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 interest since 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 data, we estimated cell cycle to explain 17% of the generalized linear model deviance (analogous to ANOVA R2) in the typical ranked gene, and 5% in the typical unranked gene. We concluded that cell cycle did not introduce substantial variability in single cell gene expression in transcriptomic studies. We were initially puzzled by the lack of concordance with BNC’s findings and hope that this letter might address these discrepancies.

First, we explored BNC’s claim that “cell cycle explains substantial proportions of the variability.” BNC’s gold-standard estimates of cell cycle-induced variability (R2 from one-way ANOVA on cell cycle using log expression, shown in supplemental figure 4b/4d), broadly agree with our previous observations2. We reproduced BNC’s ANOVA and 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.

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 whether cell cycle would be the largest contributor. We explored other covariates that might explain variability in the mESC and mouse T-cell data sets. Using ordinary principal component analysis we found that the first principal component (PC1) tracked the *geometric library size* (R2 > 0.99) – the sum of log expression values over all genes in a cell, and explained 9% and 29% of expression variance in mESC and T-cells, respectively.

The substantial influence of geometric library size is expected since BNC utilize exogenous spike-ins to normalize for unwanted technical variation in library preparation and sequencing depth, which leaves endogenous variation in library size intact, while relying on strong assumptions: technical variation affects both spike-ins and endogenous transcript uniformly for all genes and read counts. Some groups have found these assumptions difficult to verify. A recent paper by Risso et al.3 reports poor performance of ERCC spike-ins in the context of bulk RNASeq: “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.” Importantly, a recent paper by Padovan-Merhar and colleagues4 showed experimentally that single cells compensate for changes in cell volume by increasing transcript abundance. Their work implies that correcting for transcript abundance changes related to cell volume at the single cell level is an important consideration as it may be a significant source of biological variability that is not necessarily associated with the treatment effect of interest, but may be partly or even completely confounded with it. The type of normalization should primarily depend on the scientific questions considered, so no single scheme will be appropriate for all circumstances. A systematic review of normalization schemes would be of great value.

Cell size varies during cell cycle, and unsurprisingly, the scLVM factor proxies geometric size (R2=0.92, both experiments). Although cell cycle can explain 64% of the variance in the scLVM factor, the factor seems to intrinsically restate the geometric size. Within each cycle phase, geometric size remains highly correlated to the scLVM latent factor (R2 = 0.74-0.92). Importantly, the variability in geometric size within cell cycle phase is substantial, and overlaps between cycles, as seen in the PCA plot of mESC cells. We conclude that the latent factor most directly captures geometric size variability, which happens to be a suitable proxy for cell cycle in the mESC. This suggests an alternative interpretation of the correlation between the scLVM cell cycle variability and the Hoechst staining (Supplementary Figure 8). Furthermore the reduced scLVM cell cycle variance estimates in (non-cycling) terminally differentiated neurons are not necessarily evidence of scLVM specificity (Supplementary Figure 7) since 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, and BNC show that total library size normalization can 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; this has a long record of successful application in gene expression experiments5. 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). Of the top 20 modules identified as significantly enriched (q-value < 1%) using both the Reactome and GO Biological process databases, 19 were directly or indirectly related, via DNA repair or replication pathways, to cell cycle in Reactome (n=674 modules total) while 20 of the top 20 were cell-cycle or DNA repair / replication related in GO (n=825 modules total). Although scLVM purports to remove additive cell cycle effects, we conjecture that it is primarily removing geometric size effects, which are a weaker proxy for cell cycle in the T-cells than in the mESC, so cell cycle was incompletely removed. Moreover, our own analysis of the BNC corrected and uncorrected T-cell data using two common dimensionality reduction techniques failed to find compelling evidence for the existence of these two clusters of cells thus identification of these sub-populations is rather sensitive to the dimensionality reduction technique employed. In general, direct measurement via Hoechst staining could be more appropriate for investigators who require cell cycle as a covariate.

In conclusion, BNC’s gold standard calculations suggest that cell cycle comprises less than 7% of the variance in the typical (median) gene, and other normalization factors, in particular the geometric size, better explain the observed variability than cell cycle. Consequently, 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. We thank Buettner, Natarajan and colleagues for their openness in sharing their data and code, and for their willingness to contribute to this discussion. All code used to produce 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. Padovan-Merhar, O. *et al.* Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol. Cell* **58,** 339–352 (2015).

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