality or not, which is basically uncheckable. Performance evaluation of methods is therefore ultimately empirical.*

In our revision (see "Method for constructing reliable reference DEG set"), we used a very recently published RNA-seq data set with 115 qPCR validation results (Rajkumar et al. 2015. Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. BMC Genomics, 16:548.). In this analysis (Table 3), we showed that only voom and edgeR made reasonable number of calls for DEG (in hundreds), whereas others made either too few calls (NOISeq, DESeq2, GFOLD) or too many (Z-test, DESeq). Subsequently, we showed that the expected PPV of voom was relatively higher (88.9%) against edgeR (72.6%); also voom produced DEG set sizes that were less variable as well (±43 vs. ±694). This result suggests that using voom to set reference DEG set was reasonable.

R2.10. I also have some concerns about the choice of the dataset with strong biological effect. Providing the set of truly differentially expressed genes is correct, there are only 362 DEGs (less than 3% if we consider only the genes with counts). According to my experience and to the literature, I do not think this can be considered as a "strong effect". I agree that simulation experiments may have some limitations but at least they can control better the size of the effects, which by the way, it is not related only with the number of DEGs but also to the magnitude of change between phenotype classes. Could the authors provide more information about this magnitude? And a discussion on my comment?

Authors' reply: *We initial described the Bottomly and Cheung data sets as having "strong" and "mild" biological effect sizes respectively because the between phenotype variation of read count was much stronger than within phenotype variation in the former, and generally weaker in the latter, as shown in Fig. 1. However, we agree that this may not be clear to readers and the presentation would benefit from a more quantitative approach. We also agree that DEG set size may not correlate well with biological effect size. To this end, we presented new results (evaluating the T^2 statistic for DEG; see section "Characteristics of constructed reference DEG sets").*

R2.11. The authors are right when considering the importance of the biological effect size when calling DEGs. However, other factor, which is also important in any statistical analysis such as DE, is the biological variability within the same experimental group (phenotype class) when biological replicates are available. Could the authors discuss in the manuscript if the biological variability within groups is the same for both datasets, groups of replicates compared, etc. and if it is affecting some how the results of any of the compared methods. It could be useful for the readers because it is difficult to estimate the size effect in a given experiment, but variability can easily computed.

Authors' reply: *We thank the reviewer for suggesting the checking of within phenotype variation. For each data set, we performed an agreement analysis (see section "Characteristics of constructed reference DEG sets") for the variation (measured as standard deviation) of the two phenotype classes in all genes (Supplemental Material Fig. S2 for the associated Bland-Altman plot). We then checked the proportion of DEG that were within 2SD and 3SD from perfect agreement, as this would give some indication of whether a call was likely due to the presence of extreme values inflating between phenotype variation. The results indicated that the majority of DEG called had within phenotype variance within 3SD from perfect agreement.*

Validity of the findings

R2.12. The novelty of this comparison study is that it focuses in the case of non-replicated data, which is still quite common in RNA-seq. As mentioned before, the validity of the results and conclusions from the DE methods comparison may depend on the chosen data sets. It is difficult to generalize the conclusions with only two examples so the authors should carefully analyze the impact of other factors such as the magnitude of change between classes or the variability within classes. They should also provide another example in which a more widely accepted gold standard set of DEGs such as qPCR was available to reinforce their conclusions.

Authors' reply: *We have implemented additional analyses as described in the replies to R2.9, R2.10 and R2.11.*

Reviewer 3 (Hao Zheng)

Experimental design

R3.1. As a method comparison study, the design of this work is relatively straightforward. While the rationales of choosing the data sets are explicitly explained, the reasons of including such seven methods are not interpreted in detail. In the section of Introduction, and Materials and Methods, only the applicability of NIOSeq, GFOLD, and ASC with respect to biological replication (whether an analysis using one sample per phenotype can be conducted) was provided. Whether these methods in comparison are state of the art for RNA-seq analysis has not been discussed.

Authors' reply: *We appreciate this comment and which led us to compile Table 1 listing down methods for differential analysis in the literature, together with some metric (citations per year) indicating their frequency of usage in the scientific community. In the paragraphs following Table 1, we also provided the rationale for selecting the methods considered in the present study.*

R3.2. In the simulated unreplicated experiments, 27 and 124 pairs of samples were chosen for two data sets, respectively, because the ASC package encountered problems in dealing with the remaining pairs. This strategy is not preferred, for this manner of subset selection may introduce bias in comparing performance. It is better to run all combinations (or also randomly select 100 instances, like for n=3, 6, and 10) for seven methods, and to point out the suitability of individual methods.

Authors' reply: *We agree that using only the pairs that worked for ASC may potentially bias results for the other differential expression calling methods. In the revision, the methods analysed all 110 pairs for the Bottomly data set (11x10=110) and 300 pairs for the Cheung data set. We then highlighted the fact that ASC only managed to run in 27/110 and 124/300 pairs. This resulted in slight changes to sensitivity, PPV and DEG set size which did not affect the initial conclusions. As a result, all performance metrics for the case of n=1 in both data sets have been updated in Supplemental Table T3 (with corresponding updates in Figure 2a, 2d; Figure 3a, 3d)*

Validity of the findings

R3.3. The 'gold standard' used in this work for confirming DEGs is the output from implementing voom algorithm. Although the voom algorithm transforms the count data into a microarray-like data type so that the mature limma analysis could be applied, it is not convincing that the results based on voom are closer to the real situations. In fact, if voom by default outperforms the current RNA-seq analysis methods (such as the seven ones in comparison), this work of comparing these methods is not meaningful. Therefore, the statement of 'golden standard' might need rephrase or further justification, and the following implicated presentations need corresponding adjustments as well.

Authors' reply: *We have included a new section ("Limitations") to highlight the fact that when an empirical rather than simulation approach is taken to assess method performance, unless large numbers of qPCR-validated results are available to fix the reference DEG set in*

the test data (which are not available), invariably one would be forced to set the reference DEG set using a particular DEG call method. Of course, the issue then is whether the reference DEG set thus called is believable or not. We have performed additional analyses using a recently published RNA-seq data set (*Rajkumar et al. (2015). Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. BMC Genomics, 16:548.) which contained 115 genes qPCR-validated results. In this analysis, we found that only the DEG set size called using voom and edgeR were of reasonable magnitude, while that of others were either too small or too large. Between these two methods, voom showed expected PPV of about 88.9% compared to edgeR's 72.6%, and also less variable DEG set sizes (±43 vs. ±694), suggesting that using voom to set the reference DEG set for analysis of the Bottomly and Cheung data sets was reasonable.*

Since it is possible that voom's result are the best for replicated designs, we suggested taking union of either DESeq2 or edgeR's results with those of voom's; if this union set is too large, then more conservative intersection set could be considered (Second paragraph in "Limitations" section).

R3.4. The comparisons among the involved methods are presented mainly through Figures 2 and 3, and their corresponding interpretations in the section of Results and Discussion. While the numbers on the top of Page 5 are systematically reported, it is difficult to link them to the counterparts in Figure 2, especially when the figure has six panels and each panel covers seven colors. It would help if the expression of xx +/- xx % is followed by (Panel x, Figure 2). Based on the current setting, the numbers of circles in Panel (a) and (d) of Figure 2 are different from the remaining four, so listing the numbers of circles could avoid potential confusion. Because the size of identified DEGs is shown via the diameter of the corresponding circle, an illustrative legend could better convey the information.

Authors' reply: The expressions xx +/- xx % now have reference to the relevant panel in Figure 2. We have added details about the number of circles (sample pairs) for each method in Figure 2a and 2d in the caption. A new legend showing circles of certain sizes with their corresponding set size has been added into Fig. 2d.

R3.5. Meanwhile, the wording "... the mean size of DEG called was very small (6+/-11.6%) " (second paragraph on Page 5) is rather confusing. The size of DEGs should be an integer, rather than some percentage.

Authors' reply: *The percentage was a typo. This value has been updated and moved to the section "DEG set size" after the statement: "…with extremely low mean DEG set size (2.5 ± 6.8)* .

R3.6. Due to the big gap of sizes of selected DEGs, Figure 3 has to use log scale to show a normal box plot. However, it is not straightforward to map the numbers back in originally linear scale. It should largely contribute to the presentation if there is a table showing the summary statistics for the DEG sizes in the linear scale.

Authors' reply: The DEG sizes in linear scale were included in Supplemental Table T3 in the initial submission. However, we did not signpost it adequately in the DEG set size subsection. This has been fixed.

Comments for the author

R3.7. The manuscript context says there are 24 males and 17 females in Cheung data set on Page 3, but the caption of Figure 1 says there are 17 males and 24 females.

Authors' reply: *The correct sample sizes are 24 males and 17 females. Correction has been made to the caption of Figure 1.*

R3.8. In the section of Introduction on Page 2, 'through' is misspelled as 'thorugh', and the past tense of 'offer' should be 'offered' instead of 'offerred'.

Authors' reply: *Corrected.*

R3.9. In the section of Conclusion on Page 6, 'Moreover' is misspelled as 'Morever'.

Authors' reply: *Corrected.*