Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq data analysis

Chiaowen Joyce Hsiao[1,5], PoYuan Tung[2,5], John D. Blischak[1], Jonathan E. Burnett[1], Kenneth A. Barr[2], Kushal K. Dey[3], Matthew Stephens[1,4], and Yoav Gilad[1,2]

[1]Department of Human Genetics, University of Chicago
[2]Department of Medicine, University of Chicago
[3]Department of Epidemiology, Harvard School of Public Health
[4]Department of Statistics, University of Chicago
[5]These authors contributed equally to this work.

*Correspondence should be addressed to C.J.H. (joyce.hsiao1@gmail.com), M.S. (mstephens@uchicago.edu) or Y.G. (gilad@uchicago.edu).

Abstract

Cellular heterogeneity in gene expression is driven by cellular processes such as cell cycle and cell-type identity, and cellular environment such as spatial location. The cell cycle, in particular, is thought to be a key driver of cell-to-cell heterogeneity in gene expression, even in otherwise homogeneous cell populations. Recent advances in single-cell RNA-sequencing (scRNA-seq) facilitate detailed characterization of gene expression heterogeneity, and can thus shed new light on the processes driving heterogeneity. Here, we combined fluorescence imaging with scRNA-seq to measure cell cycle phase and gene expression levels in human induced pluripotent stem cells (iPSCs). Using these data, we developed a novel approach to characterize cell cycle progression. While standard methods assign cells to discrete cell cycle stages, our method goes beyond this, and quantifies cell cycle progression on a continuum. We found that, on average, scRNA-seq data from only five genes predicted a cell's position on the cell cycle continuum to within 14% of the entire cycle, and that using more genes did not improve this accuracy. Our data and predictor of cell cycle phase can directly help future studies to account for cell-cycle-related heterogeneity in iPSCs. Our results and methods also provide a foundation for future work to characterize the effects of the cell cycle on expression heterogeneity in other cell types.

Introduction

Single-cell RNA-sequencing (scRNA-seq) can help characterize cellular heterogeneity in gene expression at unprecedented resolution [1–4]. By using scRNA-seq one can study not only the mean expression level of genes across an entire cell population, but also the variation in gene expression levels among cells [5–10].

There are many reasons for differences in gene expression among cells, with arguably the most obvious candidates being differences in regulation among cell types, and differences in cell cycle phase among cells [11–13]. Cell type and cell cycle phase, while interesting to study directly, are often considered confounders in single cell studies that focus on other factors influencing gene expression [14–16], such as genotype, treatment [17], or developmental time [5, 18]. The ability to characterize, correctly classify, and correct for cell type and cell cycle phase are therefore important, even in studies that do not specifically aim to study either of these factors.

For these reasons, many studies have used single cell data to characterize the gene regulatory signatures of individual cells of different types and of cells at different cell cycle phases (e.g., [14, 19, 20]). Often the ultimate goal of such studies is to be able to develop an effective approach to account for the variation associated with cell cycle or cell type. To characterize cell cycle phase, a common strategy in scRNA-seq studies is to first use flow cytometry to sort and pool cells that are in the same phase, followed by single-cell sequencing of the different pools [14, 19]. Unfortunately, in this common study design, cell cycle phase is completely confounded with the technical batch used to process single-cell RNA. This design flaw can inflate expression differences between the pools of cells in different cell cycle phase, resulting in inaccurate estimates of multi-gene signatures of cell cycle phase. When cells are not sorted before sequencing, cell cycle phase is typically accounted for by classifying the cells into discrete states based on the expression level of a few known markers [21].

Regardless of whether or not cells are sorted, all single-cell studies to date have accounted for cell cycle by using the standard classification of cell cycle phases, which is based on the notion that a cell passes through a consecutive series of distinct phases (G1, S, G2, M, and G0) marked by irreversible abrupt transitions. This standard definition of cell phases, however, is based on physiological observations and low-resolution data.

The traditional approach to classify and sort cells into distinct cell cycle states relies on a few known markers, and quite arbitrary gating cutoffs. Most cells of any given non-synchronized culture do not, in fact, show an unambiguous signature of being in one of the standard discrete cell cycle phases [5, 22, 23]. This makes intuitive sense: while from a physiological perspective, transitions between cell cycle states can be clearly defined (the DNA is either being replicated or not; the cell is either dividing or not), this is not the case when we try to define the cell states using molecular data. Indeed, we do not expect the gene expression signature of cell state transitions to occur in abrupt steps but rather to be a continuous process. High resolution single-cell data can provide a quantitative description of cell cycle progression and thus can allow us to move beyond the arbitrary classification of cells into discrete states.

From an analysis perspective, the ability to assign cells to a more precise point on the cell cycle continuum could capture fine-scale differences in the transcriptional profiles of single cells - differences that would be masked by grouping cells into discrete categories. Our goal here is therefore to study the relationship between cell cycle progression and gene expression at high resolution in single cells, without confounding cell cycle with batch effects as in [14, 19]. To do so, we used fluorescent ubiquitination cell cycle indicators (FUCCI) [24] to measure cell cycle progression, and scRNA-seq to measure gene expression in induced pluripotent stem cells (iPSCs) from six Yoruba individuals from Ibadan, Nigeria (abbreviation: YRI). To avoid the confounding of cell cycle with batch, we did not sort the cells by cell cycle phase before we collected the RNA-seq data. Instead, we measured FUCCI fluorescence intensities on intact single cells that were sorted into the C1 Fluidigm plate, prior to the preparation of the sequencing libraries. We also used a balanced incomplete block design to avoid confounding individual effects with batch effects. Using these data, we developed an analysis approach to characterize cell cycle progression on a continuous scale. We also developed a predictor of cell cycle progression in the iPSCs based on the scRNA-seq data. Our experimental and analytical strategies can help future scRNA-seq studies to explore the complex interplay between cell cycle progression, transcriptional heterogeneity, and other cellular phenotypes.

Results

Study design and data collection

We generated FUCCI-iPSCs using six YRI iPSC lines that we had characterized previously [25] (Fig. 1; see Methods for details). FUCCI-expressing iPSCs constitutively express two fluorescent reporter constructs transcribed from a shared promoter [24, 26]. Reporters consist of either EGFP or mCherry fused to the degron domain of Geminin (geminin DNA replication inhibitor) or Cdt1 (Chromatin licensing and DNA replication factor 1). Due to their precisely-timed and specific regulation by the ubiquitin ligases APC/C and SCF, Geminin and Cdt1 are expressed in an inverse pattern throughout the cell cycle. Specifically, Geminin accumulates during S/G2/M and declines as the cell enters G1, whereas Cdt1 accumulates during G1 and declines after the onset of S phase. Thus, FUCCI reporters provide a way to assign cell cycle phase by tracking the degradation of Geminin-EGFP and Cdt1-mCherry through the enzymatic activity of their corresponding regulators, APC/C and SCF.

We collected FUCCI fluorescence images and scRNA-seq data from the same single cells using an automated system designed for the Fluidigm C1 platform (see Methods). After image capture, we prepared scRNA-seq libraries for sequencing using a SMARTer protocol adapted for iPSCs [27]. To minimize bias caused by batch effects [27, 28], we used a balanced incomplete block design in which cells from unique pairs of iPSC lines were distributed across fifteen 96-well plates on the C1 platform (see Supplemental Fig. S1 for our C1 study design). We also included data from one additional plate (containing individuals NA18855 and NA18511), which we collected as part of a pilot study in which we optimized our protocols. In total, we collected data from 1,536 scRNA-seq samples distributed across 16 C1 plates.

Single-cell RNA-sequencing

To help with quality control, we used DAPI staining to quantify the number of cells captured in each C1 well (see Methods for details). Combined with the common scRNA-seq data metrics, we determined, for our data, criteria for including single-cell samples and high quality scRNA-seq expression data (see Supplemental Fig. S2. We obtained an average of 1.7 +/- 0.6 million sequencing reads per sample (range=0.08-3.0 million). After quality control, we retained RNA-seq data from 11,040 genes measured in 888 single cells, with a range of 103 to 206 cells from each of the six individuals (see Supplemental Fig. S3, Supplemental Fig. S4). We standardized the molecule counts to counts per million (CPM) and transformed the data per gene to obtain a standardized normal distribution. We retained all genes with CPM 1 or higher in order to evaluate as many genes as possible. This resulted in a mean gene detection rate of 70 % across cells (standard deviation of 25 %, no significant difference between the six cell lines, see Supplemental Fig. S4).

We used principal components analysis (PCA) to assess the global influence of technical factors on expression, including plate, individual, and read depth (see Supplemental Fig. S5). The primary source of sample variation in our data was the proportion of genes detected (>1 $\log_2$ CPM; adj. R-squared=0.39 for PC1; 0.25 for PC2), consistent with results from previous studies [28]. Reassuringly, we found that the proportion of genes detected in our samples showed a stronger correlation with the number of reads mapped (adj R-squared=0.32) than with plate (adj. R-squared=0.01) or individual (adj. R-squared=0.09). Thus, we confirmed that further statistical adjustment to account for

batch effects will not yield noticeably different results. This demonstrates that our use of a balanced incomplete block design was an effective strategy to minimize the effects of confounding technical variables.

Quantifying continuous cell cycle phase using FUCCI intensities

Proceeding with the 888 single cells for which we had high quality RNA-seq data, we turned our attention to the corresponding FUCCI data. To summarize FUCCI intensities, we defined a fixed cell area for all sample images (100 x 100 px) in order to account for differences in cell size. We computed two FUCCI scores for each cell by individually summing the EGFP (green) and mCherry (red) intensities in the fixed cell area and correcting for background noise outside the defined cell area (see Methods for more details). Because images were captured one plate at a time, we scanned the data for evidence of batch effects. We found mean FUCCI scores to be significantly different between plates (F-test P-value $<$ 2e-16 for both EGFP and mCherry, see Supplemental Fig. S6 for comparisons between C1 plates, Supplemental Fig. S7 for comparisons between the six cell lines). We hence applied a linear model to account for plate effects on FUCCI scores without removing individual effects (FUCCI score $\sim$ plate + individual).

FUCCI intensities are commonly used to sort cells into discrete cell cycle phases. For example, cells expressing EGFP-Geminin in the absence of mCherry-Cdt1 would traditionally be assigned to G2/M, cells with the opposite pattern of expression would be assigned to G1, and cells expressing equal amounts of EGFP-Geminin and mCherry-Cdt1 would be assigned to the S/G2 transition [24]. As a representative of this approach, we applied Partition Around Medoids (PAM) from [29] to FUCCI scores to assign single-cell samples to G1, S, and G2/M phase (G1 384 cells, S 172 cells, G2/M 332 cells, see Supplemental Fig S8). Henceforth, the classification obtained from PAM is referred to as PAM-based classification. However, FUCCI intensities are known to be continuously distributed within each phase [24], suggesting that they could also be used to quantify cell cycle progression through a continuum (conventionally represented using radians in the range $[0, 2\pi]$). With this in mind, we ordered the corrected FUCCI scores by phase and plotted them on a unit circle, using the co-oscillation of mCherry-Cdt1 and EGFP-Geminin to infer an angle, or 'FUCCI phase', for each cell (Fig. 2A; see Methods). For example, Fig. 2B shows that as a cell progresses through $\pi/2$ to π radians, mCherry-Ctd1 intensity decreases from its maximum, while EGFP-Geminin intensity changes from negative to positive, suggesting progression through G1/S transition. Overall, FUCCI phase explains 87% of variation in mCherry intensity and 70% of variation in EGFP intensity.

We next sought to identify genes whose expression levels vary in a cyclic way through the cell cycle, as captured by FUCCI phase. Specifically, we used a non-parametric smoothing method, trend filtering [30], to estimate the change in expression for each gene through the cell cycle. We refer to these estimates as the "cyclic trend" for each gene. We used a permutation-based test (see Methods) to assess the significance of each inferred cyclic trend, and ranked the genes by statistical significance. Reassuringly, genes with a significant cyclic trend were strongly enriched for known cell cycle genes (using the 622 genes annotated in Whitfield et al. [31], we found Odds Ratio=25.79 for the 101 significant cyclic genes, Odds Ratio=31 for the top 5 significant cyclic genes, 30 for the top 50 genes, and 27 for the top 100 genes, Fisher's exact test P-value $<$ .001, see Supplemental Tables S1, S2 for the gene list, see Supplemental Fig. S9A,S9B for

PAM-based results). These results provide strong independent support that the inferred FUCCI phase is indeed meaningfully capturing cell cycle progression.

For illustration, Fig. 2C shows the cyclic trends for the top 5 significant cyclic genes: *CDK1*, *UBE2C*, *TOP2A*, *HIST1H4E*, *HIST1H4C*. These genes have all been previously identified as cell cycle genes in synchronization experiments of HeLa cells [31] and in scRNA-seq studies of FUCCI-sorted cells [19]. *CDK1* (Cyclin Dependent Kinase 1, also known as *CDC2*) promotes the transition to mitosis. *TOP2A* (DNA topoisomerase II-alpha) controls the topological state of DNA during cell state transitions. *UBE2C* (Ubiquitin Conjugating Enzyme E2 C) is required for the degradation of mitotic cyclins and the transition to G2 stage. Finally, *HIST1H4C*, and *HIST1H4E* (Histone gene cluster 1, H4 histone family) are replication-dependent histone genes expressed mainly during S phase.

Predicting FUCCI phase from gene expression data

Our supervised approach

Building on these results, we developed a statistical method for predicting continuous cell cycle phase from gene expression data. The intuition behind our approach is that given a set of labeled training data – cells for which we have both FUCCI phase (Y) and scRNA-seq data (X) – our trend-filtering approach learns the cyclic trend for each gene (i.e., $p(X|Y)$). We combine this with a prior for the phase ($p(Y)$) using the idea of a "naive Bayes" predictor, to predict FUCCI phase from gene expression (i.e., $p(Y|X)$). Given scRNA-seq data, X, on any additional cell without FUCCI data, we can then apply this method to predict its FUCCI phase, Y (see Methods for more details). Henceforth, our continuous predictor is referred to as peco.

To assess the performance of our predictor, we applied six-fold cross-validation. In each fold, we trained our predictor on cells from five individuals and tested its performance on cells from the remaining individual. This allowed us to assess the ability of our predictor to generalize to individuals not seen in training. We measured the prediction error as the difference between the predicted phase and the measured FUCCI phase (as a percentage of the entire cycle, 2π; see Fig. 3A). Note that since phases lie on a circle, the maximum possible error is 50% of the circle, and the expected error from random guessing would be 25% of the circle. Using our approach, on average, we were able to predict a cell's position on the cell cycle continuum to within 14% of the entire cycle (i.e., $.28\pi$ between inferred phase and FUCCI phase).

Supplemental Fig. S10A shows the performance of predictors built using between 2 and 50 genes. The genes were ranked and included in the predictors according to the significance of their cyclic trend. We observed that the mean prediction error was robust to the number of genes included in the predictor, and that the simplest predictor using only the top five genes (*CDK1*, *UBE2C*, *TOP2A*, *HIST1H4E*, *HIST1H4C*) performed just as well as the predictors with more genes.

We also checked the robustness of our predictors for data with lower effective sequencing depth compared to the C1 platform (e.g., Drop-seq and 10X). Specifically, we repeated the analysis above after thinning the test data (total sample molecule count in the unthinned data was 56,724 $\pm$ 12,762) by a factor of 2.2 (total sample molecule count 25,581 $\pm$ 15,220) and 4.4 (total sample molecule count 13,651 $\pm$ 13,577). Results in Supplemental Fig. S10C,D show that the predictors based on fewer genes (e.g. 5-15) were relatively robust to this thinning; predictors based on more genes showed worse per-

formance in the lower-count data. These results were not surprising, as we demonstrated in the unthinned data (Supplemental Fig. S9A) that adding genes with weak signals increased prediction error.

Comparisons with existing methods on our data

Several methods exist for making inferences on cell cycle from RNA-seq data. Here we consider two methods that assign cells to discrete cell cycle states (Seurat [21] and Cyclone [32]) and two methods that attempt to infer a "cyclic ordering" of cells from RNA-seq data in an unsupervised way (Oscope [19] and reCAT [33]). Coming to concrete conclusions that one analytic method is better than another is difficult in most settings, and is particularly difficult in settings where, as here, "gold standard" data are hard to come by. It is further complicated here by the fact that the methods differ in their precise goals (e.g. discrete vs continuous assignments, and supervised vs unsupervised assignments). Nonetheless, we compared the methods on both our data and on other data sets in an effort to provide some indication of their differences and commonalities.

First we ran the four other methods on our RNA-seq data from all 888 single cells, and compared their results with our FUCCI data on the same cells.

For Seurat and Cyclone, we compared the discrete classifications (G1 vs S vs G2/M) they produced from RNA-seq data with the corresponding classifications obtained from FUCCI data using the PAM-based method from [29]. Neither the Seurat nor Cyclone classifications agreed well with the FUCCI data. Treating the FUCCI results as a gold standard, Seurat misclassification rates were 78% (G1), 74% (S), 43% (G2/M); and Cyclone misclassification rates were 34% (G1), 88% (S), 31% (G2/M). See Supplemental Fig. S11A,B).

For the unsupervised methods Oscope and reCAT, we first applied each method to infer an ordering of cells from the RNA-seq data, and then assessed whether the inferred orderings produced cyclic patterns in the FUCCI scores (which one would expect if the inferred orderings accurately represented cell cycle). In both cases the ordering explained only very little variation in the FUCCI scores, with Oscope slightly higher than reCAT (Oscope: 13% EGFP, 16% mCherry; reCAT: 4% EGFP, 9% mCherry, see Supplemental Fig. S11D,F). In contrast, inferred phase from peco explained an average 29% of the variation in EGFP score and an average of 24% of the variation in mCherry score (see Supplemental Fig. S12).

To directly compare existing methods with our method requires translating results from existing methods into a continuous predictor of cell cycle phase that is comparable with our continuous predictor. For Oscope/reCAT, we did this by using their cyclic ordering to assign cells to equidistant points on the unit circle. For Seurat/Cyclone, we built a continuous predictor based on the Seurat/Cyclone phase-specific scores. Specifically, we applied the same approach used to derive FUCCI phase to transform the two Seurat scores and the three Cyclone scores to cell cycle angles (see Supplemental Fig. S13 for an example of Seurat score transformation).

This comparison will likely favor our predictor because existing methods were not optimized for continuous phase predictions (indeed, no method other than ours has been so optimized). Also our predictor was trained on the same cell types as are being used for assessment. With these caveats in mind, on these data our predictor outperformed predictors built from existing methods, with lower prediction error than all the other methods on all cell lines (and in most cases significantly lower at P-value $< .05$; see

Supplemental Fig. S14). Overall, the mean prediction error of our predictor across the six cell lines was approximately 80% of the Cyclone-based predictor and 60% of the Seurat/Oscope/reCAT-based predictors (See Figure 3B).

Visual comparisons of the results from the different methods on the top 5 cyclic genes used by peco (see Figure 3C, Supplemental Fig. S15A, S15B, S15C, S15D, S15E), suggest that on these data Oscope agrees most closely with peco than other methods; in particular results from peco and Oscope show a clearer cyclic trend in the expression levels of *HIST1H4E* and *HIST1H4C* than do other methods.

Comparisons on data from Leng et al

Leng et al. [19] collected scRNA-seq and FUCCI data on Human embryonic cells (hESCs). The cells were transfected with the same FUCCI reporters used in our study (in fact, the co-author Dr. Chris Barry generously gifted us their plasmid). However, in contrast to our study, cells were first sorted into discrete cell cycle phases based on the FUCCI data (G1, S, and G2/M, henceforth referred to as "gating-based classification"), and then cells in each phase were prepared on different 96-well C1 plates prior to RNA-seq. In contrast to our data, this design means that plate effects are confounded with cell cycle phase, which is far from ideal. In addition, the sorted cells are not a random sample of all cells across all cell cycle states, but rather represent cells whose FUCCI data place them confidently into one of three discretely-defined cell cycle states. These issues were major motivations for our own data collection efforts. However, since this is one of the very few available single-cell datasets with RNA-seq and FUCCI data on the same cells, we nonetheless compared methods on these data.

We analyzed the 247 FUCCI-expressing hESC single-cell samples from [19] that passed quality control: 91 G1 phase, 80 S phase, and 76 G2/M phase. We applied peco and the four existing methods to these data (for peco we used only the top 4 cyclic genes, because *HIST1H4E*, was not mapped in these data).

Comparing results from the three continuous assignment methods (peco, Oscope and reCAT), we found that the orderings from reCAT agree most closely with the gating-based classification (Fig. 4B). Results from peco also show strong agreement with gating-based classification, but the S-phase cells are spread out on either side of the G2/M cells, rather than only preceding them as in the reCAT results. In contrast, the ordering from Oscope show less agreement with the gating-based classification. (Quantifying these qualitative statements is not straightforward because it is not obvious how to quantitatively compare a continuous cyclic ordering with a discrete classification; nonetheless we believe the qualitative patterns are clear in Fig. 4B.)

Turning to the discrete classification methods (Seurat and Cyclone), in these data the Cyclone discrete assignments show much better agreement with the gating-based patterns than Seurat (Cyclone misclassification rates 0% G1, 2.5% S, and 0% G2/M; Seurat misclassification rates 89% G1, 25% S, and 21% G2/M).

Overall our results suggest the need for more research and better data to quantify the accuracy and relative performance of the different available methods, including ours.

Discussion

In this study we sought to characterize the effects of cell cycle progression on gene expression data from single cells (iPSCs), by jointly measuring both cell cycle phase (via

FUCCI) and expression (via scRNA-seq) on the same cells. Our study differs in two key ways from previous similar studies. First, unlike the most commonly-cited previous studies [14,19], our experimental design avoided confounding batch/plate effects with cell cycle phase. In these previous studies, cells were FACS-sorted by discrete cell cycle stage and loaded onto different C1 plates, making it difficult to decouple batch effects from cell cycle effects [28]. Second, our study focused on characterizing cell cycle progression in a continuum, rather than as abrupt transitions between discrete cell cycle phases.

We found that a simple predictor, based on 5 genes with a cyclic expression pattern (*CDK1*, *UBE2C*, *TOP2A*, *HIST1H4C*, *HIST1H4E*), was sufficient to predict cell cycle progression in our data, and that adding information from other genes did not improve prediction accuracy. That these particular genes should be helpful predictors of cell cycle is not entirely surprising, as they have been reported as potential markers in previous studies, including synchronization experiments in HeLa cells [31] and yeast [34], and in previous scRNA-seq studies of FUCCI-sorted hESCs [19]. However, our finding that additional genes did not further improve prediction accuracy is perhaps more surprising, and contrasts with the common use of dozens of genes for cell cycle prediction (e.g., Seurat [21]). Of course, our results do not imply that only these five genes are associated with cell cycle progression in iPSCs, only that additional genes provide redundant information in our data.

As noted in the Introduction, one reason to estimate cell cycle from RNA-seq data is to control for it when performing other downstream tasks. Although our methods provide a way to estimate cell cycle information, they do not dictate a specific way to control for cell cycle in downstream analyses. Indeed, how best to do this remains an interesting and open question, and the ease with which it can be achieved will inevitably depend on the downstream analyses being performed. For example, if the downstream analyses rely on Gaussian models for transformed single-cell data (e.g., [35, 36]) then it may suffice to first regress out the effects of cell cycle from the transformed data (e.g. using a non-parametric regression method such as trend filtering, to allow for the non-linear trends that must occur in any cyclic phenomenon) before applying downstream analyses to the residuals. On the other hand, if the downstream methods rely on explicit models for count data (e.g., [37]) then controlling for cell cycle may be more complicated and require further methodological development. However, we note that these issues are not unique to our approach: controlling for cell cycle within count-based analyses poses additional methodological challenges for whatever method is used to estimate cell cycle.

One important question is how well our methods will generalize beyond the data collected here. We believe that our methods should be useful in other iPSC studies because we were able to effectively predict cell cycle progression in cells from one individual using scRNA-seq data from five other individuals (that is, our approach worked well in out-of-sample prediction assessment). However, further data are required to assess how well our methods generalize to studies involving different cell types than the iPSCs studied here.

Single-cell omics technology allows us to characterize cellular heterogeneity at an ever-increasing scale and high resolution. We argue that the standard way of classifying biological states in general, and cell cycle in particular, to discrete types, is no longer sufficient for capturing the complexities of expression variation at the cellular level. Our study provides a foundation for future work to characterize the effect of the cell cycle at single-cell resolution and to study cellular heterogeneity in single-cell expression studies.

Methods

FUCCI-iPSC cell lines and cell culture

Six previously characterized YRI iPSCs [25], including three females (NA18855, NA18511, and NA18870) and three males (NA19098, NA19101, and NA19160), were used to generate FUCCI iPSC lines by the PiggyBAC insertion of a cassette encoding an EEF1A promoter-driven mCherryCDT1-IRES- EgfpGMNN double transgene (the plasmid was generously gifted by Dr. Chris Barry) [19, 24]. Transfection of these iPSCs with the plasmid and Super piggyBacTM transposase mRNA (Transposagen) was done using the Human Stem Cell Nucleofector Kit 1 (VAPH-5012) by Nucleofector 2b Device (AAB-1001, Lonza) according to the manual. Notably, single-cell suspension for the transfection was freshly prepared each time using TrypLETM Select Enzyme (1X) with no phenol red (ThermoFisher) to maintain cell viability. For standard maintenance, cells were split every 3–4 days using cell release solution (0.5 mM EDTA and NaCl in PBS) at the confluence of roughly 80%.

After two regular passages on the 6-wells, the transfected cells were submitted to fluorescence activated cell sorting (FACS) for the selection of double positive (EGFP and mCherry) single cells. To increase the cell survival after FACS, Y27632 ROCK inhibitor (Sigma) was included in E8 medium (Life Technologies) for the first day. FACS was performed on the FACSAria IIIu instrument at University of Chicago Flow Cytometry Facility. Up to 12 individual clones from each of the six iPSC lines were maintained in E8 medium on Matrigel-coated tissue culture plates with daily media feeding at 37°C with 5% (vol/vol) CO2, same as regular iPSCs. After another ten passages of the FUCCI-iPSCs, a second round of FACS was performed to confirm the activation of the FUCCI transgene before single-cell collection on the C1 platform.

Single-cell capture and image acquisition

Single-cell loading, capture, and library preparations were performed following the Fluidigm protocol (PN 100-7168) and as described in Tung et al. [27]. Specifically, the reverse transcription primer and the 1:50,000 Ambion® ERCC Spike-In Mix1 (Life Technologies) were added to the lysis buffer, and the template-switching RNA oligos which contain the UMI (6-bp random sequence) were included in the reverse transcription mix. A cell mixture of two different YRI FUCCI-iPSC lines was freshly prepared using TrypLE^{TM} at 37°C for three minutes. Cell viability and cell number were measured to have an equal number of live cells from the two FUCCI-iPSC lines. In addition, single-cell suspensions were stained with 5 uM Vybrant^{TM} DyeCycle^{TM} Violet Stain (ThermoFisher) at 37°C for five minutes right before adding the C1 suspension buffer.

After the cell sorting step on the C1 machine, the C1 IFC microfluidic chip was immediately transferred to JuLI Stage (NanoEnTek) for imaging. The JuLI stage was specifically designed as an automated single-cell observation system for C1 IFC vessel. For each cell capture site, four images were captured, including bright field, DAPI, EGFP, and mCherry. The total imaging time, together with the setup time, was roughly 45 minutes for one 96-well C1 IFC. The JuLI Stage runs a series of standardized steps for each C1 IFC and for each fluorescence channel, separately. First, the camera scans the four corners of the C1 IFC and sets the exposure setting accordingly. Then, the camera proceeds to capture images of each C1 well.

Library preparation and read mapping

For sequencing library preparation, tagmentation and isolation of 5′ fragments were performed as described in Tung et al. [27]. The sequencing libraries generated from the 96 single-cell samples of each C1 chip were pooled and then sequenced in two lanes on an Illumina HiSeq 2500 instrument using TruSeq SBS Kit v3-HS (FC-401-3002).

We mapped the reads with Subjunc [38] to a combined genome that included human genome GRCh37, ERCC RNA Spike-In Mix 1 (Invitrogen), and the mCherry and EGFP open reading frames from the FUCCI plasmid (we included the latter to ensure that the transgene was being transcribed). We focused on the protein-coding regions of the genome (RNA-seq), which have been well-established and stable for a long time. Newer genome builds improve the resolution of hard-to-map regions of the genome (e.g. lots of repeats) that would be more relevant if one is studying structural variation. Next, we extracted the UMIs from the 5′ end of each read (pre-mapping) and deduplicated the UMIs (post-mapping) with UMI-tools [39]. We counted the molecules per protein-coding gene (Ensembl 75, February 2014) with featureCounts [40]. Note that we observed quantitatively similar results when using genome build GRCh38 and gene annotations from Ensembl 96 (April 2019) (Supplemental Fig. S16). Lastly, we matched each single cell to its individual of origin with verifyBamID [41] by comparing the genetic variation present in the RNA-seq reads to the known genotypes, as previously described [27].

Filtering and normalization of gene expression data

We used DAPI staining results to inform our RNA-seq quality control analysis and establish criteria for high-quality single-cell samples in two steps. First, we used DAPI staining results to classify each C1 well into empty or non-empty wells. We then used data from the empty wells to determine filtering criteria for the non-empty wells (see Supplemental Fig. S2A): number of mapped reads, percentage of unmapped reads, percentage of ERCC reads, and percentage of genes detected to have at least one reads. Second, we determined the number of cells captured in each C1 well using linear discriminant analysis (LDA; see our previous work for the rationale [27]). We fitted two LDA models: 1) number of cells per well $\sim$ gene molecule count + concentration of cDNA amplicons, and 2) number of cells per well $\sim$ read-to-molecule conversion efficiency of ERCC spike-in controls + read-to-molecule conversion efficiency of endogeneous genes. We used DAPI staining results to determine the number of cells captured in each well. Supplemental Fig. S17 shows the results of our LDA analysis. These quality control steps have been applied and described in more details in our previous work [27].

In summary, our quality control criteria include the following:

- Only one cell observed per well
- At least one molecule mapped to EGFP (to ensure the transgene is transcribed)
- The individual assigned by verifyBamID was included on the C1 chip
- At least 1,309,921 reads mapped to the genome
- Less than 44% unmapped reads
- Less than 18% ERCC reads
- At least 6,292 genes with at least one read

After sample filtering, we excluded genes based on the following criteria.

- Over-expressed genes with more than 6^4 molecules across the samples.
- Lowly-expressed genes with sample average of CPM less than 2.

In total, we collected 20,327 genes from 1,536 scRNA-seq samples after read mapping. After the quality filtering steps described above, we were left with 888 samples and 11,040 genes. We standardized the molecule counts to CPM using per-sample total molecule count from the 20,327 genes pre-filtering.

FUCCI image data analysis and FUCCI phase

Images were segmented using the EBImage package in R/Bioconductor [42]. We used the nuclear channel (DAPI) to identify the location of cells in each C1 well. First, we normalized pixel intensities in each image and applied a 10 pixel median filter. Next, we generated a nuclear mask using the EBImage adaptive thresholding algorithm. We filled holes in the resulting binary image and smoothed borders with a single round of erosion and dilation. Finally, we identified individual nuclei using the EBImage bwlabel function. The code that implements these methods is available at `https://raw.githubusercontent.com/jdblischak/fucci-seq/master/code/create_mask.R`.

We generated 100 by 100 pixel cell images centered on each nucleus centroid for the remaining channels. For each channel, we estimated the background florescence in each image by taking the median pixel intensity of all pixels that were not within the selected 100 pixel squares. We then subtracted this background intensity from the value of each pixel within the cell images. Finally, we summed and log-transformed the background-removed fluorescence intensities across the 100 by 100 pixel cell image. This yielded two FUCCI scores summarizing fluorescence intensities of mCherry-Cdt1 and EGFP-Geminin for each cell.

We used the FUCCI scores - log10 sum of fluorescence intensity in the cell area after background correction - to infer an angle for each cell on a unit circle. Specifically, we applied PCA to transform the two FUCCI scores to orthogonal vectors. Using the PCs of the FUCCI scores, we inferred an angle for each cell where the angle is the inverse tangent function of (PC2/PC1). We refer to these angles as FUCCI phase, namely the cell cycle phase estimates based on FUCCI intensities.

Estimating cyclic trends in gene expression data

To estimate the cyclic trend of gene expression, we ordered the single-cell samples by the measured FUCCI phase and applied nonparametric trend filtering. We quantile-normalized CPM values of each gene to a standard normal distribution. This way, the samples with zero molecule count were assigned the lowest level of gene expression. We applied quadratic (second order) trend filtering using the *trendfilter* function in the *genlasso* package [30]. The *trendfilter* function implements a nonparametric smoothing method which chooses the smoothing parameter by cross-validation and fits a piecewise polynomial regression. In more specifics: The *trendfilter* method determines the folds in cross-validation in a nonrandom manner. Every k-th data point in the ordered sample is placed in the k-th fold, so the folds contain ordered subsamples. We applied five-fold cross-validation and chose the smoothing penalty using the option *lambda.1se*: among all possible values of the penalty term, the largest value such that the cross-validation standard error is within one standard error of the minimum. Furthermore, we desired that the estimated expression trend be cyclical. To encourage this, we concatenated the ordered gene expression data three times, with one added after another. The quadratic trend filtering was applied to the concatenated data series of each gene. The estimates from the

middle series were extracted and taken as the estimated cyclic trend of each gene. Using this approach, we ensured that the estimated trend be continuous at the boundaries of the ordered data: the estimates at the beginning always meet the estimates at the end of the ordered data series.

We used a permutation-based test to assess the significance of each inferred cyclic trend. For each gene, we computed the proportion of variance explained (PVE) by the inferred cyclic trend in the expression levels. Then, we constructed an empirical null distribution of PVE. We randomly chose a gene with less than 10 % of the cells observed as undetected (CPM ≥ 1) and permuted the expression levels in the selected gene 1,000 times. Each time, we fit *trendfilter* and computed PVE of the cyclic trend. We found that the significance (p-value) of the inferred cyclic trend was more conservative when the empirical null was based on a gene with low proportion of undetected cells, compared to when the empirical null was based on a gene with high proportion of detected cells ($>$ 80 %). Using these empirical p-values, we were able to assess significance of the cyclic trends for each gene.

Predicting quantitative cell cycle phase of single cells: a supervised learning approach

Our goal was to build a statistical method to predict continuous cell cycle phase from gene expression data. We implemented the method in a two-step algorithm. In the first step, we trained our predictor on data from 5 individuals and learned the cyclic trend for each gene using *trendfilter*. In the second step, we applied the predictor and used the gene-specific trends to compute the likelihood of gene expression levels in the test data for each cell. We evaluated the likelihood on grid points selected along a circle (default to 100 equally-spaced cell cycle phases). Finally, we assigned each cell in the test data to a grid point (phase) at which its likelihood reaches the maximum. Because we independently assigned each cell based on its gene expression levels, prediction accuracy does not depend on the number of cells in the test data.

Notations

- $(Y_n^{train}, \hat{\theta}_n^{train})_{n=1,\ldots,N}$: For each individual cell n in the training sample, we denote $Y_n^{train} = (Y_{1n}^{train}, \ldots, Y_{Gn}^{train})'$ as the quantile-normalized gene expression vector, and $\hat{\theta}_n$ the FUCCI-based cell cycle phases. The single-cell samples are ordered in FUCCI time, where $0 \leq \hat{\theta}_1^{train} < \ldots \hat{\theta}_N^{train} < 2\pi$.

- $(Y_m^{test}, \hat{\theta}_m^{test})_{m=1,\ldots,M}$: For each cell m in the test data, $Y_m^{test} = (Y_{1m}^{test}, \ldots, Y_{Gm}^{test})'$ denotes the $\log_2$ normalized gene expression vector. The method estimates $\hat{\theta}_m^{test}$ the cell cycle phase for each sample m.

- $(\hat{f}_g, \hat{\sigma}_g)_{g=1,\ldots,G}$: Using the training data Y^{train}, we estimate a function $\hat{f}_g$ for each gene describing the cyclic trend of gene expression levels in FUCCI phase. f is a cyclic function assumed to be continuous at 0 and 2π.

Methods

1. Estimate $(\hat{f}_g, \hat{\sigma}_g)$ using Y_g^{train} gene expression levels of gene g

(a) Sort the gene expression levels Y_g^{train} in ascending order according to the cell times $(\hat{\theta}_n^{train})_{n=1,\ldots,N}$.

(b) For each gene g, fit a piecewise polynomial function $\hat{f}_g$ using *trendfilter*. (This function uses internal 5-fold cross-validation to determine an appropriate amount of smoothing for each g).

(c) Compute the gene-specific standard error $\hat{\sigma}_g = \sqrt{\sum_{n=1}^{N}(Y_g - \hat{f}_g(\hat{\theta}_n^{train}))^2}$

2. Predict $(\theta_m^{test})_{m=1,\ldots,M}$ using the gene expression data $(Y_m)_{m=1,\ldots,M}^{test}$

(a) Choose K discrete and equally-spaced cell times between 0 to 2π. For now, we choose $K = 100$, which is pretty large considering the size of 155 cells in the test sample.

(b) Compute the likelihood of Y_m^{test} at each cell time k:

$$L_m(k) = L(\theta_m = k | Y_m^{test}, \big(\hat{f}_g(\theta_m = k), \hat{\sigma}_g\big)_{g=1,\ldots,G}\big) = \prod_{g=1}^{G} P\big(Y_{gm}^{test} | \hat{f}_g(\theta_m = k), \hat{\sigma}_g\big),$$

where $P\big(Y_{gm}^{test} | \hat{f}_g(\theta_m = k), \hat{\sigma}_g\big) \sim N(\hat{f}_g(\theta_m = k), \hat{\sigma}_g)$

(c) Maximize $L_m(k)$ over $k = 1, \ldots, 100$:

$$\hat{\theta}_m^{test} = \underset{k=1,\ldots,100}{\text{argmax}}\, L_m(k)$$

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) [43] under accession number GSE121265. We also make the processed data available at https://github.com/jhsiao999/peco-paper and https://giladlab.uchicago.edu/wp-content/uploads/2019/02/Hsiao_et_al_2019.tar.gz. All analysis results and all scripts used to produce the work are available at https://jhsiao999.github.io/peco-paper. We implement our method peco in an R package, which is available through Bioconductor. The development version of peco is also available at https://github.com/jhsiao999/peco.

Acknowledgements

We thank members of the Gilad, Stephens and Lynch laboratories for valuable discussions during the preparation of this manuscript. We also thank Natalia Gonzales for constructive feedbacks on this manuscript. This work was funded by NIH grant HG002585 to M.S. and NIH grant GM122930 to Y.G. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

CJH, PYT, YG, and MS conceived of the study, designed the experiments, and formulated the analysis framework. PYT performed the experiments with assistance from JEB. CJH and MS developed the statistical approach for predicting continuous cell cycle phase, and CJH implemented the algorithm. CJH wrote the R package with assistance from KAB and KKD. CJH analyzed the data, with assistance from KB, JDB, PYT, and MS. CJH, MS and YG wrote the original draft with input from PYT, JDB and KAB. All authors reviewed the final manuscript.

References

[1] Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18**, 35–45 (2018).

[2] Macaulay, I. C., Ponting, C. P. & Voet, T. Single-Cell multiomics: Multiple measurements from single cells. *Trends Genet.* **33**, 155–168 (2017).

[3] Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).

[4] Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: Recording the past and predicting the future. *Science* **358**, 69–75 (2017).

[5] Kowalczyk, M. S. *et al.* Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* **25**, 1860–1872 (2015).

[6] Velten, L. *et al.* Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19**, 271–281 (2017).

[7] Lu, Y. *et al.* Systematic analysis of Cell-to-Cell expression variation of T lymphocytes in a human cohort identifies aging and genetic associations. *Immunity* **45**, 1162–1175 (2016).

[8] Stubbington, M. J. T., Rozenblatt-Rosen, O., Regev, A. & Teichmann, S. A. Single-cell transcriptomics to explore the immune system in health and disease. *Science* **358**, 58–63 (2017).

[9] Skelly, D. A. *et al.* Single-Cell transcriptional profiling reveals cellular diversity and intercommunication in the mouse heart. *Cell Rep.* **22**, 600–610 (2018).

[10] Nguyen, Q. H. *et al.* Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.* **9**, 2028 (2018).

[11] Sanchez, A. & Golding, I. Genetic determinants and cellular constraints in noisy gene expression. *Science* **342**, 1188–1193 (2013).

[12] Soltani, M. & Singh, A. Effects of cell-cycle-dependent expression on random fluctuations in protein levels. *R Soc Open Sci* **3**, 160578 (2016).

[13] Keren, L. *et al.* Noise in gene expression is coupled to growth rate. *Genome Res.* **25**, 1893–1902 (2015).

[14] Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).

[15] Barron, M. & Li, J. Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data. *Sci. Rep.* **6**, 33892 (2016).

[16] Chen, M. & Zhou, X. Controlling for confounding effects in single cell RNA sequencing studies using both control and target genes. *Sci. Rep.* **7**, 13587 (2017).

[17] Kolodziejczyk, A. A. *et al.* Single cell RNA-Sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**, 471–485 (2015).

[18] Lauridsen, F. K. B. *et al.* Differences in cell cycle status underlie transcriptional heterogeneity in the HSC compartment. *Cell Rep.* **24**, 766–780 (2018).

[19] Leng, N. *et al.* Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods* **12**, 947–950 (2015).

[20] Povinelli, B. *et al.* Integrated single cell analysis reveals cell cycle and ontogeny related transcriptional heterogeneity in hscs. *Exp. Hematol.* **64**, S95–S96 (2018).

[21] Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411 (2018).

[22] Ingolia, N. T. & Murray, A. W. The ups and downs of modeling the cell cycle. *Curr. Biol.* **14**, R771–7 (2004).

[23] Pauklin, S. & Vallier, L. The Cell-Cycle state of stem cells determines cell fate propensity. *Cell* **156**, 1338 (2014).

[24] Sakaue-Sawano, A. *et al.* Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* **132**, 487–498 (2008).

[25] Banovich, N. E. *et al.* Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Res.* **28**, 122–131 (2018).

[26] Sakaue-Sawano, A. *et al.* Genetically encoded tools for optical dissection of the mammalian cell cycle. *Mol. Cell* **68**, 626–640.e5 (2017).

[27] Tung, P.-Y. *et al.* Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).

[28] Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* (2017).

[29] Kaufman, L. & Rousseeuw, P. J. Innovation and intellectual property rights. In Kaufman, L. & Rousseeuw, P. J. (eds.) *Finding Groups in Data: An Introduction to Cluster Analysis*, chap. 2, 68–125 (John Wiley Sons, Inc., Hoboken, 1990).

[30] Tibshirani, R. J. Adaptive piecewise polynomial estimation via trend filtering. *Ann. Stat.* **42**, 285–323 (2014).

[31] Whitfield, M. L. *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002).

[32] Scialdone, A. *et al.* Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).

[33] Liu, Z. *et al.* Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat. Commun.* **8**, 22 (2017).

[34] Spellman, P. T. *et al.* Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *MBoC* **9**, 3273–3297 (1998).

[35] Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**, e117 (2016).

[36] Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).

[37] Dey, K. K., Hsiao, C. J. & Stephens, M. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genet.* **13**, e1006599 (2017).

[38] Liao, Y., Smyth, G. K. & Shi, W. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108 (2013).

[39] Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).

[40] Liao, Y., Smyth, G. K. & Shi, W. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

[41] Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).

[42] Pau, G., Fuchs, F., Sklyar, O., Boutros, M. & Huber, W. EBImage–an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26**, 979–981 (2010).

[43] Edgar, R., Domrachev, M. & Lash, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research* **30**, 207–210 (2002).

[44] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. *cluster: Cluster Analysis Basics and Extensions* (2019). R package version 2.1.0 — For new features, see the 'Changelog' file (in the package source).

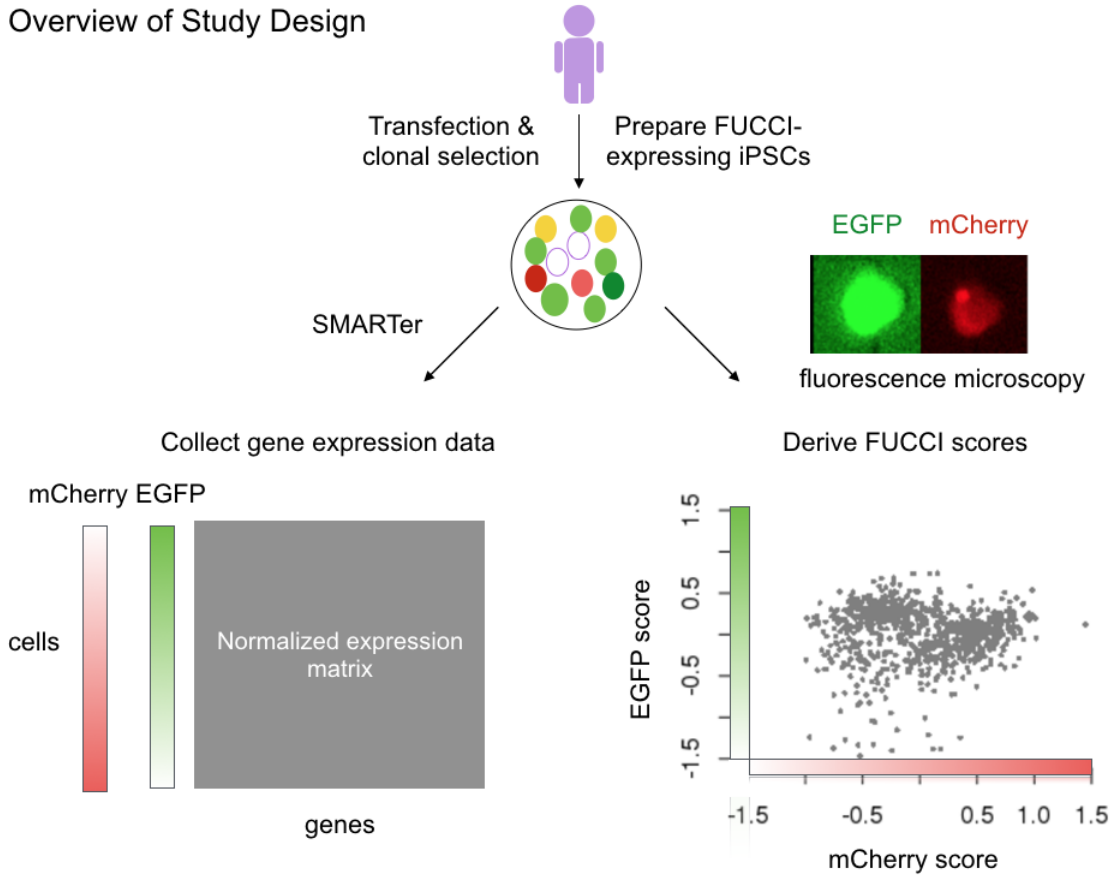


Figure 1: Overview of study design. We collected two types of data from the same single cells using FUCCI-expressing iPSCs: in situ fluorescence images and scRNA-seq. After quality control, we obtained 888 single cells for which we had high quality RNA-seq data. We computed two FUCCI scores for each cell by individually summing the EGFP (green) and mCherry (red) intensities in a fixed cell area (100 x 100 px), correcting for background noise outside the defined cell area, and then taking the log10 transformation of the sum of corrected intensities. In the bottom-right scatter plot, we show the FUCCI scores for the 888 high quality single-cell samples, i.e., mCherry and EGFP log10 sum intensities after background noise correction. Finally, we standardized the molecule counts to counts per million (CPM) and transformed the data per gene to a standard normal distribution.

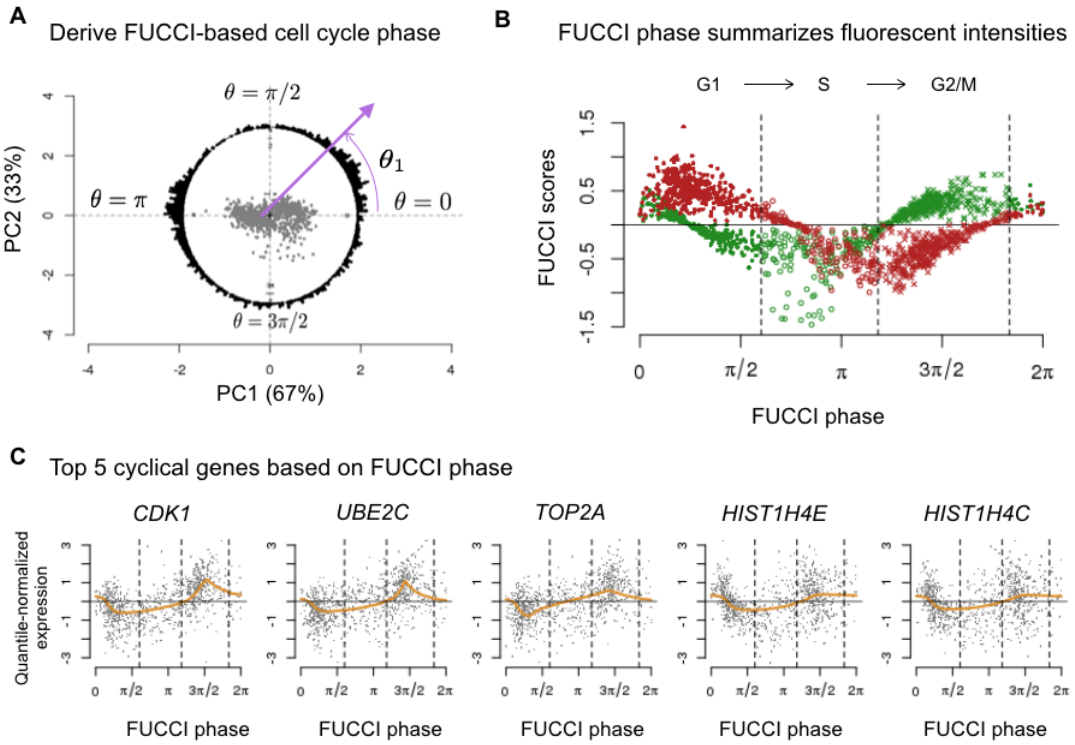


Figure 2: Characterizing cell cycle phase using FUCCI fluorescence intensities. (A) We inferred FUCCI phase (angles in a circle) based on FUCCI scores of EGFP and mCherry. The points in center correspond to PC scores based on EGFP and mCherry scores, and the circle histogram shows the corresponding FUCCI phase distribution. For example, we inferred θ_1 based on the PC scores derived from the cell's FUCCI scores. (B) We ordered FUCCI scores of EGFP and mCherry by FUCCI phase to visualize the co-oscillation of EGFP and mCherry along the cell cycle. Red and green points correspond to EGFP and mCherry scores, respectively. The vertical lines correspond to phase boundaries derived from the PAM-based classification (G1 384 cells, S 172 cells, G2/M 332 cells). (C) Given FUCCI phase, we ordered cells along the cell cycle to estimate the cyclic trend of gene expression levels for each gene. We identified these 5 genes as the top 5 cyclic genes in the data: *CDK1*, *UBE2C*, *TOP2A*, *HIST1H4E*, and *HIST1H4C*. Each plot shows the expression levels of 888 single-cell samples and the estimated cyclic trend (orange line). All 5 genes were previously identified as related to cell cycle regulation. The vertical lines correspond to phase boundaries derived from the PAM-based classification.

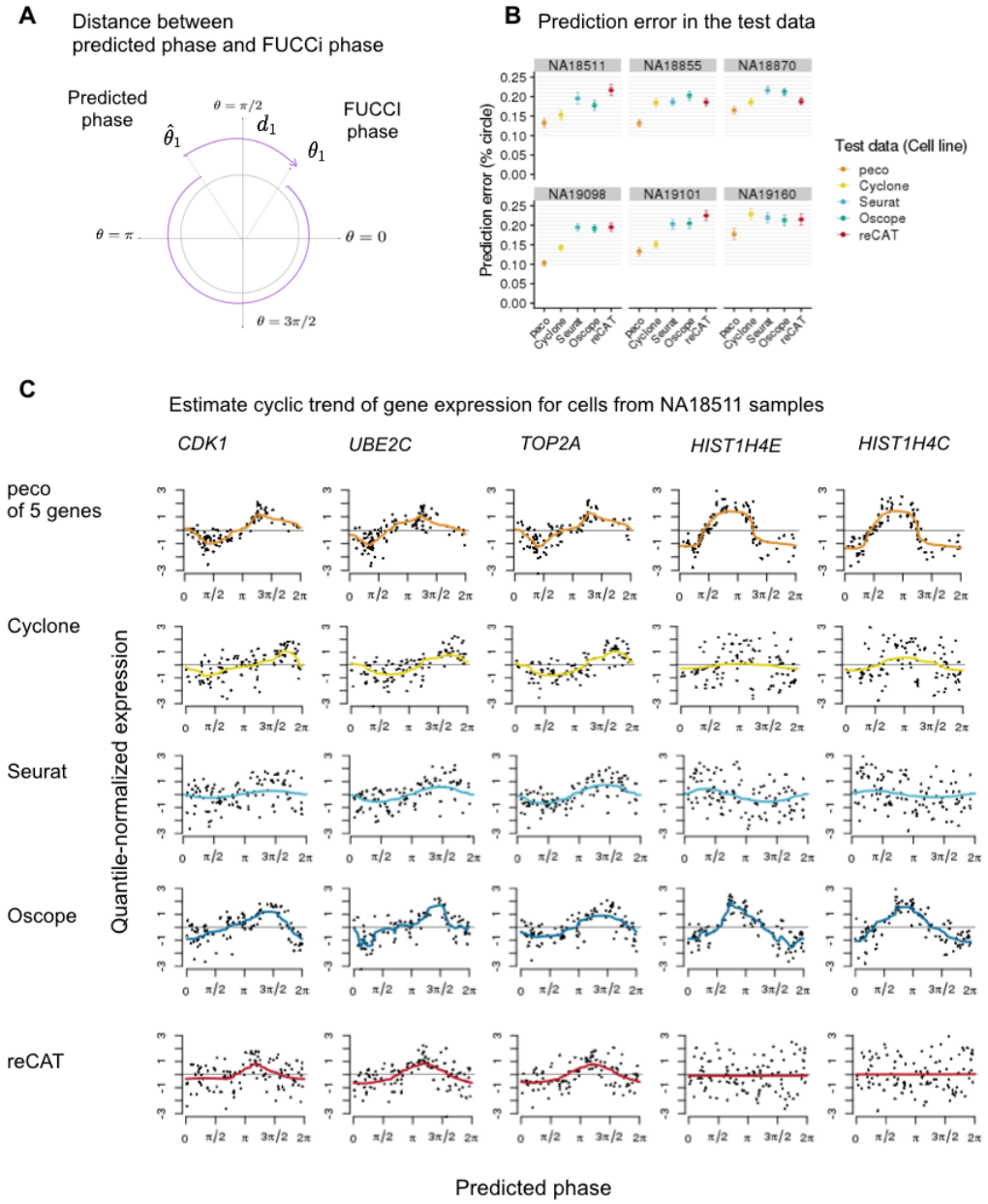
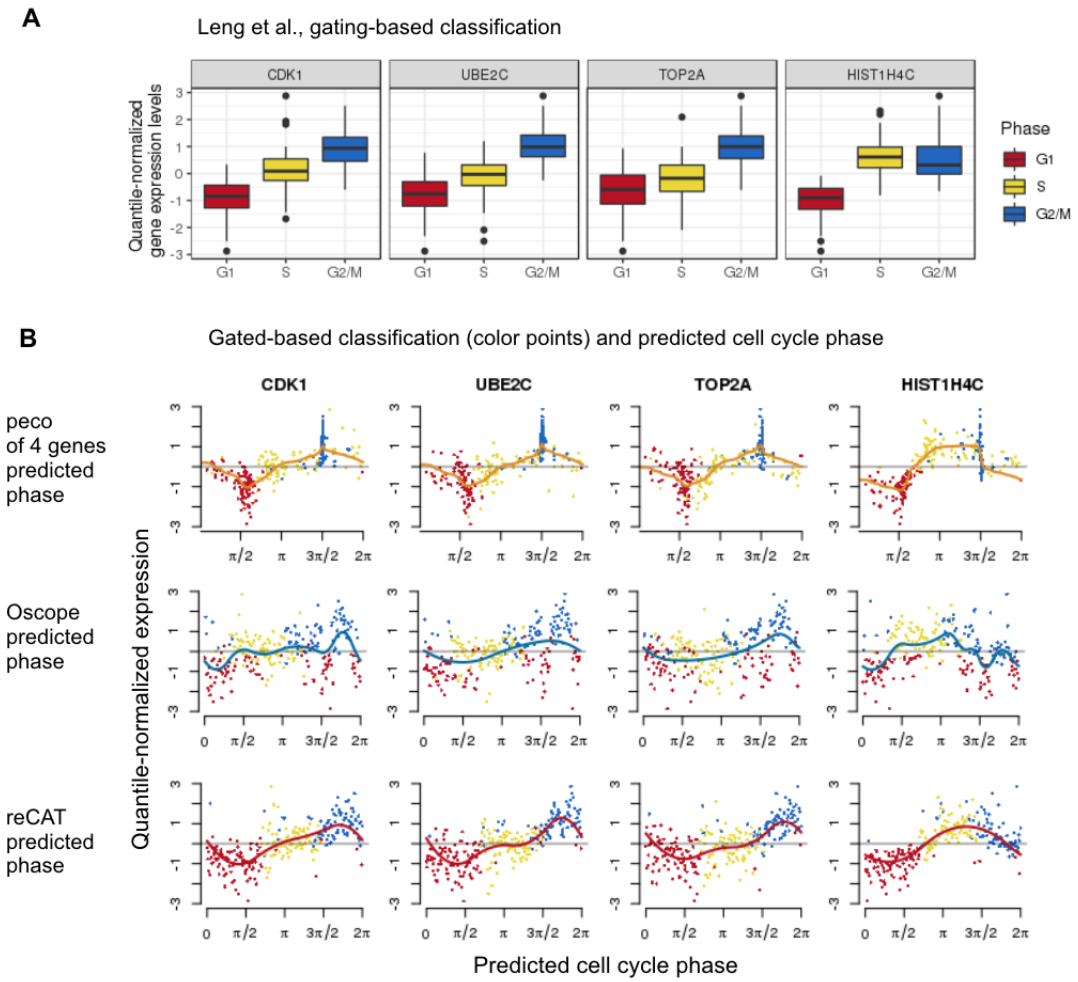


Figure 3: Inferring cell cycle phase from scRNA-seq data. (A) We defined prediction error as the distance between predicted phase and FUCCI phase (as a percentage of the entire cycle, 2π). (B) We applied six-fold cross-validation to test the performance of our predictor. The six panels correspond to performances in the six folds. Each panel compares the mean prediction error of FUCCI phase in the test data (error bars correspond to standard error) using our method and existing tools (Seurat [21], Cyclone [32], reCAT [33], and Oscope [19]). (C) Estimated cyclic trend of top 5 cyclic genes for samples from individual cell line NA18511. Rows correspond to the results of the five methods. For example, we show the results of peco predicted phase in the first row. We ordered the samples according to the predicted phase and used *trendfilter* to estimate cyclic trend of gene expression. The colored line corresponds to the estimated cyclic expression level along the predicted phase.



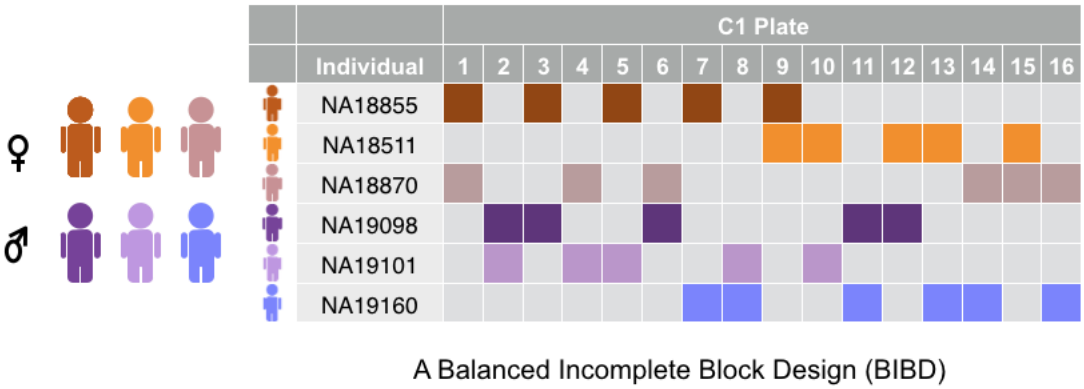
C

Gating-based classification vs Seurat/Cyclone-based classification

Gating based classification	Seurat classification			Cyclone classification		
	G1	S	G2M	G1	S	G2M
G1	10	28	53	91	0	0
S	20	60	0	1	78	2
G2M	11	5	60	0	0	76

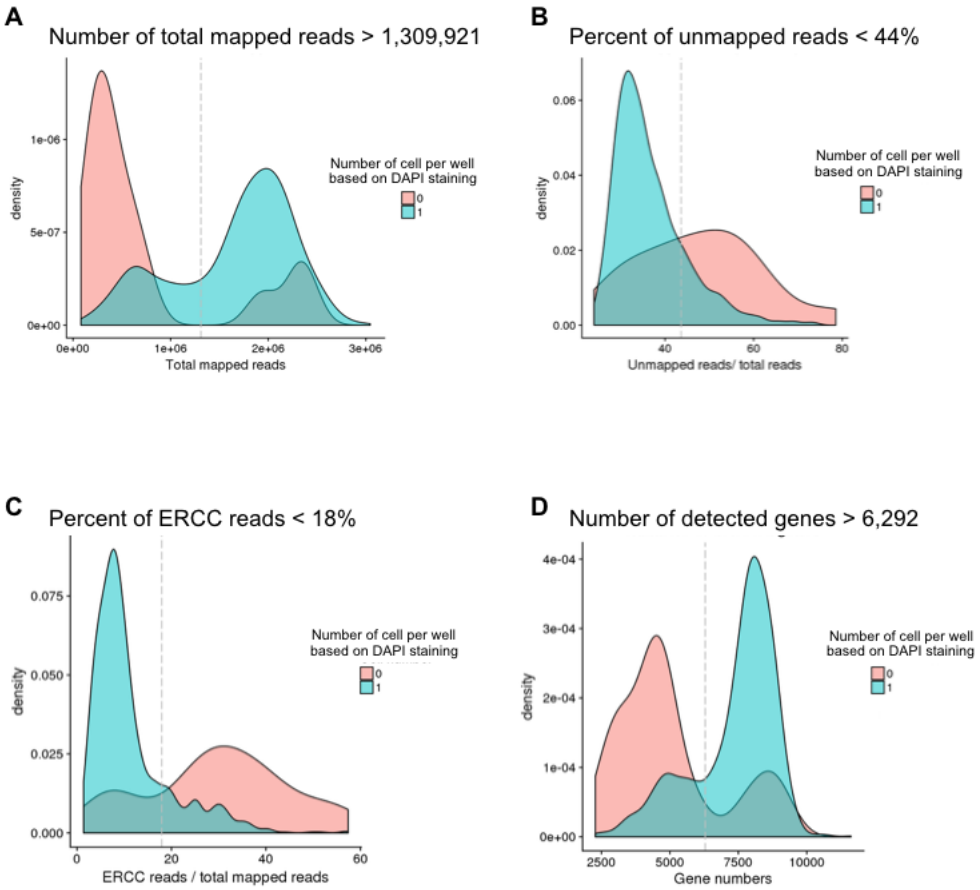
Figure 4: Applying peco and existing tools to Leng et al. data [19]. The single-cell samples in this data were sorted into G1, S and G2M phase. (A) We plot the distribution of gene expression for the top 4 cyclic genes per cell cycle phase. (B) We compare predicted phases based on peco, Oscope and reCAT. Rows correspond to prediction results based on the three methods. Specifically, we sort the single-cell samples according to the predicted phase, and color the sample points according to the gated phase. For example, in the first row, we show that the peak expression profile of peco prediction is consistent with results based on gating. The orange line corresponds to the cyclic trend of expression levels. (C) We compare the phase assignment based on gating with Seurat/Cyclone-based classification.

Supplemental Figure S1



Supplemental Fig. S1: C1 study design. The table displays the distribution of cells from six individual cell lines across sixteen C1 96-well plates, with rows corresponding to cell lines and columns corresponding to C1 plates. Specifically, we used a balanced incomplete block design (BIBD) in which cells from unique pairs of individuals were distributed across fifteen 96-well C1 plates on the C1 platform. We also included data from one additional plate (containing individuals NA18855 and NA18511), which we collected as part of a pilot study. In total, we collected data from 1,536 scRNA-seq samples distributed across sixteen C1 plates.

Supplemental Figure S2



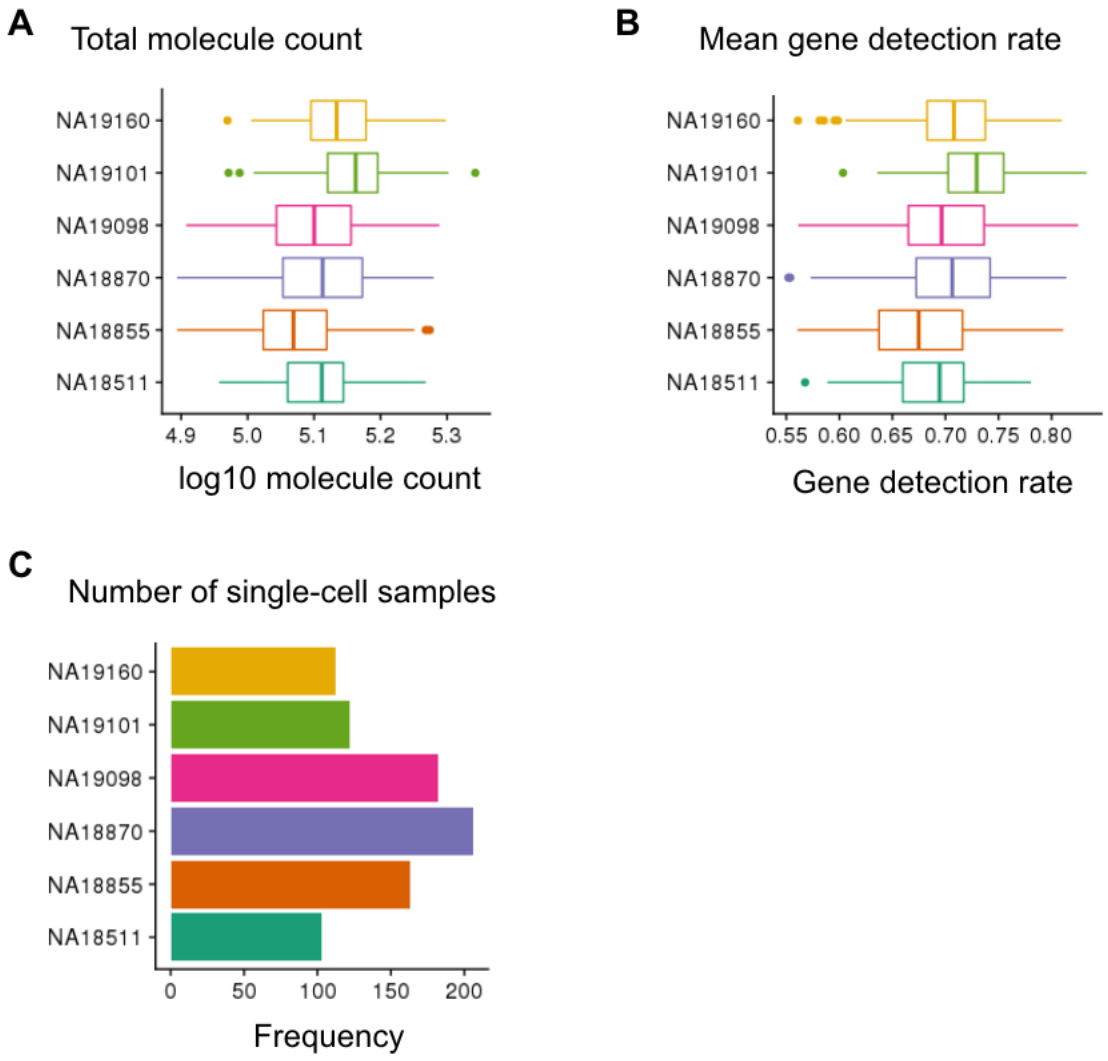
Supplemental Fig. S2: Filtering criteria for including single-cell samples. We used DAPI to determine the number of cells captured in each C1 well and compared common scRNA-seq data metrics between empty wells and single-cell wells to determine filtering criteria for single-cell samples. Using this approach, we determined filtering criteria for (A) the number of total mapped reads (>= 1,309,921), (B) the percentage of unmapped reads (< 44%), (C) the percentage of ERCC reads (< 18%), and (D) the number of detected genes (>= 6,292 genes with least one read).

Supplemental Figure S3

	Mapped Reads (million)	Unmapped Reads Proportion (%)	Number of singleton samples	ERCC reads proportion (%)	Molecule count (million)
Overall	2.0 (0.30)	33.0 (4.12)	888	7.6 (3.15)	.13 (.024)
NA18511	1.9 (0.25)	35.0 (4.31)	103	9.0 (2.43)	.13 (.019)
NA18855	1.9 (0.31)	34.9 (4.17)	163	8.2 (3.58)	.12 (.022)
NA18870	2.0 (0.30)	32.5 (4.30)	206	7.4 (3.47)	.13 (.023)
NA19098	2.1 (0.32)	31.3 (2.95)	182	6.8 (2.25)	.13 (.024)
NA19101	2.0 (0.29)	33.3 (4.27)	122	5.9 (3.27)	.15 (.022)
NA19160	2.0 (0.27)	32.0 (3.14)	112	8.8 (2.33)	.14 (.022)

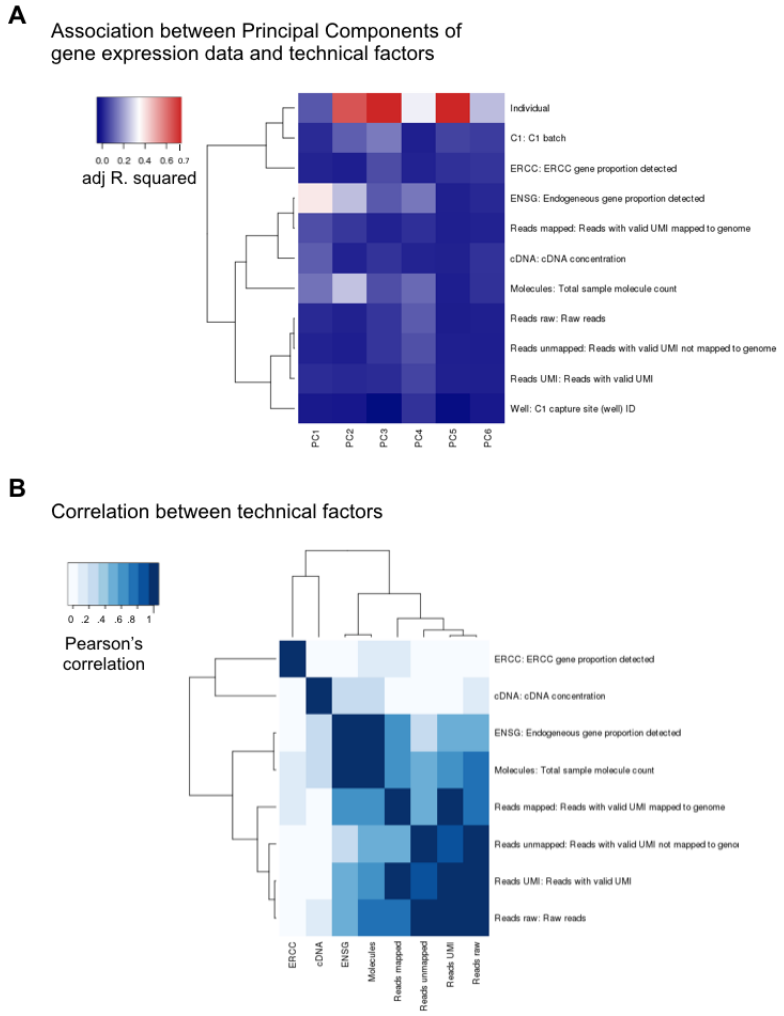
Supplemental Fig. S3: Summary table of scRNA-seq quality metrics. We computed means and standard deviations (in parentheses) of these metrics across single-cell samples for each of the six cell lines and also for the entire dataset.

Supplemental Figure S4



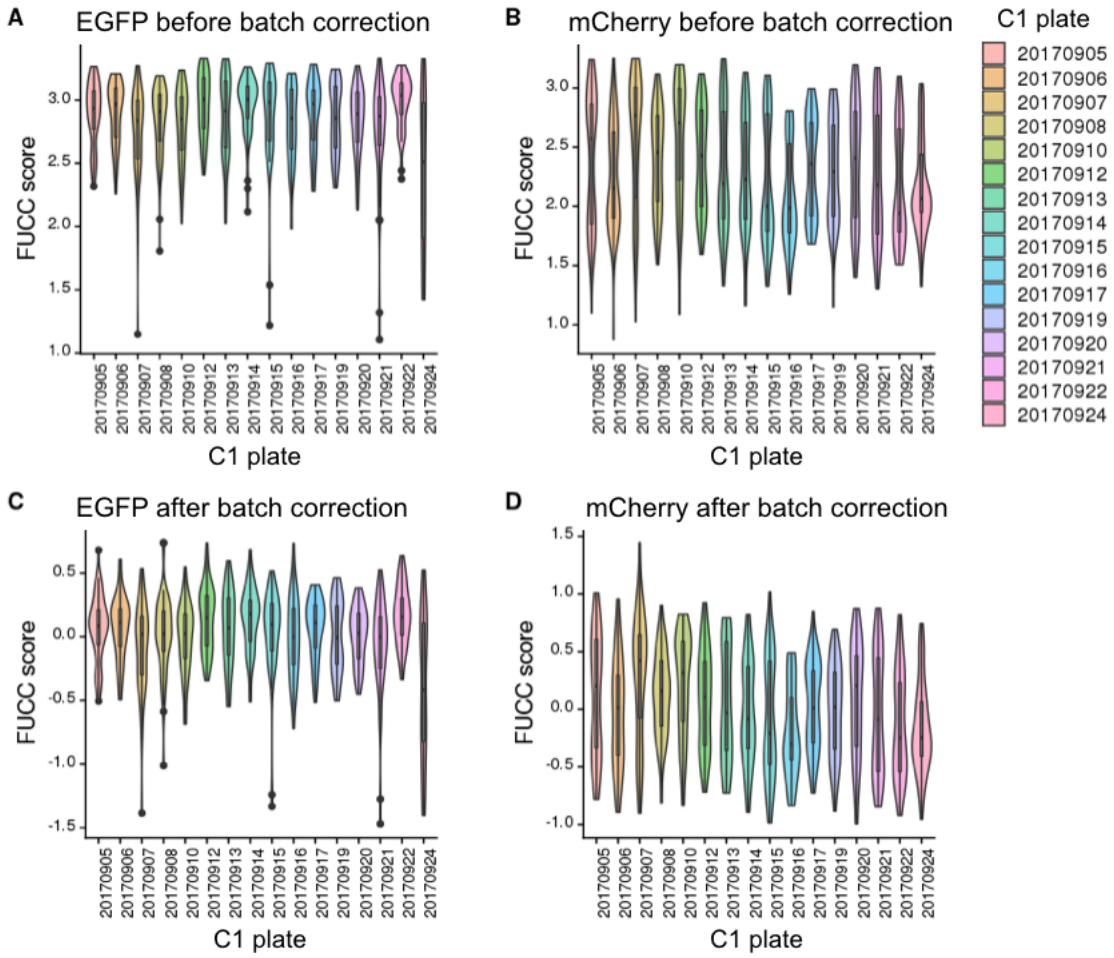
Supplemental Fig. S4: Distribution of scRNA-seq quality metrics for the six cell lines. We show the distribution of single-cell samples in (A) the total molecule count, (B) the mean gene detection rate (i.e., fraction of genes with at least one read), (C) the number of single-cell samples.

Supplemental Figure S5



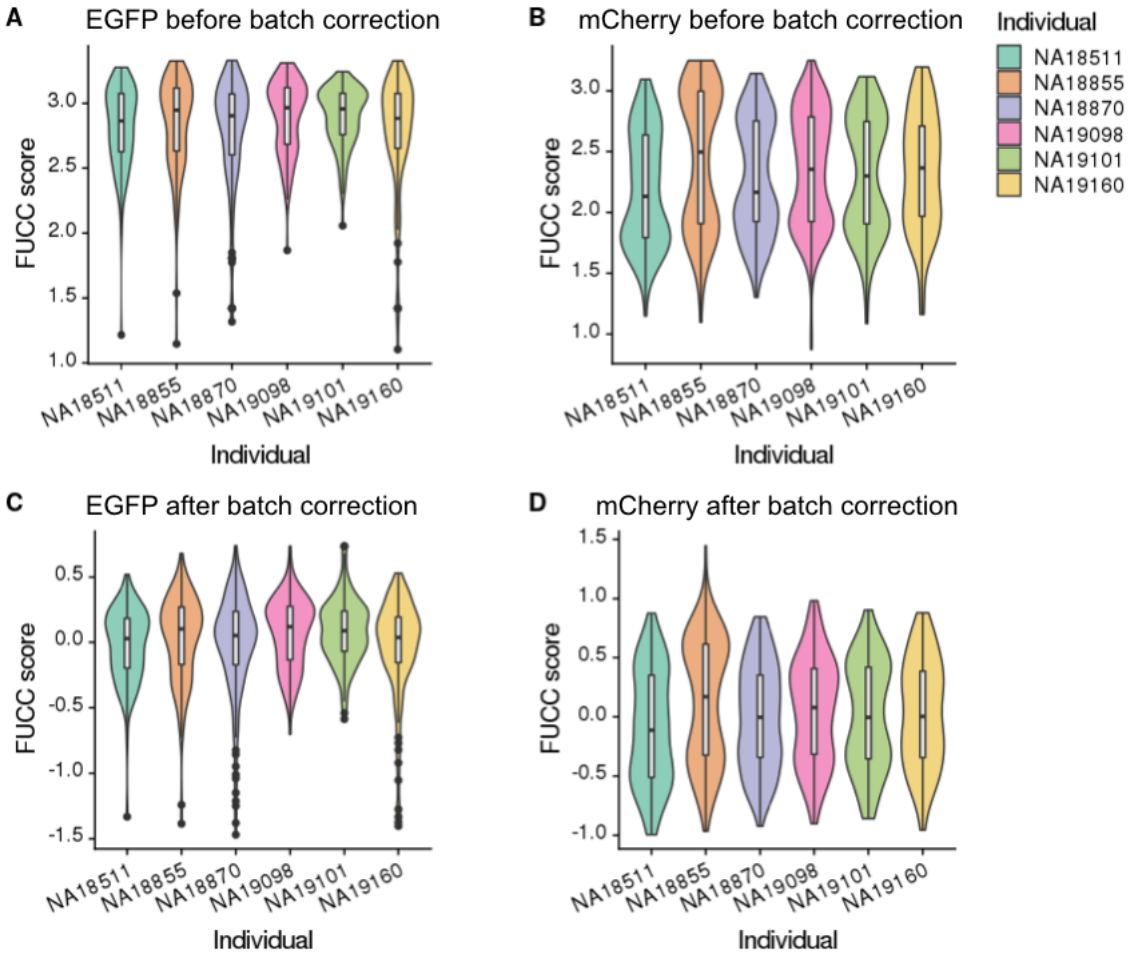
Supplemental Fig. S5: Major sources of variation in our gene expression data of 888 quality samples and 11,040 genes. (A) Principal Component Analysis (PCA) was applied to the $\log_2$ CPM of the gene expression data. We computed the proportion of variance explained (i.e., adjusted R-squared) in each of the principal components by: individual identity of the single-cell sample (Individual), C1 processing batch (C1), capture site or well (Well), fraction of ERCC genes detected (ERCC), fraction of endogenous genes detected (ENSG), cDNA concentration (cDNA), sample total molecule count (Molecules), number of raw reads (Reads raw), number of raw reads with valid UMI (Reads UMI), number of reads with valid UMI mapped to the genome (Reads mapped), and number of reads with valid UMI not mapped to the genome (Reads unmapped). (B) Pearson's correlation between technical factors that are known to influence sample variation in gene expression data.

Supplemental Figure S6



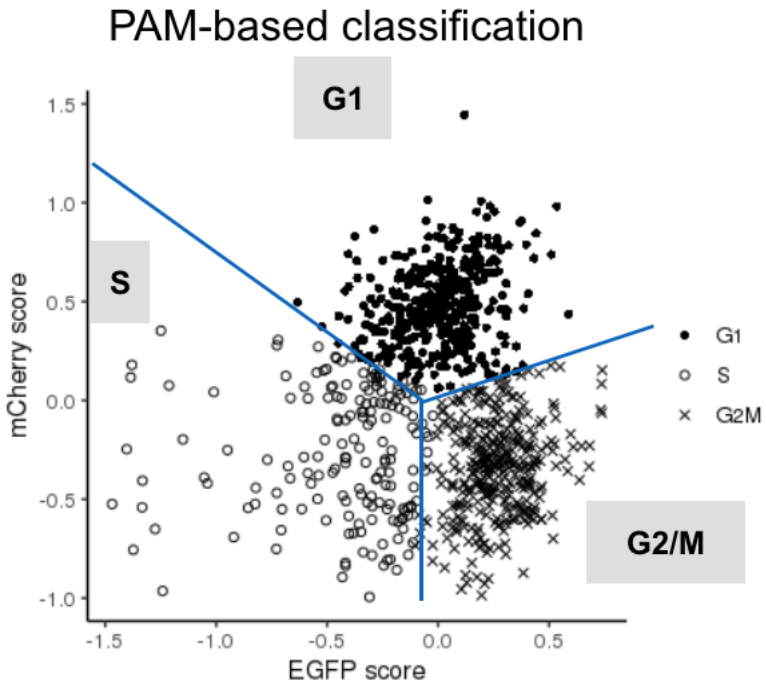
Supplemental Fig. S6: FUCCI scores for the sixteen C1 plates before and after correcting for C1 plate effect. We computed two FUCCI scores - log10 sum of fluorescence intensity in the predefined cell area after background noise correction - corresponding to EGFP and mCherry intensities. (A) and (B) show FUCCI scores before correcting for C1 plate effect. (C) and (D) show FUCCI scores after correcting for C1 plate effect. We found mean FUCCI scores to be significantly different between plates and applied a linear model to account for plate effects on FUCCI scores without removing individual effects.

Supplemental Figure S7



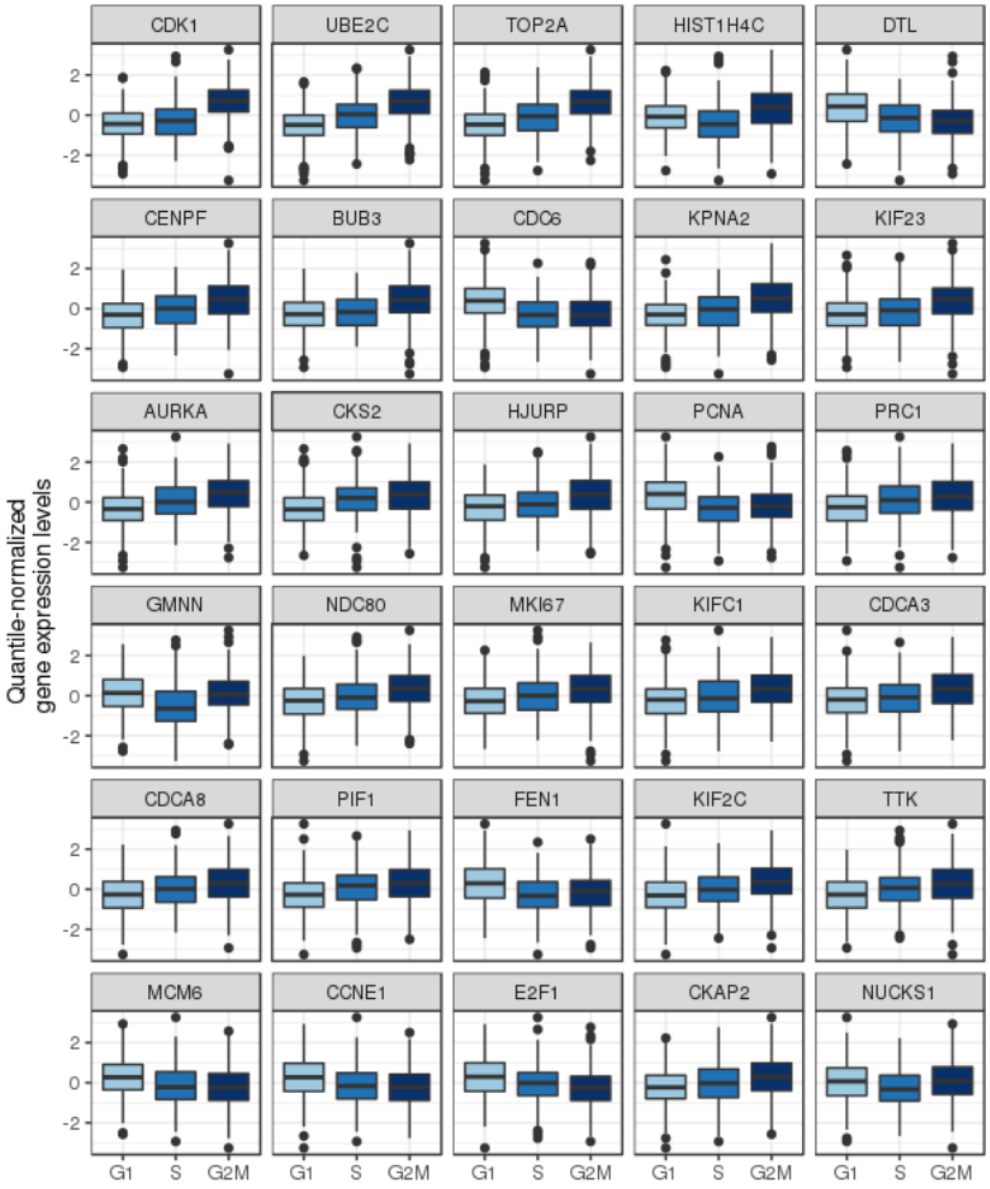
Supplemental Fig. S7: FUCCI scores for the six cell lines before and after correcting for C1 plate effect. We computed two FUCCI scores - log10 sum of fluorescence intensity in the predefined cell area after background noise correction - corresponding to EGFP and mCherry intensities. (A) and (B) show FUCCI scores before correcting for C1 plate effect. (C) and (D) show FUCCI scores after correcting for C1 plate effect. We found mean FUCCI scores to be significantly different between plates and applied a linear model to account for plate effects on FUCCI scores without removing individual effects.

Supplemental Figure S8

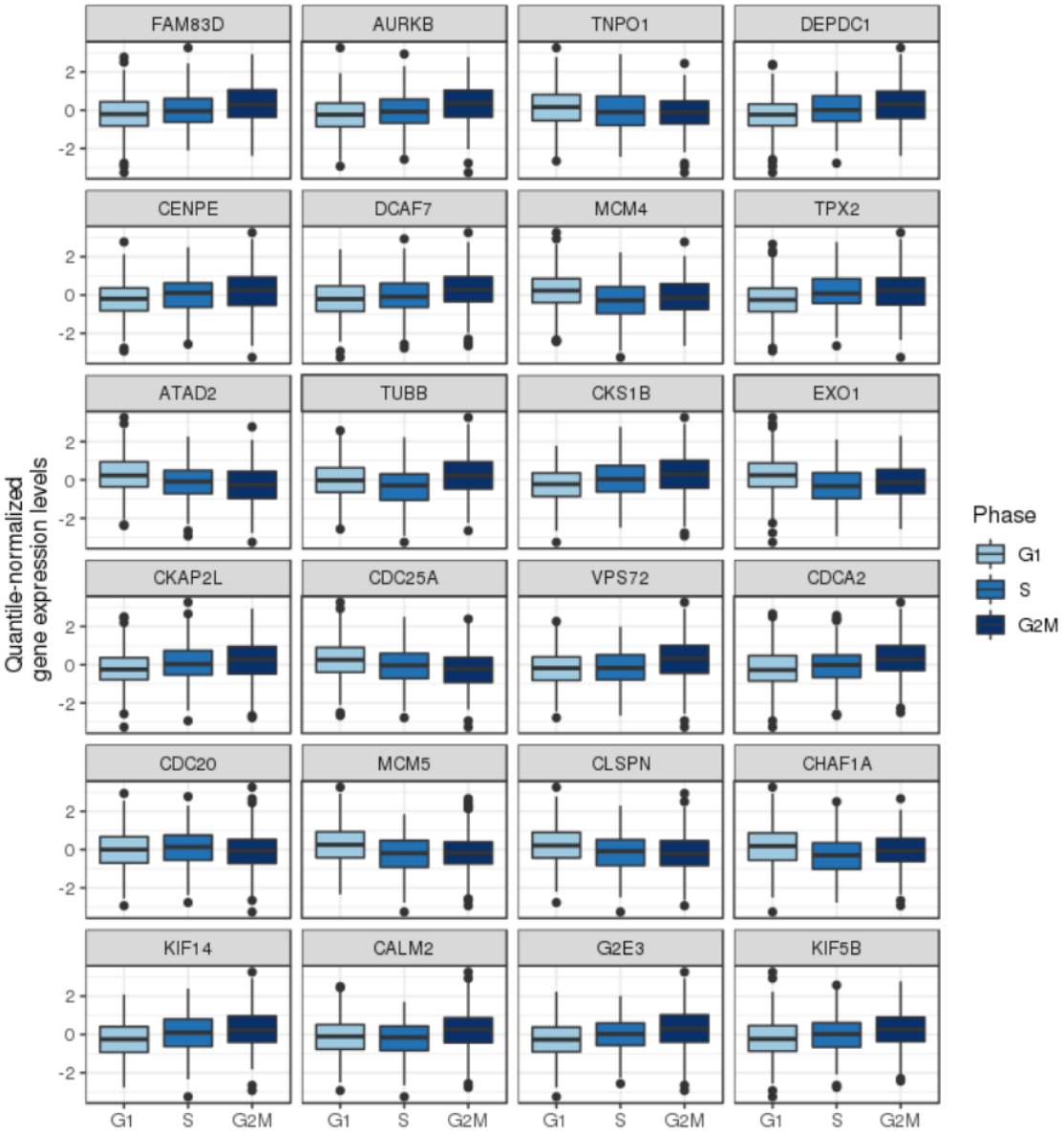


Supplemental Fig. S8: Classification obtained from the PAM-based method. We applied Partition Around Medoids (PAM) to FUCCI scores to cluster the 888 single-cell samples into G1, S, or G2/M phase (384, 172 332 cells in each phase, respectively), using *pam* function in the R package *clust* [44]. X- and Y-axis correspond to EGFP and mCherry score, respectively. The blue lines represent phase boundaries obtained from the PAM method.

Supplemental Figure S9A

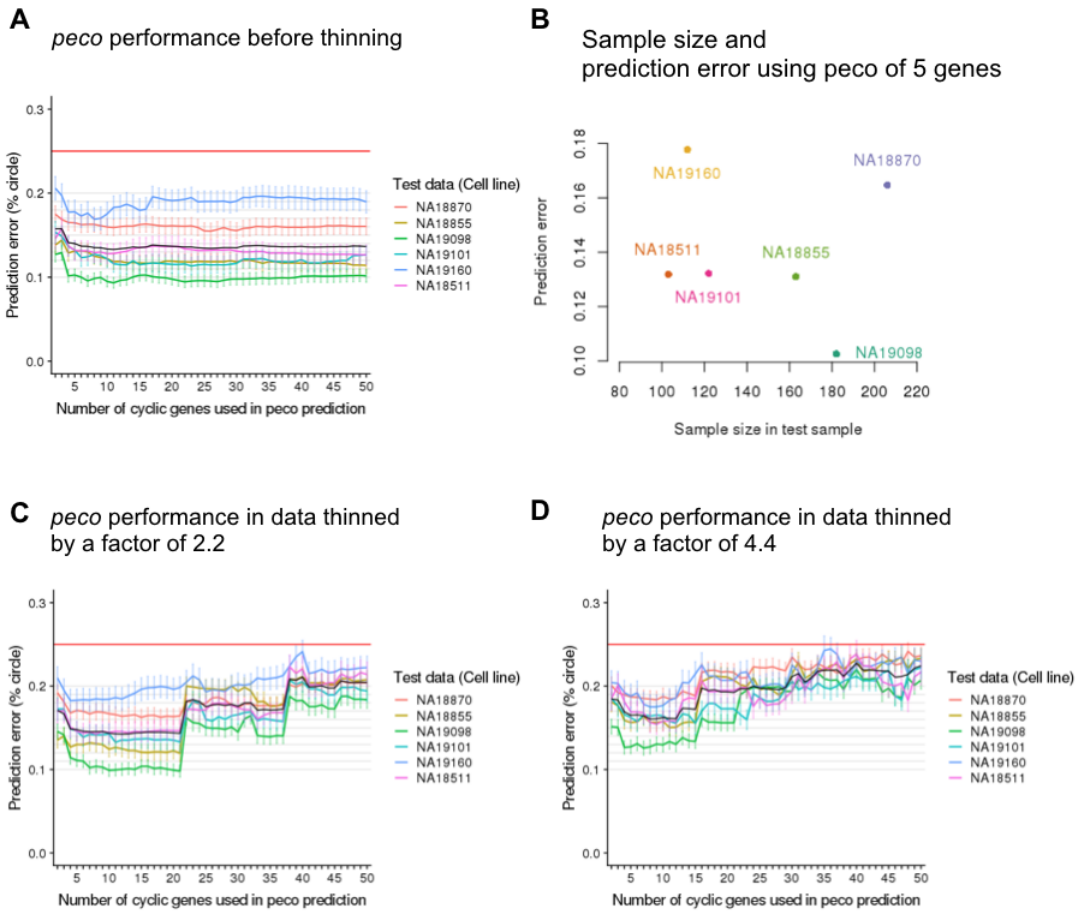


Supplemental Fig. S9A: We identified 54 significant cyclic genes that are also known to be cell cycle genes in Whitfield et al. [31]. This figure shows the top 30 of the 54 genes (from top row, left to right). We applied the PAM-based method to FUCCI scores to classify the 888 single-cell samples to G1, S, or G2/M phase (384, 172, and 332 cells in each phase, respectively). In each boxplot, we plot the distribution of expression levels of the single-cell samples in each phase. Using Analysis of Variance, we found significant differences between phases in average gene expression levels for most genes (ANOVA, P-value $< .01$ for 53 genes and and P-value $= .27$ for *CDC6*).



Supplemental Fig. S9B: We identified 54 significant cyclic genes that are also known to be cell cycle genes in Whitfield et al. [31]. This figure shows the bottom 24 of the 54 genes (from top row, left to right). We applied the PAM-based method to FUCCI scores to classify the 888 single-cell samples to G1, S, or G2/M phase (384, 172, and 332 cells in each phase, respectively). In each boxplot, we plot the distribution of expression levels of the single-cell samples in each phase. Using Analysis of Variance, we found significant differences between phases in average gene expression levels for most genes (ANOVA, P-value $< .01$ for 53 genes and and P-value $= .27$ for *CDC6*).

Supplemental Figure S10



Supplemental Fig. S10: Performance of peco in unthinned and thinned data. We applied six-fold cross-validation. In each fold, we trained our predictor on cells from five individuals and tested its performance on cells from the remaining individual. In panel (A), (C), (D), Y-axis corresponds to prediction error (between 0 to 25%, or $\pi/4$), and X-axis corresponds to the number of top cyclic genes used in the predictor. The six lines correspond to performances in the six folds, specifically average prediction error among cells in the test samples, and error bars correspond to standard errors. (A) The performance of our predictor built between 5 to 50 genes in unthinned data. In (C) and (D), we repeated the analysis in (A) after thinning the test data (total sample molecule count in the un-thinned data was 56,724 $\pm$ 12,762) by a factor of 2.2 (total sample molecule count 25,581 $\pm$ 15,220) and 4.4 (total sample molecule count 13,651$\pm$13,577. (C) and (D) show the performance of our predictor in data thinned by a factor of 2.2 and 4.4, respectively. (B) shows that number of cells was not correlated with prediction error of FUCCI phase using our predictor of 5 genes.

Supplemental Figure S11

A

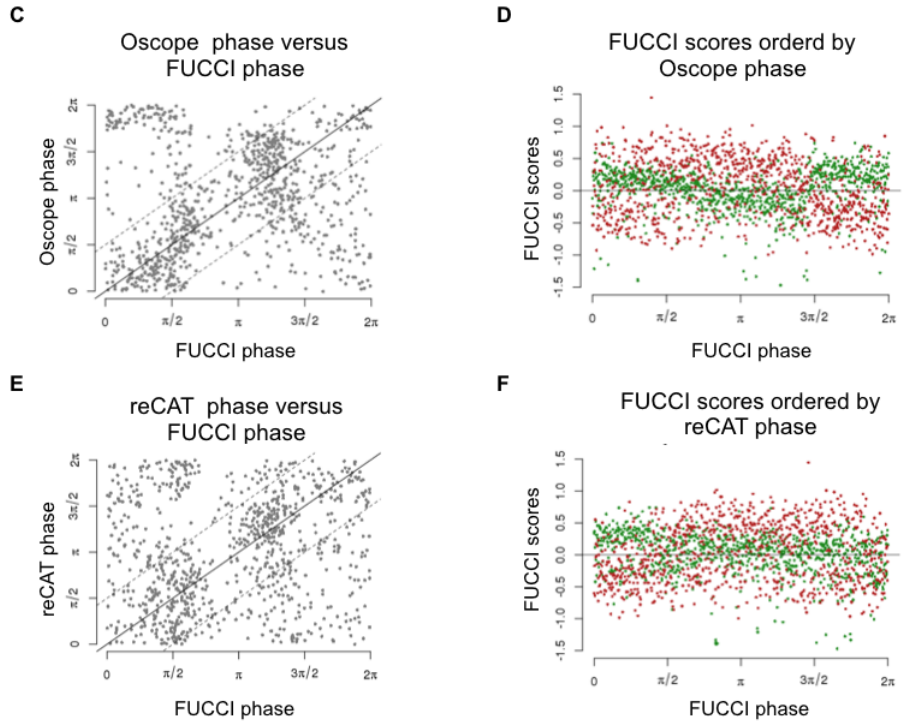
PAM based discrete classification

Cyclone discrete phase	G1	S	G2M
G1	250	76	44
S	84	21	58
G2M	50	75	230
% Misclass-ified	34%	88%	31%

B

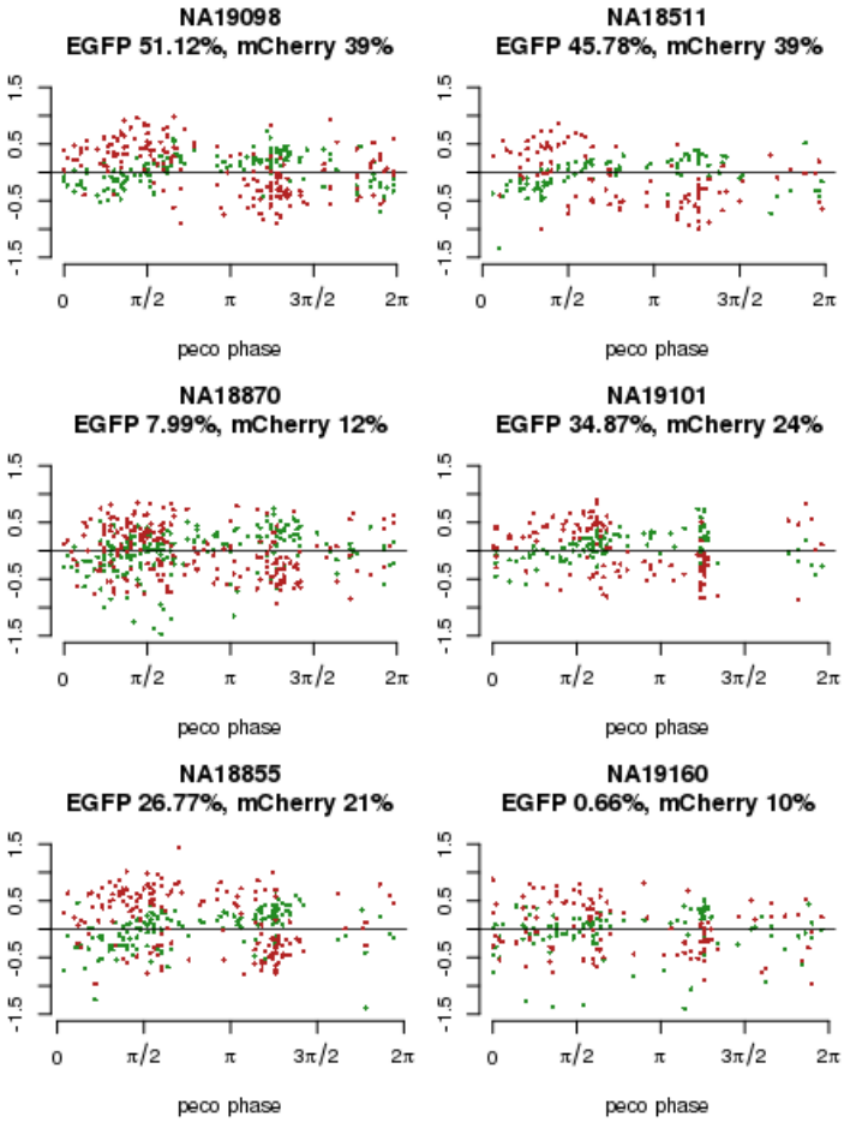
PAM based discrete classification

Seurat discrete phase	G1	S	G2M
G1	86	36	68
S	238	44	76
G2M	60	92	188
% Misclass-ified	78%	74%	43%



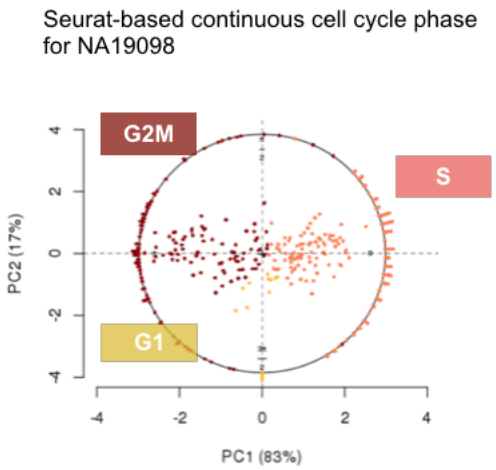
Supplemental Fig. S11: Comparison of FUCCI phase with phase assignment of existing tools. In (A) and (B), classification obtained from PAM is compared with Seurat/Cyclone-based classification. In (C) and (E), we plot FUCCI phase (X-axis) against continuous phase (Y-axis) based on Oscope and reCAT, respectively. In (D) and (F), we order FUCCI scores by Oscope/Cyclone based phase, respectively. Red and green points represent FUCCI scores of EGFP and mCherry, respectively.

Supplemental Figure S12



Supplemental Fig. S12: FUCCI scores and inferred phase from peco in individual cell lines. In (A) to (F), we order FUCCI scores by inferred phase from peco in each individual cell line (using our cross-validation results). We also estimated proportion of variance explained by inferred phase from peco in EGFP and mCherry scores (as shown on top of each plot). Red and green points represent FUCCI scores of EGFP and mCherry, respectively.

Supplemental Figure S13



Supplemental Fig. S13: Continuous cell cycle phase assignment based on the two Seurat phase-specific scores for samples from cell line NA19098. Seurat uses the mean expression levels of 43 S-phase marker genes and 54 G2/M phase genes to compute two phase-specific scores for each cell. We applied PCA to transform the two phase-specific scores to PC scores. X- and Y-axis correspond to PC1 and PC2 score. The dots inside the circle correspond to the cells and their PC scores. We transformed these scores to angles on the unit circle (see Methods for details). The colors correspond to Seurat-based G1, S, G2/M phase assignment.

Supplemental Figure S14

Comparison of prediction error on our data.

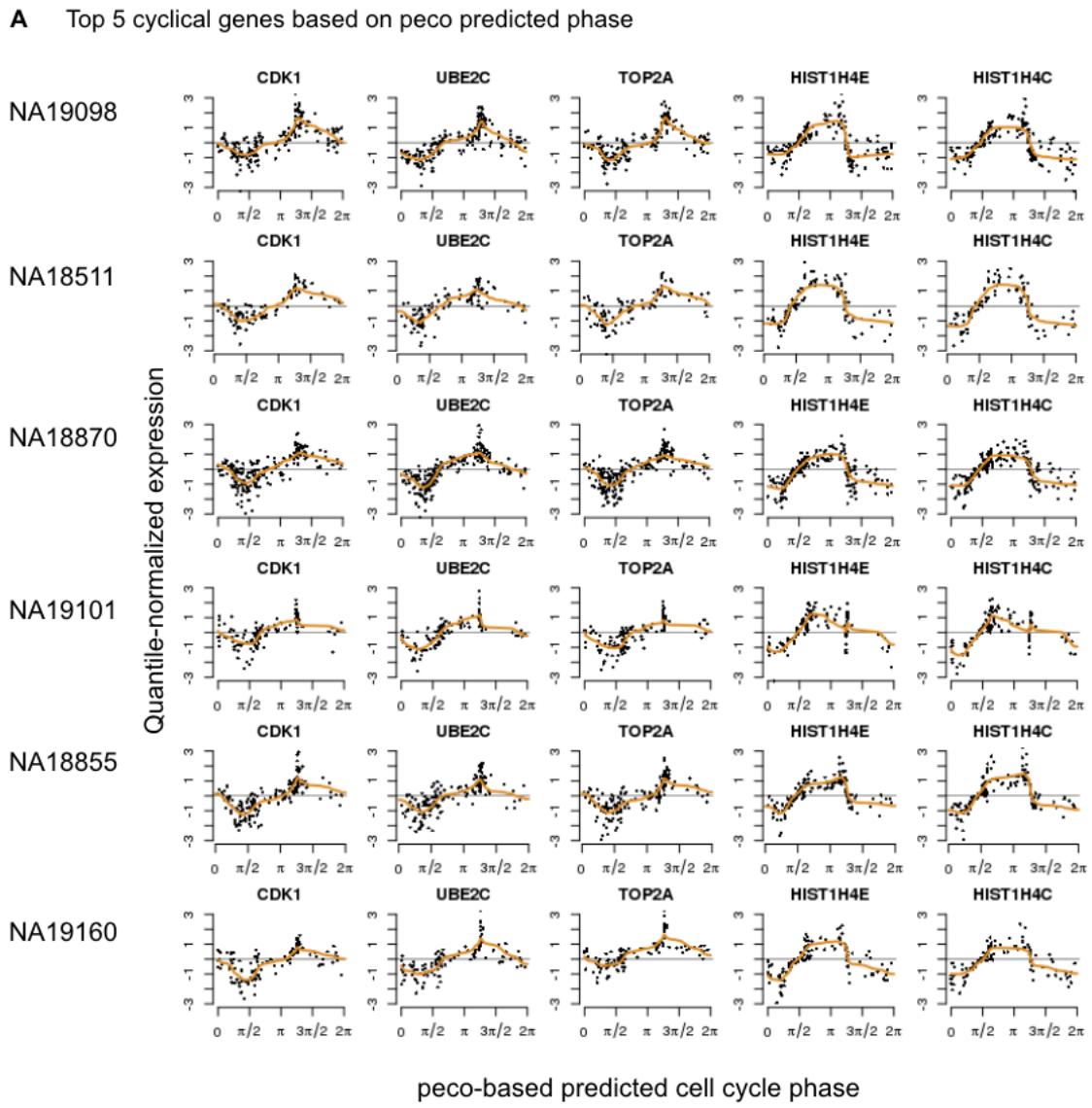
	peco	Cyclone	Seurat	Oscope	reCAT
NA18511	10.3(.72)	15.2(1.27)	19.6(1.51)*	20.4(1.39)*	21.4(1.39)*
NA18855	13.2(1.21)	18.4(1.02)*	18.6(.01)*	15.9(.99)*	21.5(1.08)*
NA18870	16.5(.96)	18.6(.98)	21.7(.99)*	22.9(1.03)*	23.9(1.01)*
NA19098	13.2(1.09)	14.3(.79)*	19.4(.96)*	24.5(1.04)*	20.9(.96)*
NA19101	13.1(.97)	15.1(1.03)	20.2(1.28)*	18.8(1.17)*	23.3(1.34)*
NA19160	17.8(1.44)	22.9(1.41)*	21.9(1.35)*	24.1(1.35)*	24.3(1.37)*

Note. Reported in each cell are percent of a unit cirlce with standard error sin parentheses

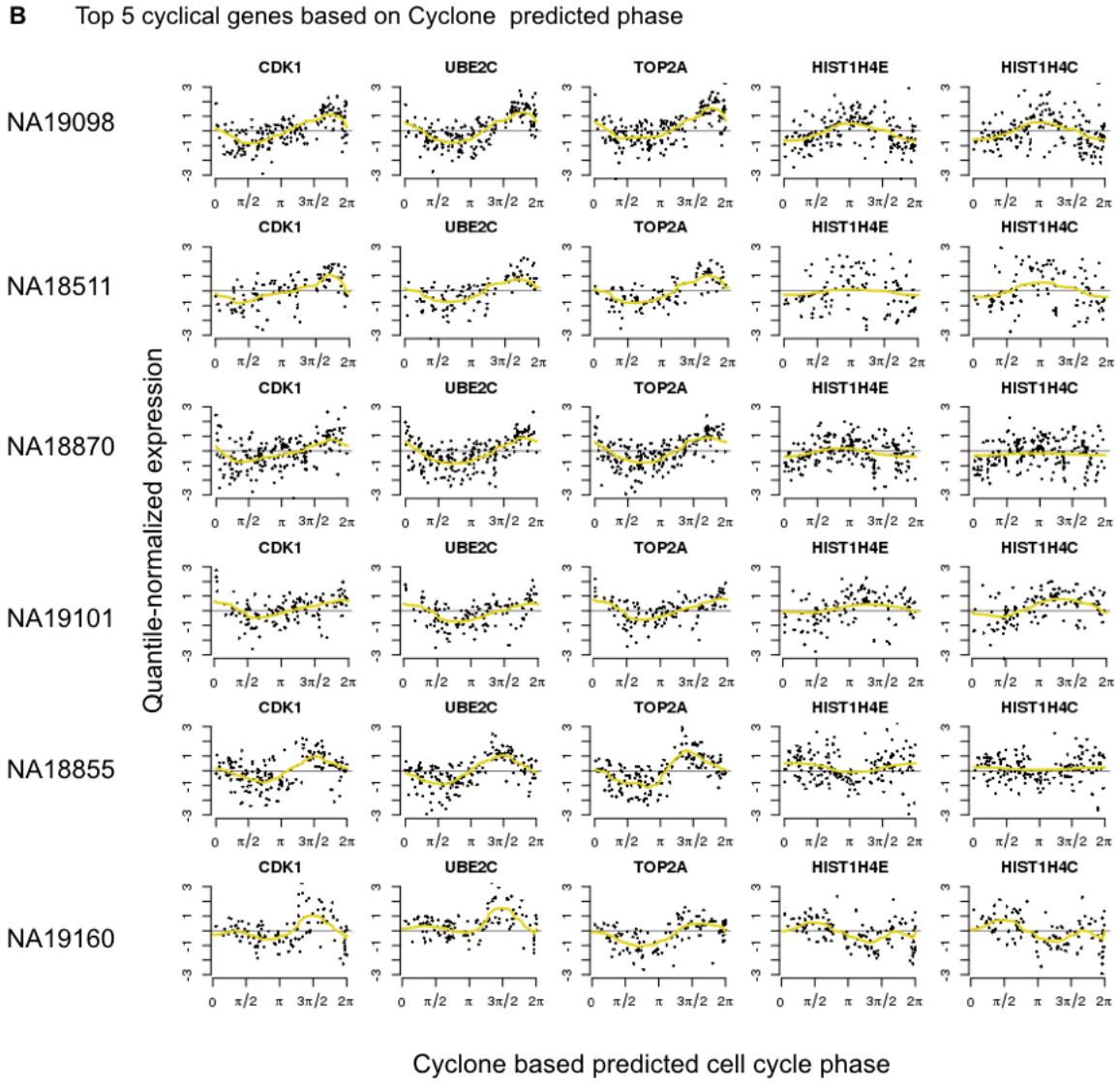
*We compared prediction error between methods and peco using Wilcoxon test and determined signficance at P-value < .05

Supplemental Fig. S14: Comparison of prediction error on our data in cross-validation. We performed Wilcoxon test to compare prediction error between peco and each method in each test data set (samples from an individual cell line). We report prediction error as percentage of the unit circle, along with standard error associated with mean prediciton error in parentheses.

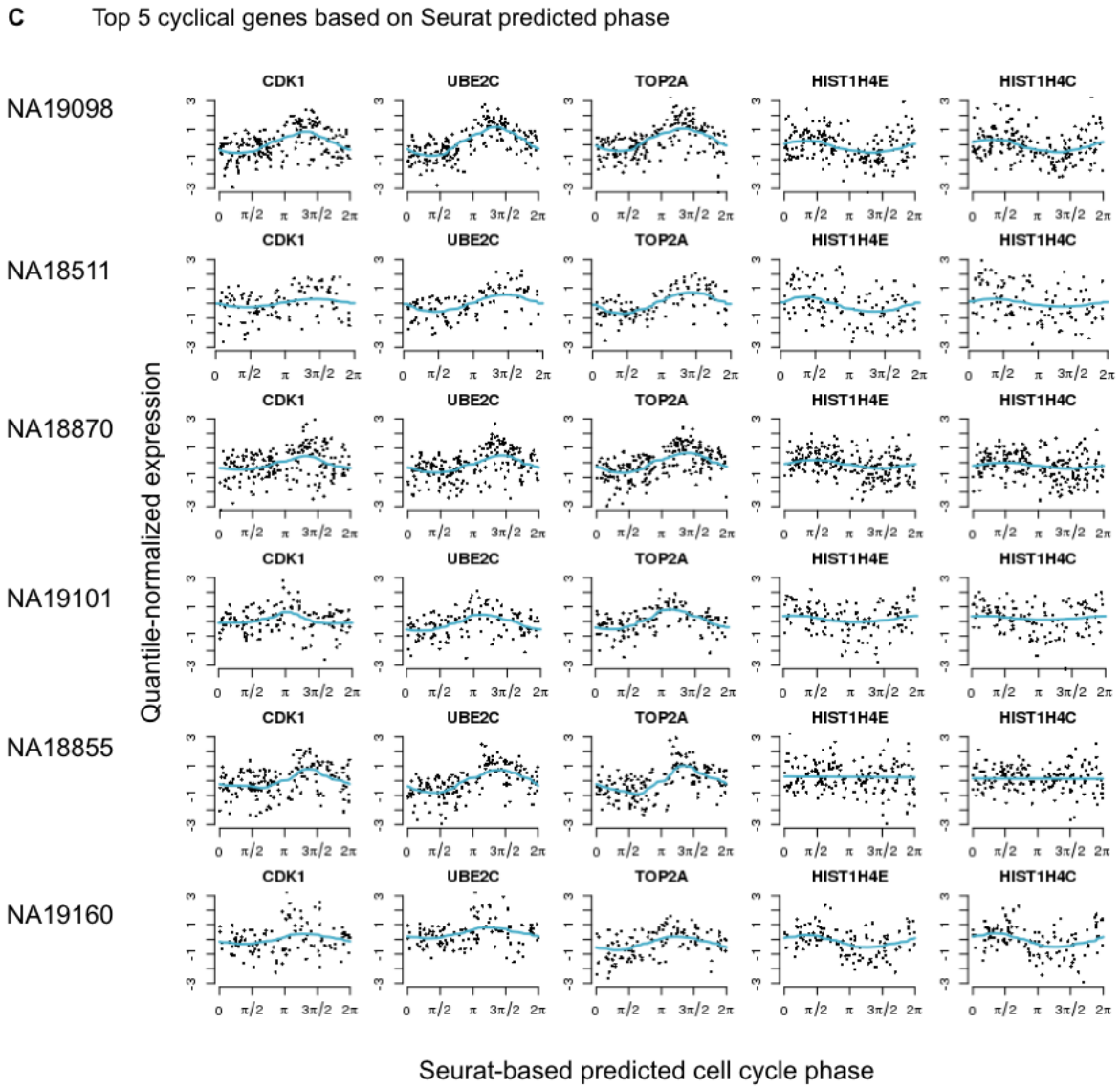
Supplemental Figure S15A-E



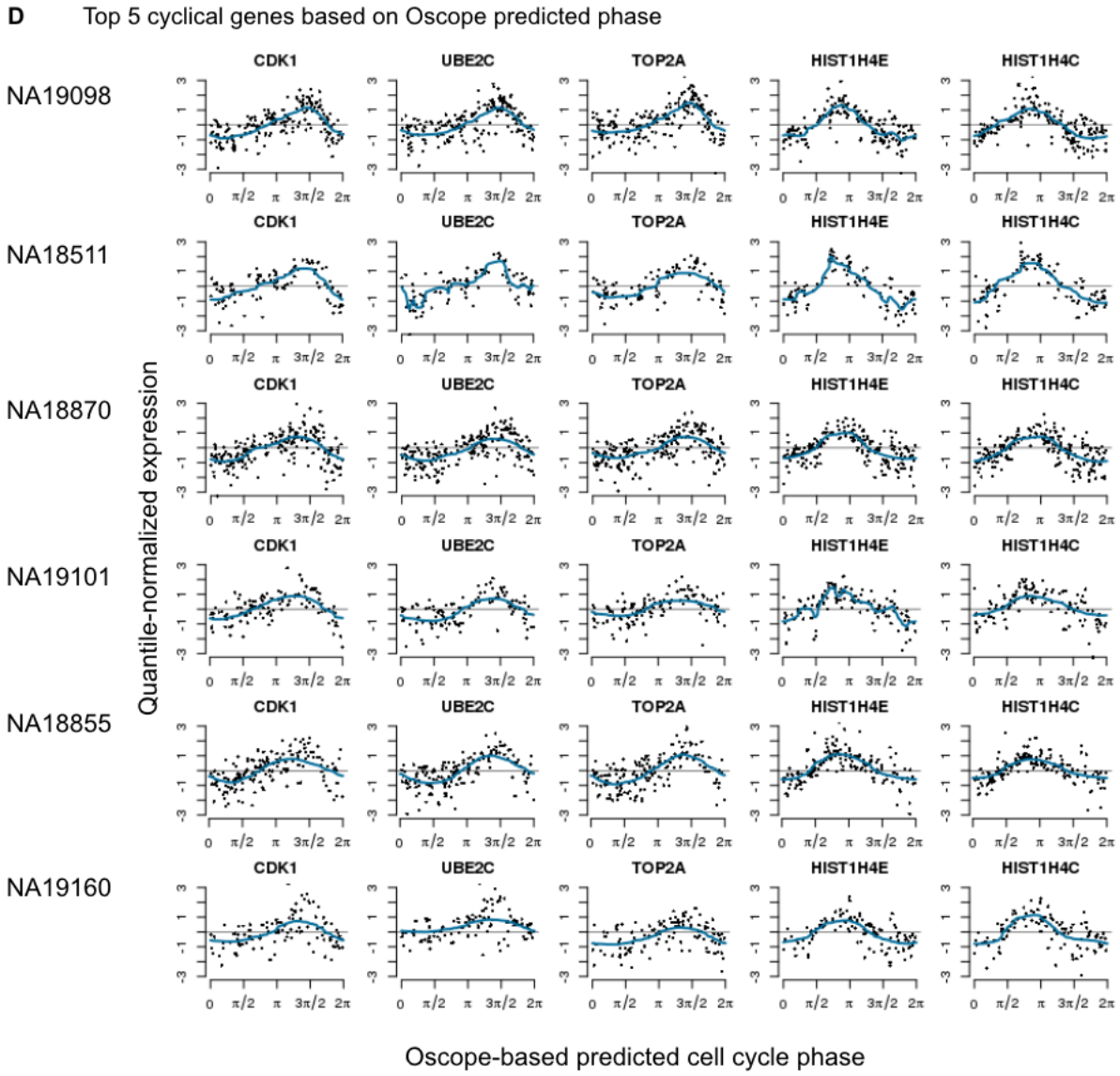
Supplemental Fig. S15A: peco prediction results for the six cell lines, using the simple predictor of 5 genes (*CDK1*, *UBE2C*, *TOP2A*, *HIST1H4E*, *HIST1H4C*). Rows correspond to results for individual cell lines. For example, for cell line NA19098, we ordered samples by FUCCI phase and used *trendfilter* to estimate the cyclic trend of gene expression in the top 5 cyclic genes. The colored line represents the predicted cyclic trend.



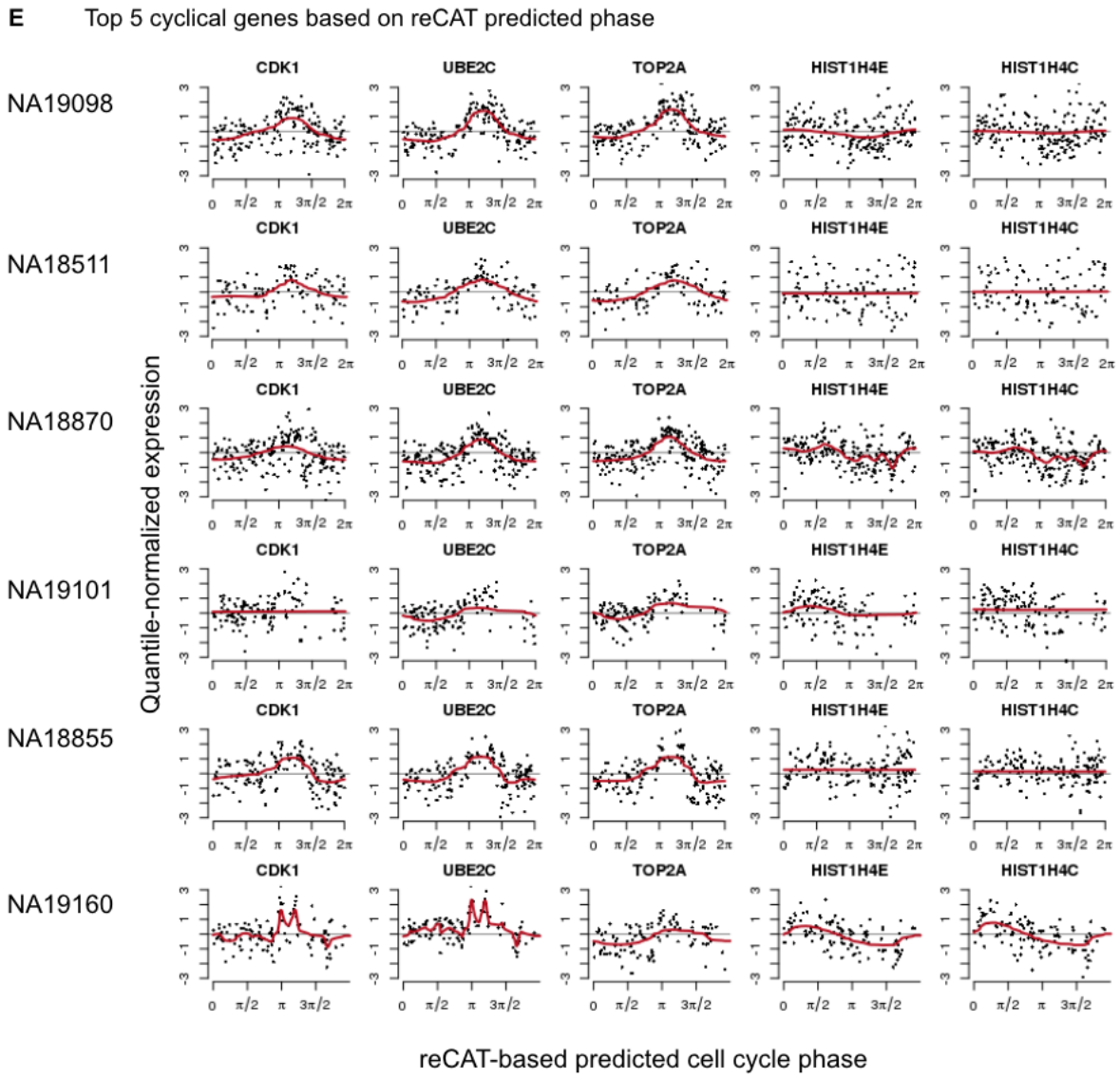
Supplemental Fig. S15B: Cyclone prediction results for the six cell lines. As described in the Results, we transform the three phase-specific Cyclone scores to angles on the unit cirlce, using the same approach for deriving FUCCI phase from FUCCI scores. Rows correspond to results for individual cell lines. For example, for cell line NA19098, we ordered samples by Cyclone-based predicted phase and used *trendfilter* to estimate the cyclic trend of gene expression in the top 5 cyclic genes.



Supplemental Fig. S15C: Seurat prediction results for the six cell lines. As described in the Results, we transform the two phase-specific Seurat scores to angles on the unit cirlce, using the same approach for deriving FUCCI phase from FUCCI scores. Rows correspond to results for individual cell lines. For example, for cell line NA19098, we ordered samples by Seurat-based predicted phase and used *trendfilter* to estimate the cyclic trend of gene expression in the top 5 cyclic genes.

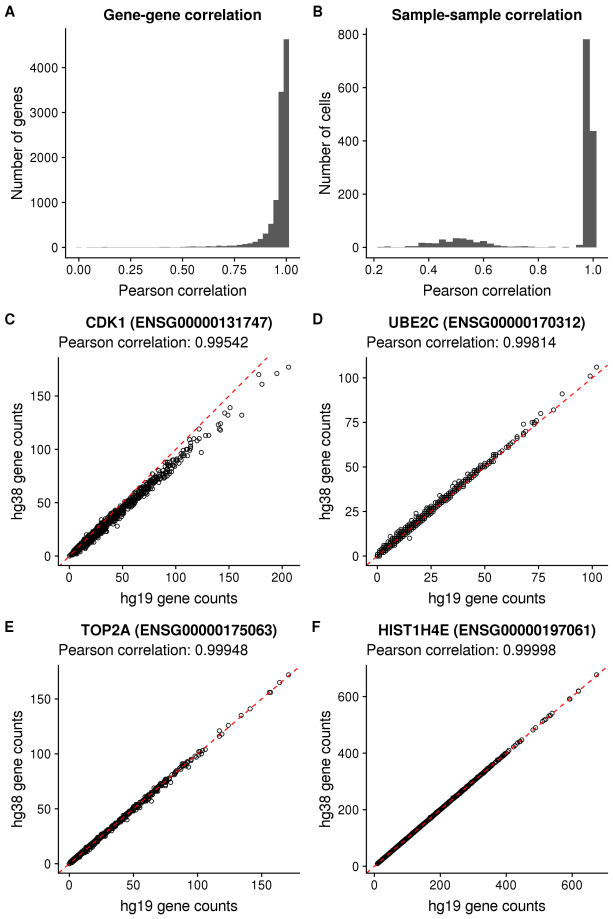


Supplemental Fig. S15D: Oscope prediction results for the six cell lines. As described in the Results, we estimated the cyclic ordering of cells across the 888 high-quality single-cell samples in the data. We then assigned each cell an angle on the unit circle based on the ordering per individual cell line. Rows correspond to results for individual cell lines. For example, for cell line NA19098, we ordered samples by Oscope predicted phase and used *trendfilter* to estimate the cyclic trend of gene expression in the top 5 cyclic genes.



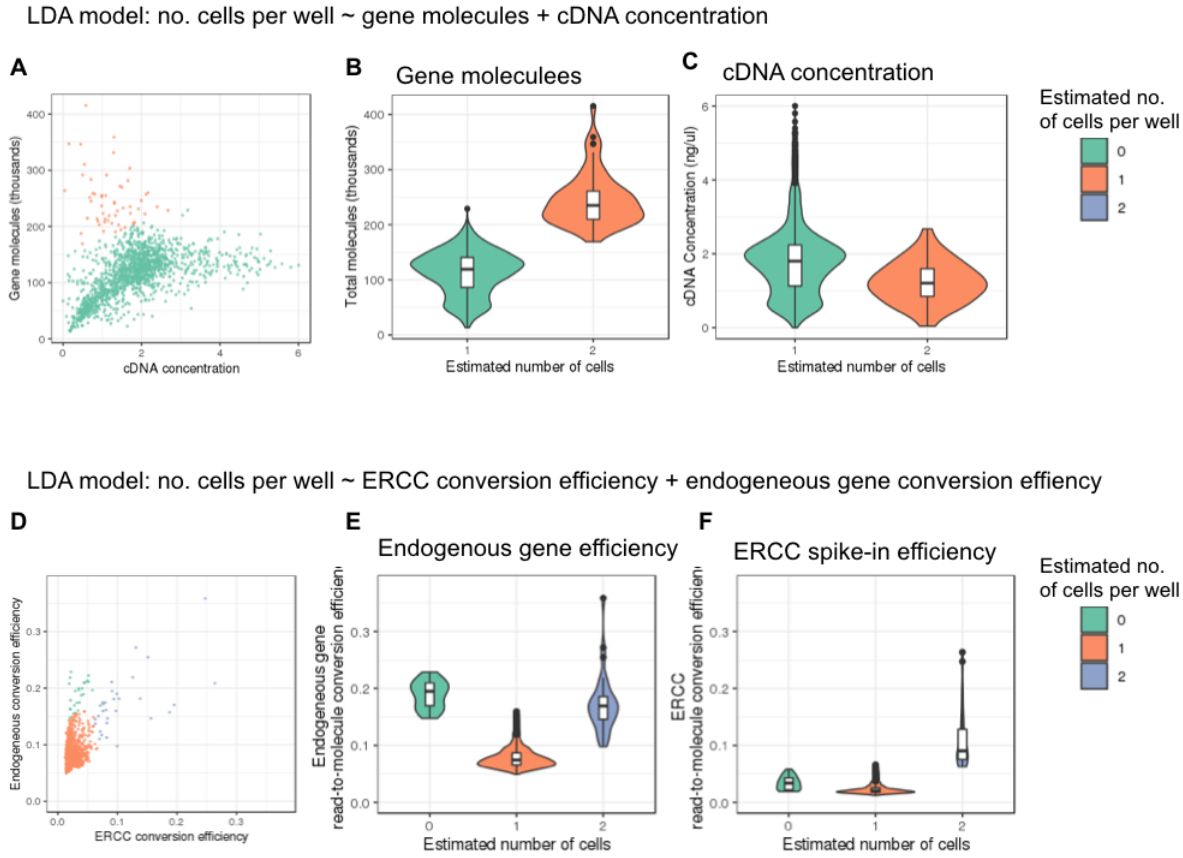
Supplemental Fig. S15E: reCAT prediction results for the six cell lines. As described in the Results, we estimated the cyclic ordering of cells across the 888 high-quality single-cell samples in the data. We then assigned each cell an angle on the unit circle based on the ordering per individual cell line. Rows correspond to results for individual cell lines. For example, for cell line NA19098, we ordered samples by reCAT predicted phase and used *trendfilter* to estimate the cyclic trend of gene expression in the top 5 cyclic genes.

Supplemental Figure 16



Supplemental Fig. S16: Comparison of global gene expression profiles obtain from genome build hg19 vs hg38 (A) Histogram of Pearson correlation for each gene comparing its counts across single cells when mapping to genome build hg19 vs hg38. The median correlation was 0.983919, and 1,070 genes (9.8%) had a correlation less than 0.9. (B) Histogram of Pearson correlation for each single-cell sample comparing its gene counts when mapped to genome build hg19 vs hg38. The median correlation was 0.9849, and 306 single cells (19.9%) had a correlation less than 0.9. (C) A scatterplot of cell cycle gene *CDK1* (ENSG00000131747) gene counts for hg19 (x-axis) vs hg38 (y-axis). The Pearson correlation across single cells was 0.99542. (D) A scatterplot of cell cycle gene *UBE2C* (ENSG00000170312) gene counts for hg19 (x-axis) versus hg38 (y-axis). The Pearson correlation across single cells was 0.99814. (E) A scatterplot of cell cycle gene *TOP2A* (ENSG00000175063) gene counts for hg19 (x-axis) versus hg38 (y-axis). The Pearson correlation across single cells was 0.99948. (F) A scatterplot of cell cycle gene *HIST1H4E* (ENSG00000197061) gene counts for hg19 (x-axis) versus hg38 (y-axis). The Pearson correlation across single cells was 0.99998. The dashed red line represents the 1-1 line. All plots used data for the 10,297 protein-coding genes that were shared between genome builds hg19 and hg38. Note that cell cycle gene *HIST1H4C* (ENSG00000198518) was deprecated in the annotation for genome build hg38.

Supplemental Figure 17



Supplemental Fig. S17: As a part of the quality control analysis, we used LDA to determine the number of cells captured in each well. Specifically, we fitted two LDA models: 1) number of cells $\sim$ gene molecule count + cDNA amplicons concentration, and 2) number of cells $\sim$ ERCC spike-in control read-to-molecule conversion efficiency + endogenous gene read-to-molecule conversion efficiency. We determined the observed number of cells captured in each C1 well based on DAPI staining results. (A) plots the relationship between cDNA concentration and gene molecule, with sample points colored by the predicted number of cells per well in LDA analysis. (B) and (C) show the distribution of gene molecule and cDNA concentration in wells predicted to have 1 cell and wells predicted to have 2 cells. (D) plots the relationship between the read-to-molecule conversion efficiency of ERCC controls and endogenous genes, with sample points colored by the predicted number of cells per well in LDA analysis. (E) and (F) show the distribution of ERCC and endogenous gene read-to-molecule conversion efficiency in wells predicted to have 0, 1, and 2 cells in LDA analysis.