# Reply to reviewers

## General comments

We have made two changes that we wish to point out explicitly.

First, we have added a short section and an appendix providing details on the actual distribution of the F-statistic after two-way ANOVA based batch adjustment. We think this should help address some of the questions and detail the effect of this kind of batch adjustment more clearly.

The second change, although handled briefly in the manuscript, is somewhat critical with respect to the discussion of the experiment published by Towfic et al. (Experiment 1). We contacted the authors of this paper, both to notify them of the potential problem and to verify our re-analysis. We were then informed that the description of the data and analysis method is not entirely correct. The data set they cite in the article (GEO: GSE40566), which we analysed, contains 107 samples (in the raw data file from which 2 are later excluded). Two subsets of these samples are then compared. There was some disagreement in how these treatment subsets were annotated in GEO relative to the article, which we have now corrected in our re-analysis of the data.

However, the crucial issue is that their article actually performed the analyses on a different version of these data, which has now been made available (GEO: GSE61901). The main difference between the cited data set (GSE40566) and the one they performed their analyses on (GSE61901) is that the latter contains two technical replicates per sample: these correspond to the pairs of slots on the bead-arrays which are physically connected on the slide. As such, their data set contains twice as many “samples”. Their analysis of treatment effect was performed including these techical replicates as independent samples.

We have informed them that this is not an appropriate approach, although they do not seem to acknowledge this. However, we did not wish to bring this discussion into the manuscript, and so decided to keep the original analysis using the cited data set (GSE40566) in which these technical replicates have been combined into one. We have pointed out, following the discussion of this experiment, and in as factual a tone as we could, how the analysis by Towfic et al. differs from our re-analysis. On their request, we have included a reference to a similar analysis which was performed on the cited data set (GSE40566) by Bakshi et al. How the batch adjustment was done on this is not fully clear, but we suspect it must have used a two-way ANOVA approach as implemented in Partek, and thus would suffer the same problems as displayed in our re-analysis.

Ideally, we could have started over, looking for more examples in the published literature. However, the amount of effort it takes to actually reproduce published finding is rather discouraging. The Towfic et al. article was picked because it provided enough detail to allow us to verify without too much effort that it was in the “target group”: it was clear from the available data and methods description that it had an unbalanced design and used ComBat with treatment group as covariate. Often, even this basic information may be hard to find.

## Reviewer 1

(4) As the author of the sva package, I have now updated the vignette to encourage users not to include group differences when using ComBat in response to the results of this paper. This change has been pushed to Github (https://github.com/jtleek/sva-devel) should propagate to the development version of sva in a couple of days, and should propagate to the release version on the next release.

We are most pleased to see such a quick update to the sva package, and have added a note in the manuscript noting that this has already been done.

## Reviewer 2

1. Most importantly, the authors do an excellent job defining the problem and illustrating it with practical real-data examples. However, they make no attempt to actually solve the problem. For example, I agree that most people (incorrectly) use the distributional assumptions in 3.10 in their 2-step approaches. However, can the authors derive an expression for the actual distribution of the data, and then use this to provide and validate a more appropriate (and justified) two-step procedure. I believe that this is crucial for transforming this paper from an editorial that describes and presents a problem to an research paper that actually solves one.

The main problem, as the reviewer is correct in emphasising, is the two-step procedure itself when applied to for heavily group—batch unbalanced data sets. Unfortunately, the only fully appropriate solution we can see is either to avoid it by ensuring the design is fairly balanced, or to include batch as a covariate in all subsequent analyses.

We have pondered if there might be a two-step solution in which group and batch effects are estimated using a two-way ANOVA, and then batch effect is only partially removed: by just the right amount to ensure the remaining variability from the batch effect compensates for the missing variability in the group effect estimates. However, we decided this was not a reliable approach. It might work for one particular analysis, e.g. comparison of differences between two groups, but would be unreliable for other analyses.

Perhaps of more practical use, we discuss briefly how to assess the effect of batch-group unbalancedness in particular cases: e.g. the v\_0/v factor or adjustment factor to the F-statistic. For analyses that produce a statistic similar to the F-statistic, one might try to compensate for the problem by adjusting the statistic by a factor (like the factor v\_0/v). This is perhaps the most general solution we can offer, but its adequacy still needs to be evaluated in each separate case.

We have determined the actual distribution of the common one-way ANOVA F-statistic when applied to data which have first been batch adjusted using two-way ANOVA. In theory, this could of course be used to recompute P-values using the appropriate F-distribution, but this requires some effort and in effect is equivalent to doing the two-way ANOVA in the first place.

We have added a short section in the main manuscript summarising the effects on the F-statistic, and placed the derivation of the distribution in an appendix. We have tried to keep the derivation brief, but hope it is still sufficiently clear.

2. Alternatively, this problem can be considered in the context of degrees of freedom. In a one-step procedure (e.g. ANOVA) the residual degrees of freedom is reduced by every parameter that is included in the model. So if a model includes treatment and batch, the t-test (or F-test) for significance of treatment will be less than the case that excludes batch (as is done in the two-step). Therefore, could the ‘effective’ degrees of freedom equal to the one-step residual df be used in the two-step procedure? Would this correct the distributional issues in Figure 2?

It is true that the degrees of freedom in the F-distribution are affected by the two-way ANOVA batch adjustment, but this is actually not the main effect. The main effect is that the F-statistic gets enlargened by a factor which is independent of the sample size. We hope that the derivation of the actual distribution of the F-statistic (and a more convenient approximation in terms of the F-distribution) may help clearify this.

An R script performing these analyses is now included at GitHub which compares simulated data to both the F-distribution assumed by a one-way ANOVA and our approximated F-distribution. This also helps illustrate how the computations (of q-tilde and sigma-square-tilde) are done.

3. Furthermore, could the uncertainty ratio of v\_0/v defined by Jenson’s be used to achieve an ‘effective’ degrees of freedom?

See 1 and 2.

4. The authors claim that the problem of underestimating uncertainty is independent of size. However, while I expect that the problem will not go away as sample size gets larger, I will be surprised if the impact of the problem doesn’t reduce as the sample size gets larger. The authors need to better justify this result with more technical/theoretical descriptions and with data examples (can be simulated if needed here).

We have added a simulation with different sample sizes to illustrate that the problem does not decrease with increasing sample size. Actually, when using ComBat on small samples, it will tend to shrink the batch adjustments, and so when using ComBat the problem will often be less pronounced on small samples than on big ones.

5. Figure 1 very confusing and not clear. I have looked at this several times and for several minutes and I still don’t understand what the authors are trying to portray. The authors should reconsider how these data are presented and revamp this figure to make it clearer.

We have worked on this, both to make the example simpler (although perhaps a bit more designed), and to make the figurative presentation clearer. I hope we have succeeded.

6. In the Abstract, the authors state ”which mostly likely has contributed to false discoveries being presented in the literature” this is a strong statement and should be backed up by some concrete examples.

We have moderated this claim slightly since we, admittedly, do not have many examples where things have demonstratively gone wrong. Unfortunately, as we also point out, presentations of how batch adjustments have been performed are often so sparing that it is difficult to make an assessment.

7. The authors focus their concern on ComBat, but other methods that adjust for technical heterogeneity, such as batch mean adjustment (without covariates) or SVA may also have the same concerns if used in a similar two-step process. This should be more clearly described and discussed. The problem is not with ComBat in particular, rather with the fact that uncertainty can easily be underestimated in two-step approaches.

This is a very good point indeed, and extremely important to convey to the reader. We have tried to make it more clear that this is a general problem with two-step procedures for batch adjustment in cases where outcomes differ greatly between batches and batch effects thus might be a strong confounder.

8. I agree with the statement “…systematically underestimate the statistical uncertainties and exaggerate the confidence of group differences…” I believe that in the introduction the authors do an excellent job heuristically describing the problem with two-step approaches.

9. In many parts of the manuscript, the authors discuss the ambiguous idea of underestimating ‘uncertainty’, but the never precisely define what they mean by uncertainty (e.g. variance, degrees of freedom, etc). This should be more clearly defined throughout the manuscript.

We have now for the most part replaced ‘uncertainty’ with less ambiguous terms, mostly in terms of variance of estimators.

10. In Section 2.1, the authors describe the full ComBat model, but then immediately reduce their theoretical discussion to cases with only one gene and no differences in variance. As such, I think most of this subsection is largely irrelevant and confusing (i.e. defining a more general model and then never using it). It would be better to just define the models used in the paper, and then give heuristic descriptions of the differences between this an ComBat.

We agree that equation 2.2 presenting the model used by ComBat is not required for discussing the effects of batch adjustment using the two-way ANOVA model. However, for readers either familiar with ComBat or similar models, or who want to refer to the ComBat article, we think it is convenient to present what that model is and how it relates to the simpler model we analyse. It also has some relevance for discussing the similarities and differences between the pure two-way ANOVA method and ComBat.

11. Sections 2.2 and 3.2 are contain most of the technically important material, but their both contain multiple ideas/themes/results. These would be much easier to understand in the authors split these into smaller more direct subsections.

We have now split 2.2 and 3.2 into smaller subsections and hope this will make it more structured.

Minor concerns:

1. In section 3.3.1, the authors should include more detail/description of the actual experimental design for this experiment (how many batches, how many treatment/control samples are there in each batch, etc).

We have added some more detail to this.

2. In Section 3.3.1 it might be better (and more) to use p-value cutoff and not FDR…

In this kind of analysis, FDR has become the de facto standard. The main reason for this is that the number of tests (genes tested) is so high that the smallest P values tend to get quite small owing to the number of tests alone, while Bonferroni correction tends to kill off most findings.

However, we do agree that the number of genes with FDR below a given threshold is a bad statistic for assessing findings or comparing methods.

Still, we decided to follow the analyses provided in the publication by Towfic et al. as closely as we could. However, we do comment that the use of FDR makes the analyses more sensitive to false positives.

3. The “6+2, 3+4”, etc represenations are not clear and should be presented in a different way

This has been adjusted.