Response: Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells

Andrew McDavid, Greg Finak and Raphael Gottardo

We read Buettner, Natarajan and coauthors’ (BNC) recent paper1 with great interest, given that we have also recently studied this question2. We profiled gene expression in 930 cells targeting canonical cell cycle genes (“ranked” genes) and genes without known cell-cycle annotation (“unranked” genes) across three cell lines. In our own data, we estimated cell cycle to explain 17% of the generalized linear model deviance (analogous to ANOVA R^2) in the typical ranked gene, and 5% of the deviance in the typical unranked gene. We concluded that cell cycle was not introducing substantial variability in single cell gene expression studies when studying the transcriptome at-large. We were initially puzzled by the lack of concordance between the two studies and hope that this letter might resolve some of the discrepancy.

First, we explored BNC’s claim that “the cell cycle explains substantial proportions of the variability.” BNC’s gold-standard estimates of cell cycle-induced variability (R^2 from one-way ANOVA on cell cycle using log expression, results partially shown in supplemental figure 4b/4d), broadly agree with our deviance estimates. We reproduced BNC’s ANOVA and we found the variability attributable to cell cycle in the 8949 unranked genes ranges from 3%-15%, (median-90th percentile gene), and 8%-26% in the 622 ranked genes. The magnitude of these estimates generally agrees with our own observations2.

Somewhat contradictory to the ANOVA results, BNC attribute greater than 30% of the variability to cell cycle via the “scLVM” method (Figure 3 and Supplementary Figure 21 of BNC). We conjectured that the scLVM latent factor would track the largest sources of variability in the data, but it was unclear from the ANOVA whether cell cycle would be the largest contributor. Therefore, we explored other factors that might explain variability in the mESC and mouse T-cell data sets. Using ordinary principal component analysis we found, in both experiments, that the first principal component (PC1), tracked (R^2 > .99) the *geometric library size* – the sum of log expression values over all genes in a cell. Geometric size explained 9% and 29% of expression variance in mESC and T-cells, respectively. This was not surprising given that BNC rely on external spike-ins to account for unwanted technical variation in library preparation and sequencing depth. Normalization based solely on spike-ins assumes that technical variation affects the spike-ins in the same way as it affects endogenous transcript, and that this effect is constant for all genes and read counts. Other authors have found these assumptions difficult to verify. A recent paper by Risso et al.3 reports poor performance of ERCC spike-ins: “Unfortunately, given the troubling behavior of the ERCC spike-ins in our two data sets (Fig. 4), global-scaling normalization factors based on these were unrealistic and led to poorly normalized counts.”

Since the scLVM estimate is tantamount to regularized factor analysis, it comes as no surprise that the scLVM factor also proxies geometric size (R^2=.92, both experiments). Although the geometric size does vary by cell cycle in the mESC experiment, cell cycle only explains 53% of the variance in geometric size and 64% of the variance in the scLVM factor. We therefore conclude that the latent factor estimated by scLVM most directly captures geometric size variability, which happens to be a suitable proxy for cell cycle in the mESC. One might alternatively interpret the correlation between the scLVM cell cycle variability and the Hoechst staining (Supplementary Figure 8) to be a consequence of this partial confounding that geometric size has with cell cycle.

BNC report a reduced scLVM cell cycle variance estimate in (non-cycling) terminally-differentiated neurons as evidence for the specificity of the scLVM method (Supplementary Figure 7). However, this conclusion may not be entirely warranted, as the data sets being compared are not consistently normalized. The neurons are normalized by total library size, while cycling cells are normalized by ERCC spike-ins. Total library size normalization is shown to greatly decrease the variance estimates attributable to cell cycle (Supplementary Figure 21) in the T-cells. A comparison of the performance of scLVM on a non-cycling cell line using ERCC vs. global normalization would identify the degree to which differences in variability are due to biology (cell cycle) or normalization.

BNC also use the scLVM latent factor to derive cell cycle-adjusted expression values. Another interpretation of these adjusted expression values is that a source of nuisance variability (geometric size differences) has been regressed out. Adjustment for nuisance variability using regression has a long record of successful application in gene expression experiments4. To further elucidate the effect of the scLVM adjustment we considered gene set enrichment analysis (GSEA) comparing the two clusters BNC identified in the cycle-adjusted T-cell data (corresponding roughly to corrected expression levels of the differentiation factor GATA3). The top modules identified were: Mitotic Prometaphase, E2F Mediated Regulation of DNA Replication, Deposition of New CENPA Containing Nucleosomes at the Centromere, G2/M Checkpoints, G1/S Specific Transcription and Cell Cycle (all FDR q < .1%). This is somewhat perplexing, given that scLVM purports to remove additive effects due to cell cycle. One explanation could be that geometric size is a weaker proxy for cell cycle in the T-cells than it was in the mESC, so cell cycle was incompletely removed. In general, direct measurement via Hoechst staining could be more appropriate for investigators who require cell cycle as a covariate.

In conclusion, based on BNC’s gold standard calculations, cell cycle comprises less than 7% of the variance in the typical (median) gene. Normalization factors, in particular the sum of log-expression – the geometric size – better explain the variability and that caution may be warranted when using spike-ins for RNA quantitation. It would be of interest to see what biological factors, besides cell cycle, are associated to the geometric size and how the efficiency of rate – limiting steps (e.g. lysis and reverse transcription) affect this factor. All code used to produce some of the results discussed here is available at https://github.com/RGLab/BNCResponse.

1. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* **33,** 155–160 (2015).

2. McDavid, A. *et al.* Modeling Bi-modality Improves Characterization of Cell Cycle on Gene Expression in Single Cells. *PLoS Comput. Biol.* **10,** e1003696 (2014).

3. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology* **32,** 896–902 (2014).

4. Gagnon-Bartsch, J. A. & Speed, T. P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13,** 539–552 (2012).