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

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

[1]Department of Human Genetics, University of Chicago
[2]Department of Statistics, University of Chicago
[3]Department of Medicine, University of Chicago
[4]These authors contribute equally to this work.

*Correspondence should be addressed to C.J.H. (joyce.hsiao1@gmail.com) 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 and 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 phase (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 at 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 low resolution data.

Indeed, 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: we do not expect the transition of cells between 'phases' 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 (iPSC) 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] (Figure 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 (Chromotin 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.

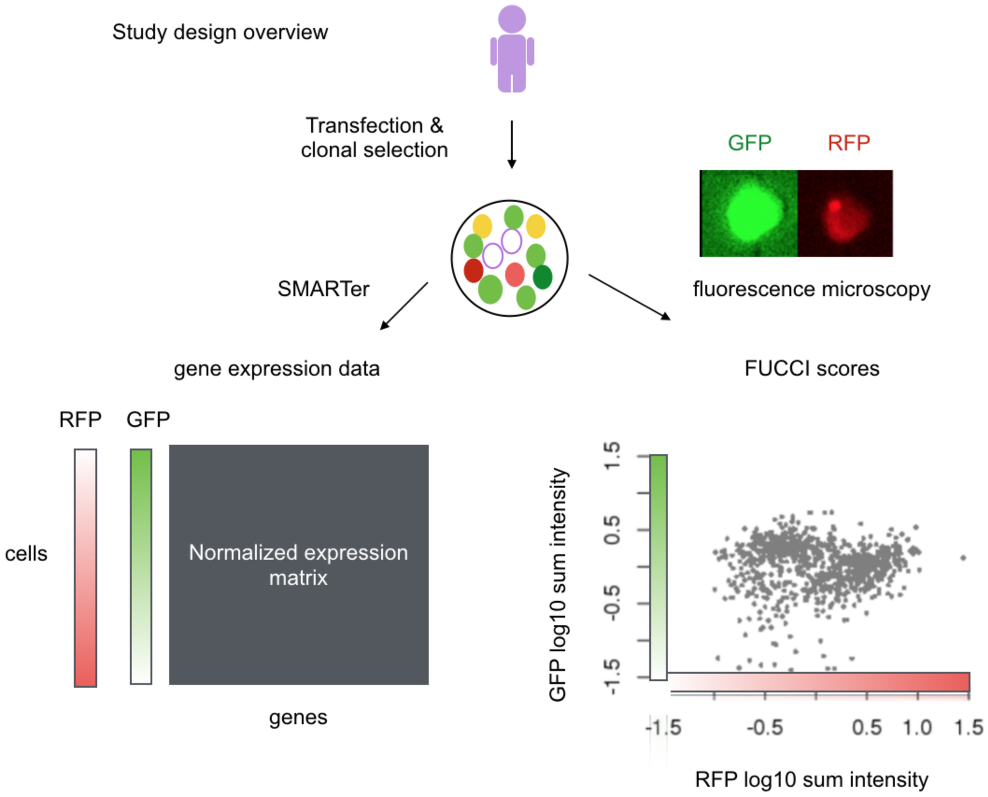

Figure 1: Study design overview. We collected two types of data from FUCCI-expressing iPSCs, including in situ fluorescent imaging and single-cell RNA-seq, and from the same single cells. FUCCI scores correspond to the standardized and background-corrected fluorescence intensities for all sample images of Gemini-EGFP (GFP) and Cdt1-mCherry (RFP), namely the log10 sum of background-corrected fluorescence intensities in the defined cell area (100 pixel squares) for each C1 capture site.

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 Supp. Figure 1 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

We obtained an average of 1.7 +/- 0.6 million sequencing reads per sample (range=0.08-3.0 million). After quality control (see Methods for details), 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 (Supp. Figure 2, Supp. Figure 3). We standardized the molecule counts per individual to $\log_2$ counts per million (CPM) and transformed the data per gene to obtain a standardized normal distribution. We retained all genes with >1 $\log_2$ CPM 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 %, Supp. Figure 3).

We used principal components analysis (PCA) to assess the global influence of technical factors on expression, including plate, individual, and read depth (Supp. Figure 4). 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 $p < 2e\text{-}16$ for both EGFP and mCherry; see Supp. Figure 5, Supp. Figure 6). 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]. 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 (Figure 2A; see Methods). For example, Figure 2B shows that as a cell progresses through 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.

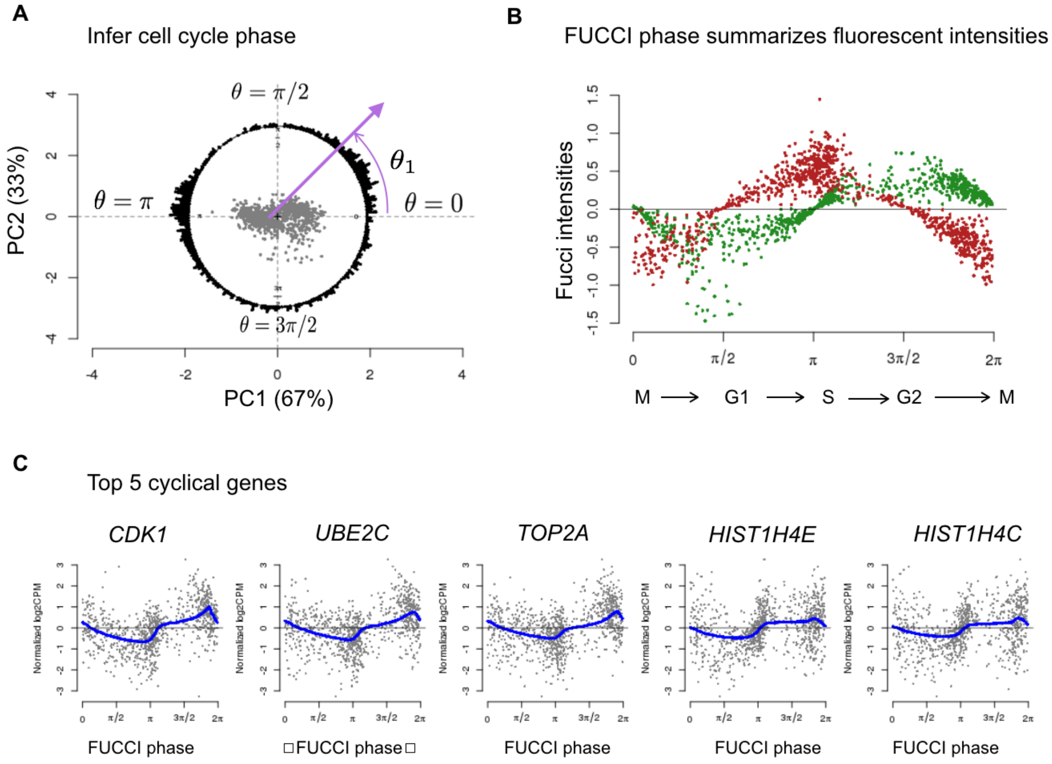

Figure 2: Characterizing cell cycle phase using FUCCI fluorescence intensities. (A) We inferred FUCCI phase (angles in a circle) based on the FUCCI scores of GFP and RFP. For example, we inferred θ_1 based on the PC scores derived from the cell's FUCCI scores. (B) FUCCI phase summarizes the co-oscillation of GFP and RFP along the cell cycle. (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 dot represents a single-cell sample. Blue line indicates the estimated cyclic trend. All 5 genes were previously identified as related to cell-cycle regulation.

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 [29], 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 (622 genes annotated in Whitfield et al., 2002 [30], Odds Ratio=31 for top 5 genes, 30 for top 50 genes, 27 for top 100 genes; Fisher's exact test P-value $< .001$). These results provide strong independent support that the inferred FUCCI phase is indeed meaningfully capturing cell cycle progression.

For illustration, Figure 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 [30] 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 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

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.

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 Figure 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%. 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.

Figure 3B shows the performance of predictors built using between 5 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 (i.e., those in Figure 2C) performed just as well as the predictors with more genes (Supp. Figure 7A shows results up to 500 genes).

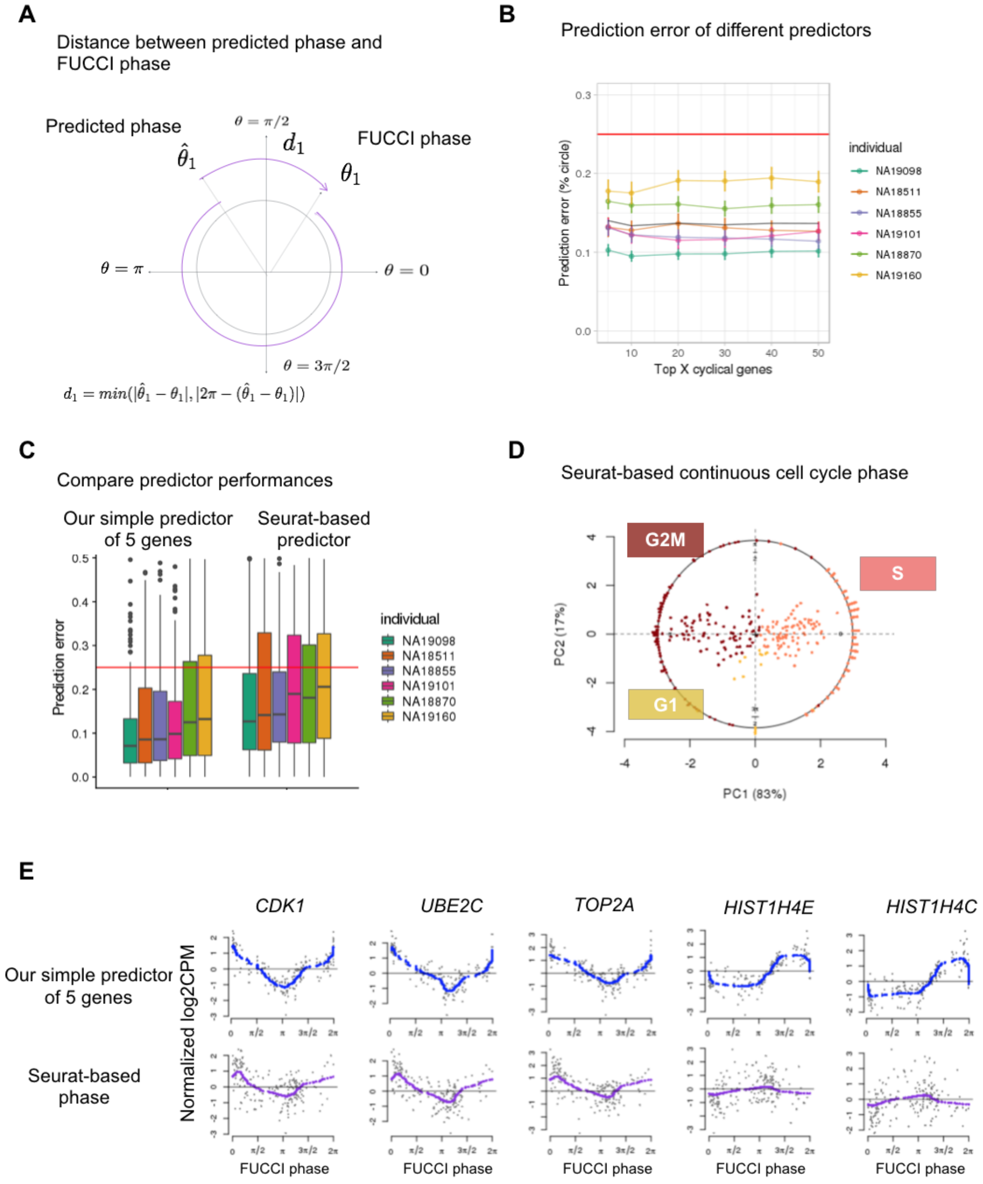

Figure 3: Inferring cell cycle phase from scRNA-seq data. (A) We defined prediction error as the distance between predicted phase and FUCCI phase (as a percentage of the entire cycle, 2π. (B) We compared the performance of our predictors built using 5, 10, 20, 30, 40, and 50 top cyclic genes. The red line indicates the the expected prediction error from random guessing, which would be 25 %. (C) Performance of our predictor of 5 genes (left) and the Seurat-based predictor (right). (D) We inferred Seurat-based cell cycle phase in a continuum based on Seurat's cell cycle scores (for individual NA19098). (E) We compared the cell cycle phase predictions (for individual NA19098) using our predictor of 5 genes (top panel) with predictions using the Seurat-based predictor (bottom panels). The blue and purple lines correspond to the estimated cyclic trends for each gene.

Comparing the performance of our cell cycle predictor to existing methods is complicated by the fact that existing methods typically assign each cell to a discrete cell cycle phase, and not a continuum. For example, Seurat [21] uses the mean expression levels of 43 “S phase" marker genes and 54 “G2/M phase" marker genes to compute standardized scores for S and G2/M phase for each cell, and then assigns each cell to the phase with the highest score (if both scores are negative, the cell is assigned to G1 phase, see Supp. Figure 8 for an example). In contrast, our approach explicitly assumes a continuous process rather than discrete states. Thus, to provide a basic benchmark against which to compare our method, we built a continuous predictor based on Seurat scores. Specifically, we applied the same approach used to derive FUCCI phase to transform the two Seurat scores to cell cycle angles (Figure 3D). This is not a terribly fair comparison because Seurat was not optimized for continuous phase predictions, but no method other than ours has been so optimized. With that caveat in mind, our predictor outperformed the Seurat-based predictor on all cell lines (Figure 3C,E, Wilcoxon test P-value $< .05$ for NA19160, P-value $< .005$ for all other cell lines). Overall, the mean prediction accuracy across the six cell lines was 1.5 times larger for our predictor than the Seurat-based predictor.

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*, *HIST1C*, *HIST1E*), 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 [30] and yeast [31], and in previous scRNA-seq studies of FUCCI-sorted human embryonic stem cells [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.

We believe that our observations can be of use 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). Further efforts will be necessary to ascertain to what extent our predictor will be helpful for predicting cell cycle phase in future studies that involve 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 Fluorescent ubiquitination cell cycle indicator (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], 29 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, 2017 [27]. Specifically, 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$^{\mathrm{TM}}$ at 37°C for three minutes. Cell viability and cell number were measured to have equal number of live cells from the two FUCCI-iPSC lines. In addition, single cell suspension was stained with 5 uM Vybrant$^{\mathrm{TM}}$ DyeCycle$^{\mathrm{TM}}$ 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, GFP, and RFP. 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 for each capture site.

Library preparation and read mapping

For sequencing library preparation, tagmentation and isolation of 5′ fragments were performed as described in Tung et al., 2017 [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 [32] 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 extracted the UMIs from the 5′ end of each read (pre-mapping) and deduplicated the UMIs (post-mapping) with UMI-tools [33]. We counted the molecules per protein-coding gene (Ensembl 75) with featureCounts [34]. Lastly, we matched each single cell to its individual of origin with verifyBamID [35] by comparing the genetic variation present in the RNA-seq reads to the known genotypes, as previously described [27].

Filtering gene expression data and normalization

We filtered the samples based on imaging data and RNA-sequencing data. Specifically, we used DAPI fluorescence imaging data to determine the number of cells captured in each C1 well. Then, we compared data collected from wells observed having zero cells versus wells observed having one or more cells present to established various sequencing data quality metrics. The quality control analyses are listed below and described in more details in our previous work (Tung et al., 2017 [27]).

- 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

In addition, we performed linear discriminant analysis (LDA) to infer the number of cells in each well using 1) gene molecule count and RNA concentration, and 2) the conversion efficiency of the endogenous genes and of the ERCC spike-ins. Supp. Figure 9 shows the observed versus the inferred number of cells in each well. 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 $\log_2$ counts-per-million (CPM) using a pseudo count of 1 and 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 [36]. 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 gives us 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 principal component analysis to transform the two FUCCI scores to orthogonal vectors. Using the PCs of the FUCCI scores, we infer 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 transformed the $\log_2$ CPM values of gene expression levels to a standard normal distribution for each gene. This way, the samples with zero molecule-count are assigned the lowest level of gene expression. We applied quadratic (second order) trend filtering using the *trendfilter* function in the *genlasso* package [29]. 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 (log2CPM < 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 applied six-fold cross-validation to data collected from 6 YRI individuals. In each fold, we trained our predictor on data from 5 individuals (training data) and tested its performance on cells from the remaining individual (test data). Thus, we assessed our predictor's ability to generalize to individuals not seen in training. We implemented the method in a two-step algorithm. In the first step, we used the training data to characterize how gene expression levels vary in a cyclic way through the cell cycle. We applied these gene-specific cyclic trends to the predictor in the second step to estimate the cyclic trends of gene expression levels in the test data. In the second step, we computed the likelihood of test data gene expression levels on grid points selected along a circle. The grid points were selected to span 100 equally-spaced cell cycle phases. Finally, the cells in the test data were assigned to cell cycle phases to maximize the likelihood of gene expression.

Notations

- $(Y_n^{train}, \hat{\theta}_n^{train})_{n=1,\dots,N}$: For each individual cell n in the training sample, we denote $Y_n^{train} = (Y_{1n}^{train}, \dots, Y_{Gn}^{train})'$ as the $\log_2$ normalized gene expression vector, and $\hat{\theta}_n$ the FUCCI-based cell cycle phases. The single-cell samples are ordered in FUCCI time, where $0 \leq \hat{\theta}_1^{train} < \dots \hat{\theta}_N^{train} < 2\pi$.

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

- $(\hat{f}_g, \hat{\sigma}_g)_{g=1,\dots,G}$: 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 an internal 5-fold cross-validation. The degree of smoothing is allowed to vary between the genes.
   (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:
   $$L_m(k) = L(\theta_m = k | Y_m^{test}, \big(\hat{f}_g(\theta_m = k), \hat{\sigma}_g)_{g=1,\ldots,G}\big) = \prod_{g=1}^{G} P\big(Y_{gm}^{test} | \hat{f}_g(\theta_m = k), \hat{\sigma}_g\big)$$
   (c) Maximize $L_m(k)$ over $k = 1, \ldots, K$:
   $$\hat{\theta}_m^{test} = \operatorname*{argmax}_{k=1,\ldots,K} L_m(k)$$

The standard approach to cell cycle phase assignment

Seurat [21] provides *CellCycleScoring* function to assign single cells to discrete cell cycle phases. They used 97 cell cycle marker genes that were identified in Tirosh et al., 2016 [37], which included 43 for S phase and 44 for G2M phase. We outlined the steps for assigning Seurat-based cell cycle phase as follows:

1. Standardize gene expression levels to $\log_2$ CPM.
2. Group all genes in the data based on their mean expression levels and assign genes to 25 equal-sized bins. Compute the mean expression level of each bin. We used the default setting of 25 equal-sized bins.
3. For each cell, standardize the S-phase genes by subtracting mean expression of the corresponding bin from the observed gene expression levels. Repeat for G2M genes. This step computes standardized S-score and G2M-score for each cell.
4. The cells are assigned to G1 if both S-score and G2M-score are lower than average (i.e., negative). Otherwise, the cells are assigned to S or G2M, depending on whichever score is higher.

   To compare with our continuous cell cycle phase, we used S-scores and G2M scores to compute Seurat-based continuous cell cycle phase. We applied the

Seurat-based cell cycle phase to train a Seurat-based predictor for cell cycle phase - Seurat-based predictor. Specifically, we performed PCA analysis on S scores and G2M scores, and used the PC1 and PC2 score from the PCA analysis to infer an angle for each cell, which we referred to as Seurat-based cell cycle phase.

Code and data availability

The data have been deposited in NCBI's Gene Expression Omnibus [38] and are accessible through GEO Series accession number GSE121265 (`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121265`). The processed data are available at `https://github.com/jdblischak/fucci-seq`, including molecule count data, summarized fluorescence intensities and sample phenotype information. The results of our analysis are viewable at `https://jdblischak.github.io/fucci-seq`. The peco package for predicting cell cycle progression in iPSCs is avaiable 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. 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 KB 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 KB. 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] Tibshirani, R. J. Adaptive piecewise polynomial estimation via trend filtering. *Ann. Stat.* **42**, 285–323 (2014).

[30] 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).

[31] 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).

[32] 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).

[33] 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).

[34] 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).

[35] 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).

[36] 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).

[37] Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).

[38] 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).