Dear Jeff, Dimitris,

Thank you for the helpful comments. We have made major improvements to our manuscript to address your and the reviewers' concerns. In particular, we have completely redesigned the simulations to make them more representative of real single-cell RNA-seq data. We have assessed performance of the various methods under a greater variety of simulation settings involving variable library sizes, decreased plate effects and zero inflation for the counts. We have also expanded our analysis of the real data – we now include a label-swapping analysis to assess type I error in a real data set, and we show an improvement in the biological relevance of the results (based on DE of known pluripotency factors) when summation is performed. We believe that these modifications provide a more rigorous justification of our proposed methodology.

We have also streamlined the manuscript considerably, reducing its length by almost 5 pages by removing the quantile adjustment section (given that it provides no benefit over direct summation) and by trimming other sections. We hope that this will improve the readability and focus of the manuscript.

Yours sincerely,

Aaron Lun and John Marioni

**REVIEWER 1:**

**1) “To avoid detection of spurious DE, a summation approach is proposed whereby counts for all cells on each plate were summed prior to the DE analysis.” Then variability of cells within the same plate/group cannot be modelled, which is the very advantage that single-cell sequencing technology provides.**

In our paper, we focus on DE analyses which are very widely applied in the context of scRNA-seq studies. Here, the goal is to find genes that show a difference in mean expression across cells. Consequently, the variability across observations can be treated as a nuisance parameter in this context. As long as the variance is modelled accurately across observations – be they counts for single cells, or count sums for plates – then the DE analysis will be successful.

**2) In short, the authors propose summation to solve the confounding plate effect seen in single cell**

**sequencing. The summation can be either direct or quantile-adjusted summation. This is the authors’**

**major discovery, the quality and quantity of which does not qualify for publication in a professional**

**statistical journal. Besides, the authors use existing software for the actual calling of differential genes, which can be considered just a repetition of previous researchers’ work.**

We thank the reviewer for their comments on the novelty of our manuscript. However, we believe that the proper application of existing methods is just as important as the development of new methods. It is clear that, for single-cell RNA-seq data sets containing plate effects, popular DE analysis methods are being misapplied. This is the point of our work – to rigorously demonstrate that this misapplication has negative effects on the DE analysis, and that it can be resolved by performing the analysis on count sums. To our knowledge, this has not been investigated before, and so we feel that it warrants publication.

**3) In realistic settings, the library sizes of cells in any sequencing-type of experiment cannot be assumed to be equal across conditions (groups) or even replicates. Therefore, all the simulation work before section 4.2 will be invalid, in which equal library size is assumed.**

The purpose of the simulation with equal library sizes was to provide the simplest, easiest scenario in which all methods should perform well. That the methods still fail to control the type I error rate indicates that there is a fundamental statistical problem, rather than any problem with normalization. Nonetheless, we have redesigned the simulations to be more realistic such that a distribution of library sizes is present. (See Supplementary Figure S1a for the distribution of size factors across cells, as estimated from real data. Variability in the size factors between cells leads to variability in the sizes of the corresponding libraries ).

**4) “In all cases, library sizes were only modified for a fraction of cells and in a subset of plates. This**

**subtlety is necessary because the mean-variance relationship will be the same between plates if there is a consistent increase in the library size for all cells on a plate, or if there is an increase in the library size for the same proportion of cells on all plates.” The “same proportion” restriction of this simulation work is unrealistic. This simulation cannot be considered valid.**

These simulations were intentionally unrealistic, as they were designed to represent pathological scenarios where summation would be expected to perform poorly due to changes in the mean-variance relationship between observations. That summation still maintains type I error control indicates that it can still perform well in challenging situations. Nonetheless, we have removed these simulations in favour of a more realistic spread of library sizes and numbers of cells per plate (see Supplementary Figures S2, S3).

**5) The authors seem to only have validated their method on one simulation (Fig. 3) in which genuine DE is assumed. This simulation assumed equal library size, and is therefore not valid according to my point 3. Therefore, throughout the paper, there is not a single simulation in which the authors simulated real differences and calculated percentages of false negatives to compare and control type II errors of the summation approach. “Recall that the single-cell analyses fail to maintain type I error control in the simulations. This suggests that the increased numbers in Table 1 correspond to detection of false positives rather than genuine DE genes”. Then for the reasons stated above, the authors cannot claim the decrease in called DE genes by the summation approach is due to well-controlled type I error, rather than inflated type II errors.**

We have redesigned the simulations based on parameter estimates from real data (including the observed distribution of library sizes), and we observe similar results. We have also repeated the real data analysis after swapping sample labels between one set of replicates for the groups being compared (see Figure 4). Under the assumption that no genes should be DE in the swapped analysis, we show that the DE methods applied to single cells fail to control the type I error rate, while the same methods applied to the summed data perform well. This indicates that differences in error rate control are still present between summed and single-cell analyses of the real data, consistent with our simulations.

Furthermore, the liberalness observed in the label-swapping analysis supports our claim that the increased number of detected genes in Table 1 for methods applied to single-cell counts is caused by increased detection of false positives, rather than an increase in detection of genuine DE genes.

**6) I find the writing of this manuscript to be too lengthy and lack of focus**

We have addressed this by trimming the manuscript considerably. In particular, we have incorporated the GLMM analysis into the main simulations, and we have removed the description of the quantile adjustment method as it provides no advantage over direct summation.

**REVIEWER 2:**

**1. In section 2.2, \sigma^2 is set to be either 0.5 (strong plate effect) or 0 (no plate effect). It would be helpful to try different values of \sigma^2 (e.g. 0, 0.1, 0.2, 0.3, 0.4, 0.5) and replace the current Figure 1 with a line plot showing the trend of type I error rate with increasing plate specific effect. Also is there a possible way to somehow estimate the \sigma^2 in some real data so readers could have a rough idea how serious the plate effect could be in real datasets?**

We have redesigned the simulations to use a sigma^2 estimate from the real mESC data set (estimated by fitting GLMMs to all genes and taking the variance of the random effect). We estimate this to be about 0.3, which is comparable to the value used in the original simulations. We have also repeated the simulations after halving sigma^2 (see Supplementary Figure S2), to test whether the results are robust to changes in the magnitude of the plate effect. We show that all methods are substantially liberal, even in the presence of a smaller plate effect.

**2. In section 2.2 (and section 4.2), the counts are generated using a negative binomial distribution. However, due to the drop out events and low sequencing depth per cell in scRNA-seq data, it would be more reasonable to generate the counts using a zero inflated negative binomial distribution. i.e. the counts has a probability p to be exactly zero and a probability 1-p to be generated from a negative binomial distribution.**

We have repeated the simulations after sampling counts from a ZINB distribution, with parameters for the count and zero components estimated from real data. We show that similar results are obtained as those with the standard simulation (see Supplementary Figure S2).

**3. More generally, evaluation of plate effects and the methods that adjust for such effects should carried out in simulation settings similar to real single-cell RNA-seq data. All current simulations are more similar to bulk RNA-seq data, and they do not reflect data characteristics typically seen in single-cell RNA-seq (e.g., drop-out, half of genes may have zero count within each sample, etc.)**

The simulations have been redesigned using parameter estimates derived from the mESC data set. This includes estimates of the (conditional) NB dispersion and mean for each gene; the distribution of library sizes across cells; the variance of the plate effect; and for the ZINB simulations, the drop-out rate for each gene. See Supplementary Figure S1 for plots of the parameter estimates used in the standard simulations. We find that our results are unchanged, and that all methods are still liberal in these improved simulations.

**4. The real data example (section 5) is not convincing enough. The authors merely compared the number of genes detected by each method. However, whether the new method (DE based on summation) really matters in term of true biology or downstream applications is not clear. For example, some people may only be interested in a few top genes. For them, a biased type I error estimate may not be a very serious problem. Can you provide a real example showing that the new method really makes a difference in terms of biological discoveries? Some potential criteria the author could consider include but are not limited to: (a) to compare the DE genes obtained from scRNA-seq (original and after summation) with DE genes obtained from bulk RNA-seq (bulk RNA-seq is available for the mESC dataset described in the manuscript) and treat the DE genes from bulk RNA-seq as gold standard genes. (b) Check whether the identified scRNA-seq DE genes (original and after summation) are related to the true biology. For example the author can perform GO analysis just as described in the mESC paper and should be able to reproduce the GO analysis results in the mESC paper to some extent using the new summation method. Also, does the new method result in more meaningful GO terms? The DE state for key transcription factors should also be reproduced using the new summation method (from the original mESC paper, main article: “Importantly, Oct4, Sox2, and Klf4 are not differentially expressed between 2i and a2i. Additionally, while Nanog is significantly differentially expressed, the expression level difference between 2i and a2i is smaller than between 2i and serum”)**

We have compared the ranking of key pluripotency factors (listed in Supplementary Figure S5 of the mESC paper) between the single-cell and summed DE analyses of the mESC data set. We find that the rankings for these factors are consistently improved – that is, the factors are closer to the top of the DE list – in the summed analysis compared to the single-cell analysis (Supplementary Figure S6, p-value of 0.0002). This suggests that summation can provide more biologically relevant discoveries in analyses of real data, even when only the top set of genes are of interest.

Additionally, even if the rankings were unchanged, accurate control of the error rate with the summation approach would still be advantageous. For example, the FDR threshold provides a principled way to decide whether or not a putative DE gene warrants follow-up study, especially if it lies outside the top set of DE genes. This depends on proper control of the FDR below the nominal threshold, which is not possible with the analyses based on single-cell counts in the presence of plate effects.

**5. In some situations the summation will not work. For example, suppose there are only two plates: plate 1 only contains cells from treatment A and plate 2 only contains cells from treatment B, and differential analysis is to be performed between treatment A and B. After summation, there will be only one observation for A and one observation for B. In this case it is not possible to perform differential tests (DEseq2, edgeR, etc.) without making additional assumptions. Will this be a limitation of the summation method?**

This is an interesting question. As the reviewer suggests, the summation approach will not be easily applied for a design with one plate per group, as there is no replication within each group. In such cases, performing the DE analysis on single-cell counts is more straightforward as the variance can be modelled across cells. However, the latter analysis is only valid if one assumes that there is no plate effect. This is because the variance estimate only accounts for the variance within each plate, not that between plates – no protection is provided against spurious DE between groups due to plate effects. The only way to verify this assumption would be to generate data from replicate plates and populations, in which case summation becomes applicable. We have added a discussion about this to the manuscript.

**6. In section 2.2, the statements “to represent the decreasing mean-dispersion trend observed in real data.” and “consistent with high technical and biological variability in scRNA-seq counts.” need to be justified. The author would need to either give appropriate reference to existing literature or provide some real data examples.**

These statements are no longer necessary in our new manuscript, as the simulations have been redesigned to use dispersion estimates from real data. Incidentally, we still see large dispersion estimates (> 1 for the majority of genes) and a decreasing mean-dispersion trend in Supplementary Figure S1.

**7. The monocle approaches is also involved when evaluating different DE methods. However, monocle performs a different kind of DE analysis i.e. it tests whether the expression of a gene is constant or not across a pseudo temporal path instead of testing the differences of gene expression averages between two or multiple predefined sample groups. The author may need to add some discussion that monocle is different from other canonical methods like DEseq2, edgeR and voom.**

We have added a comment about the intended design of Monocle for testing DE across a pseudo-temporal ordering. Nonetheless, we believe that the statistical framework of Monocle is generalizable to our experimental designs. The differentialGeneTest documentation describes the purpose of the function as:

*Tests each gene for differential expression as a function of progress through a biological process, or according to other covariates as specified.*

This suggests that arbitrary covariates – in our case, biological groupings – can be used. This should not pose a problem for likelihood ratio tests, which are applicable to many experimental designs and contrasts. Indeed, we observe that Monocle performs well in the simulations where plate effects are absent.

**8. Figure legends should be added to Figure 4 showing the meaning of black circles and grey diamonds.**

All figures now have appropriate legends.