HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2018 May 04.

Published in final edited form as:

Nat Biotechnol. 2016 June 09; 34(6): 591–593. doi:10.1038/nbt.3498.

The contribution of cell cycle to heterogeneity in single-cell RNA-seq data

Andrew McDavid[*], **Greg Finak**[*], and **Raphael Gottardo**[*]

[*]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle 98109, WA, USA

In the February 2015 issue, Buettner *et al*.[1] reported a computational approach to estimate and remove latent sources of variation, such as cell cycle stage, on gene expression in single cells. Here, we suggest that this variation is largely explained by geometric library size, rather than cell cycle stage. Furthermore, we argue that the exogenous spike-ins utilized by Buettner *et al*.[1] to adjust for technical variation in library preparation and sequencing depth may have led to poorly normalized read counts.

Recently, 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[2]. We estimated that cell cycle explained 17% of the generalized linear model deviance (analogous to ANOVA R^2) in the typical ranked gene, and 5% in the typical unranked gene. On the basis of these results, we concluded that cell cycle did not cause substantial variability in single cell gene expression. Our findings were not concordant with those reported in Buettner *et al*.[1] and therefore we sought to explain these discrepancies.

First, we explored the claim by Buettner *et al*.[1] that cell cycle explains a substantial proportion of the variability in single cell gene expression. Their gold-standard estimates of cell cycle-induced variability (R^2 from one-way ANOVA on cell cycle using log expression, shown in supplemental Fig. 4b/4d of their paper), broadly agree with our previous observations[2]. We found the variability attributable to cell cycle in the 8,949 unranked genes ranges from 3%–15%, (median-90th percentile gene), and 8%–26% in the 622 ranked genes. However, the scLVM method attributes more than 30% of the variability to its latent factor, which is putatively the cell cycle (Figure 3 and Supplementary Figure 21). We conjectured that the scLVM latent factor would track the principal component 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 mouse embryonic stem cell (mESC) and mouse T-cell data sets.

Using principal component analysis (PCA), we found that the first principal component (PC1) tracked the geometric library size ($R^2 > 0.99$) – 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 Buettner et al.[1] utilize exogenous spike-ins to normalize for unwanted technical variation in library

Correspondence to: Raphael Gottardo.

preparation and sequencing depth. This approach leaves endogenous variation in library size intact, while relying on the assumption that technical variation affects both spike-ins and endogenous transcript uniformly for all genes and read counts. These assumptions have been difficult to verify. A recent paper by Risso et al.[3] reported that ERCC spike-ins led to poorly normalized counts in the context of bulk RNASeq. Furthermore, a recent paper by Padovan-Merhar and colleagues[4] showed experimentally that single cells compensate for changes in cell volume by increasing transcript abundance. Their work implies that correcting for transcript abundance is important because variability in cell volume may be a substantial source of undesired or uninteresting biological variability. The type of normalization should depend primarily on the scientific questions considered, so no single scheme will be appropriate for all circumstances. Finding the optimal method for normalization remains a major challenge in the field.

Cell size varies during cell cycle and the scLVM factor proxies geometric size (R^2=0.92, in 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 (R^2 = 0.74–0.92). Notably, the variability in geometric size within cell cycle phase is substantial, and overlaps between cycles, as seen in the PCA plot of mESC cells (Fig. 1). On the basis of this information, 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 that the correlation between the scLVM cell cycle variability and the Hoechst staining (Supplementary Fig. 8 in ref 1) is a result of cell cycle causing geometric size differences detected as variation in the scLVM latent factor. Furthermore the reduced scLVM cell cycle variance estimates in (noncycling) terminally differentiated neurons are not necessarily evidence of scLVM specificity (Supplementary Fig. 7 in ref 1) because the data sets being compared are not consistently normalized. The neurons are normalized by total library size, whereas cycling cells are normalized by ERCC spike-ins. Buettner *et al.*[1] show that total library size normalization can greatly decrease the variance estimates attributable to cell cycle (Supplementary Fig. 21 in ref 1) in the T-cells, in fact to levels comparable to those observed in the noncycling neurons. A comparison of the performance of scLVM on a noncycling cell line using ERCC versus global normalization would identify the degree to which differences in variability are due to biology (cell cycle) or normalization.

Buettner *et al.*[1] 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 approach has a long record of successful application in gene expression experiments[5]. To further elucidate the effect of the scLVM adjustment, we considered gene set enrichment analysis (GSEA) comparing the two clusters Buettner *et al* 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 the Reactome or GO Biological process databases, 15 were directly or indirectly related, via DNA repair or replication pathways, to cell cycle in Reactome (n=674 modules total) whereas 17 of 20 were related to cell-cycle or DNA repair or replication 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 mESCs, and therefore cell cycle was incompletely removed. Moreover, our own analysis of the 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. Identification of these subpopulations appears to be dependent on 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, we believe that the data presented by Buettner *et al.*[1] suggest that cell cycle comprises less than 7% of the variance in the typical (median) gene, and that geometric library size rather than cell cycle may better explain the observed variability in gene expression. Consequently, caution may be warranted when using spike-ins for RNA quantification. It would be of interest to investigate what biological factors, besides cell cycle, are associated with the geometric size and how the efficiency of rate–limiting steps (such as lysis and reverse transcription) affect these factors. 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.

References

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. 2015; 33:155–160.
2. McDavid A, et al. Modeling Bi-modality Improves Characterization of Cell Cycle on Gene Expression in Single Cells. PLoS Comput Biol. 2014; 10:e1003696. [PubMed: 25032992]
3. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nature Biotechnology. 2014; 32:896–902.
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. 2015; 58:339–352. [PubMed: 25866248]
5. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012; 13:539–552. [PubMed: 22101192]

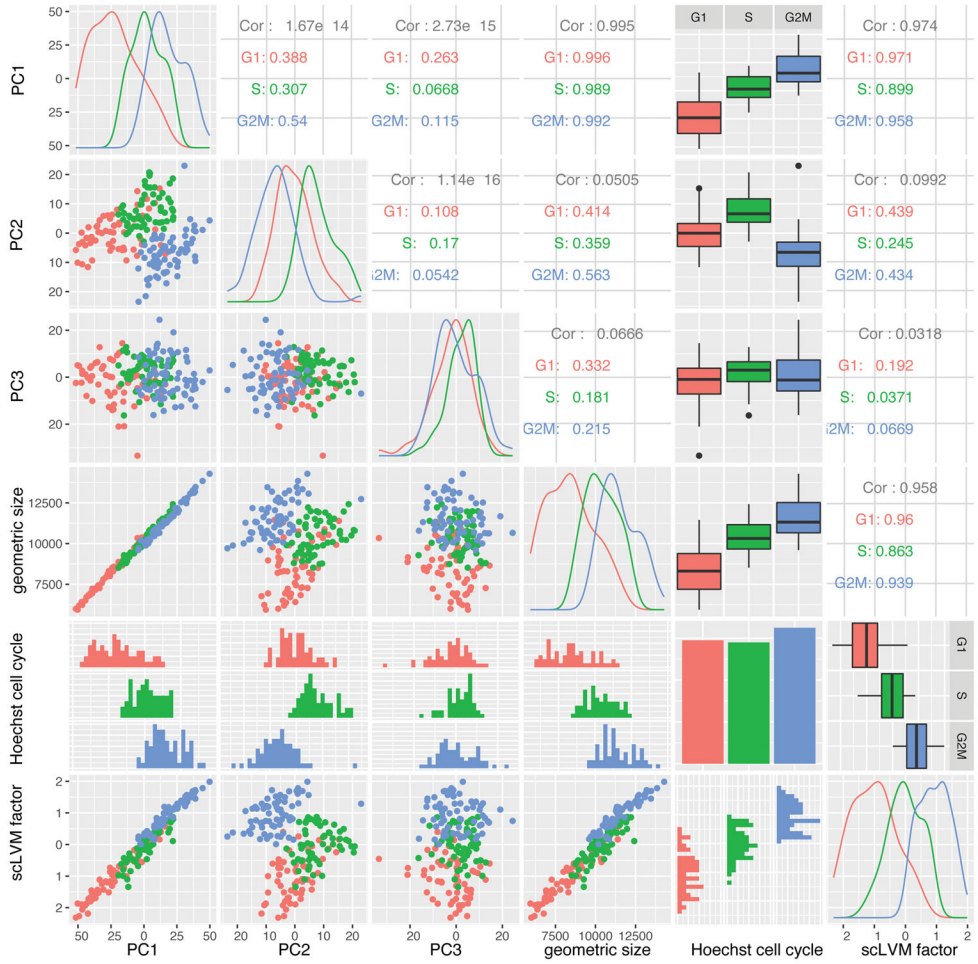

Figure 1.

Correlations and pairwise relationships between principal components analysis (PCA) components 1–3, geometric library size, cell cycle (derived through Hoechst staining), and the scLVM latent factor in the mESC data set.