We read Buettner, Natarajan and coauthors’ (BNC) recent paper with great interest, since we agree that many technical challenges still remain before single-cell gene expression assays can be fully exploited, and we share their hope that statistical modeling may help resolve some of these challenges. BNC argued that cell cycle can be a large source of variability in single-cell RNA-seq data, which if not accounted for, masks real biological signal. They present a computational approach to estimate and remove unwanted cell-cycle variability.

We found this work intriguing, given that we have also recently studied this question[[1]](#footnote-1). 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. As with BNC, we used the online database “Cyclebase.org” and gene ontology sources to define ranked and unranked sets. Like BNC’s staged mouse embryonic stem cells (mESCs), our cells were also flow sorted by cell cycle using DNA content. 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 of all, BNC’s own 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. When we reproduced BNC’s “gold-standard” ANOVA, 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 reader is left to judge whether these figures warrant the claim that “that the cell cycle explains substantial proportions of the variability.”

We conjectured that the latent factor estimated via “scLVM” would track the largest sources of variability in the data. It was unclear from the ANOVA if cell cycle would be the largest factor; therefore we sought other explanatory factors in the mESC and mouse T-cell data sets. When we considered ordinary principal component analysis, we found, in both experiments, that the first principal component (PC1), simply tracks the *geometric library size* (R^2 > .99) – the sum of log expression values over all genes in a cell. PC1/geometric size explains 9% and 29% of expression variance in mESC and T-cells, respectively.

It is unsurprising that geometric library size tracks the leading principal component, 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[[2]](#footnote-2). 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 latent factor estimation is tantamount to regularized factor analysis, it comes as no surprise that the scLVM factor 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 captures almost as much technical variability, available via principal component analysis, as it does cell-cycle variability. Adjustment for technical variability using regression has a long record of successful application in gene expression experiments2[[3]](#footnote-3)[[4]](#footnote-4), however we found it unfortunate that BNC did not offer any discussion on the benefits the scLVM method might offer compared to previously described approaches.

To further elucidate the effect of the scLVM normalization, we explored the two clusters identified in the corrected Tcell data, wherein the component of expression that scLVM ascribes to cell cycle has been subtracted. We considered gene set enrichment analysis (GSEA) comparing the two clusters BNC identified (corresponding roughly to corrected expression levels of the differentiation factor GATA3). Despite the expectation that the corrected data would not exhibit cell cycle differences, 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%).

On the other hand, when we considered a GSEA on GATA3 expression using uncorrected relative expression values (log counts per million), then the top GSEA modules are: Downregulation Of Erbb2/ Erbb3 Signaling, Transferrin Endocytosis and Recycling, Latent Infection with Mycobacterium Tuberculosis, and Beta Defensins, although none are FDR significant. This suggests *in silico* adjustment for cell cycle should be used with caution, and that direct measurement via Horsch 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 is warranted when using absolute RNA quantitation from single cell RNAseq. 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 (eg lysis and reverse transcription) affects this factor.

1. McDavid, Dennis, Gottardo, 2014. Modeling Bi-modality Improves Characterization of Cell Cycle on Gene Expression in Single Cells [↑](#footnote-ref-1)
2. Risso, Ngai, Speed, 2004. Normalization of RNA-seq data using factor analysis of control genes or samples. [↑](#footnote-ref-2)
3. Gagnon-Bartsch, Speed 2011. Using control genes to correct for unwanted variation in microarray data. [↑](#footnote-ref-3)
4. W. Evan Johnson Cheng Li 2006. Adjusting batch effects in microarray expression data using empirical Bayes methods. [↑](#footnote-ref-4)