Cambridge, January 8th, 2016

Dear Biostatistics editors,

We wish to submit our manuscript, “Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data”, for consideration as a research paper in Biostatistics.

Single-cell RNA sequencing is a powerful tool that can provide unparalleled insights into the molecular biology of individual cells. Count data from this technique can be used to characterize novel subpopulations via clustering; to identify highly variable genes driving cellular heterogeneity; and, of course, to identify differentially expressed (DE) genes associated with phenotypic differences between cells. However, DE analyses are not easily applied with most existing study designs, where the logistics of the experimental protocol means that cells must be processed in batches, i.e., plate-by-plate. This means that the biological conditions of interest are confounded with plate effects, such that genuine DE between conditions cannot be easily distinguished from spurious plate-based differences.

In the absence of a suitable method to remove plate effects, most existing studies have simply ignored them when performing DE analyses on the single-cell counts. However, it is unclear whether such an approach is statistically valid. In this article, we demonstrate that failure to account for the plate effect with existing DE analysis methods results in loss of type I error control on simulated data. To avoid this, we propose a simple yet effective solution, in which counts for all cells on each plate are summed and the count sums are used for the DE analysis. We show that our summation strategy restores type I error control without compromising detection power for DE genes, and is robust to different numbers and sizes of cells on each plate. We also observe that the differences between single-cell and summed analyses are recapitulated in a real single-cell RNA sequencing data set, generated from mouse embryonic stem cells cultured under different serum conditions.

In summary, we believe that the method presented in this manuscript will improve the statistical rigour of DE analyses of single-cell RNA sequencing data. This will improve the quality of the biological conclusions that are drawn from such analyses. We anticipate that this will of great importance for future studies of gene expression at the single cell level.

We hope that our paper will be of interest to you and to Biostatistics. For reviewers, we suggest that Mark Robinson (University of Zurich, [mark.robinson@imls.uzh.ch](mailto:mark.robinson@imls.uzh.ch)), Rafael Irizarry (Harvard University, [rafa@jimmy.harvard.edu](mailto:rafa@jimmy.harvard.edu)) and Sandrine Dudoit (University of California, Berkeley, [sandrine@stat.berkeley.edu](mailto:sandrine@stat.berkeley.edu)) be considered.

Yours sincerely,



John Marioni