Cambridge, January 27th, 2016

Dear Genome Biology editors,

We wish to submit our manuscript, “Computing normalization factors for single-cell RNA-seq data: avoiding problems with zero

counts”, for consideration as a research paper in Genome Biology.

Single-cell RNA sequencing is a powerful tool for quantifying genome-wide expression in 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 to identify differentially expressed genes associated with phenotypic differences between cells. However, differences in capture efficiency, sequencing depth and other technical effects lead to cell-specific biases that confound direct comparisons between the counts for different cells. Thus, accurate normalization of the count data is critical to removing these biases prior to any downstream analysis.

A number of strategies have been proposed for normalizing RNA sequencing data, e.g., DESeq normalization, trimmed mean of M-values. However, many of these existing methods were designed for analyses of bulk data. Single-cell data is more problematic due to the presence of large numbers of zero counts. We show that existing methods fail to accurately normalize cell-specific biases, using simulations with many zeroes and DE genes. We propose a novel normalization strategy based on normalizing pools of cells, rather than based on the individual cells themselves. Cell-specific normalization factors are then derived from the pool-based estimates. This “deconvolution” approach is robust to high proportions of zero counts and provides accurate normalization of cell-specific biases in our simulations. We observe similar differences in the behaviour of deconvolution compared to existing methods on several real data sets. In particular, we demonstrate that the differences between normalization methods have a substantial impact on the results of downstream quantitative analyses, such as in the detection of differentially expressed or highly variable genes.

In summary, we believe that the method presented in this manuscript will improve the accuracy of normalization for of single-cell RNA sequencing data. This will improve the accuracy of downstream analyses and the validity of any ensuing biological conclusions. 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 Genome Biology. 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