Computational Cell Cycle Analysis of Single Cell RNA-Seq Data

Marmar Moussa[1(✉)] and Ion I. Măndoiu[2]

[1] School of Medicine, University of Connecticut, Farmington, CT, USA
marmar.moussa@uconn.edu
[2] University of Connecticut, Storrs, CT, USA
ion@engr.uconn.edu

Abstract. The variation in gene expression profiles of cells captured in different phases of the cell cycle can interfere with cell type identification and functional analysis of single cell RNA-Seq (scRNA-Seq) data. In this paper, we introduce SC1CC (**SC1 C**ell **C**ycle analysis tool), a computational approach for clustering and ordering single cell transcriptional profiles according to their progression along cell cycle phases. We also introduce a new robust metric, Gene Smoothness Score (GSS) for assessing the cell cycle based order of the cells. SC1CC is available as part of the SC1 web-based scRNA-Seq analysis pipeline, publicly accessible at https://sc1.engr.uconn.edu/.

Keywords: scRNA-Seq · Cell cycle · Cell order · Gene smoothness score

1 Background and Motivation

The variation in gene expression profiles of single cells that are captured in different phases of the cell cycle can interfere with cell type identification and functional analysis of single cell transcriptomic data. In particular, it is important to differentiate between cell type and cell cycle effects when analyzing single cell RNA-Seq data. A first challenge in the computational analysis of cell cycle effects in single cell transcriptomics is to differentiate between cells that are actively proliferating and those that are quiescent, i.e., cells that do not actively divide but retain the ability to re-enter a proliferative state. A second computational challenge is to correctly label individual cells or cell clusters according to their phase in the cell cycle. The main cell cycle phases are G1 (where metabolic changes prepare the cell for division), S (where DNA synthesis replicates the genetic material), G2 (where molecular components needed for mitosis and cytokinesis are assembled), and M (where a nuclear division followed by cytokinesis occurs), although transition phases G1/S and G2/M are also commonly identified [4]. Such cell labels coupled with existing biological knowledge of genes associated with each of the cell cycle phases can assist functional analysis of single cell

transcriptional profiles and interpretation of unsupervised scRNA-Seq clustering results. Finally, a third computational challenge is to order individual cells according to their progression along the cell cycle.

Although there are several existing methods for cell cycle analysis of single cell RNA-Seq data, most of them attempt to address one of the above-mentioned challenges in isolation. Our proposed SC1CC method enables a comprehensive analysis of the cell cycle effects that can be performed independently of cell type/functional annotation, hence avoiding hazardous manipulation of the single cell transcription data that could lead to misleading analysis results. Specifically, SC1CC can be used to distinguish proliferating from quiescent cells and provides the ability to annotate cell populations based on the cell cycle phase. Additionally, the cells are also ordered based on their progression along the cell cycle phases.

In the remainder of this section we briefly review some representative methods for individually addressing the above challenges in cell cycle analysis of scRNA-Seq.

ccRemover. The ccRemover tool [2] attempts to *remove* the cell-cycle effects from the single cell transcriptional profiles. This is done by identifying those principal components that, based on their loadings, capture mostly cell cycle effects in a low dimensional principal component analysis (PCA) projection of the scRNA-Seq data. Subtracting these components is expected to enhance gene expression variation due to differences in cell type. We performed an initial test to determine the effectiveness of ccRemover at removing cell cycle effects by running it with default settings on a dataset consisting of a 50%–50% mixture of Jurkat and 293T single cells that was previously profiled using the 10x Genomics droplet-based scRNA-Seq platform. This dataset is comprised of cells of two different types (T lymphocyte and human embryonic kidney cells) that are well separated according to their original scRNA-Seq profiles (Fig. 1a). However, after processing the scRNA-Seq data using ccRemover the two cell types appear nearly indistinguishable in the 3D t-SNE (t-Distributed Stochastic Neighbor Embedding) plot (Fig. 1b). This suggests that attempting to subtract the cell cycle signal using ccRemover without careful parameter tuning could result in inadvertently subtracting the cell type signal. For this reason, ccRemover was not included in further method comparisons in this paper.

Cyclone. Cyclone [16] uses a classification algorithm based on selecting pairs of genes whose relative expression has a sign that changes with the cell-cycle phase in the training data. The learned gene pairs are used to quantify the evidence that a given cell is in one of three cell cycle phases (G1, S, or G2M). Specifically, under the recommended approach, Cyclone calculates for each cell a score between 0 and 1 for two of these phases, G1 and G2M. Cells with G1 or G2M scores above 0.5 are assigned to the G1 or G2M phases, respectively (if both scores are above 0.5, then the higher score is used to make the assignment). Cells with both G1 and G2M scores below 0.5 are assigned by default to the S phase. The method allows users to override these thresholds, but we used
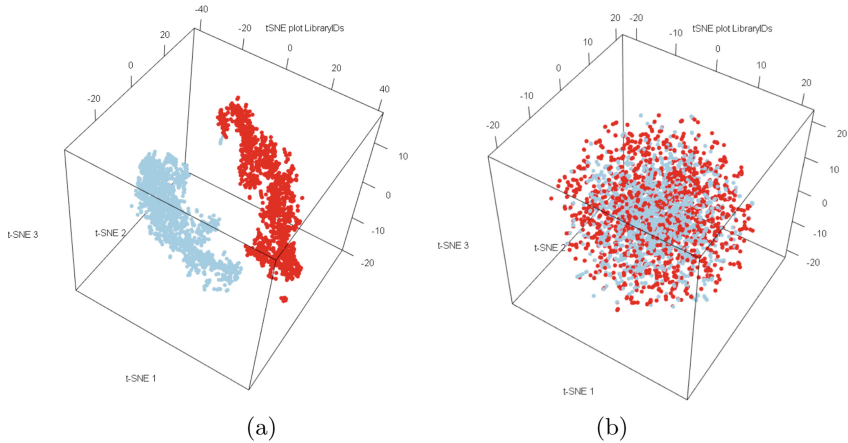
Fig. 1. ccRemover effect on data variability. (a) 3D t-SNE plot of Jurkat-239T data (Blue and red colors distinguish the Jurkat and 239T cells respectively). (b) 3D t-SNE plot of the Jurkat-293T dataset after applying ccRemover with default settings. As in [19] we inferred the cluster/library labels based on the expression of cell type-specific markers, The blue cluster corresponds to Jurkat cells (preferentially expressing CD3D), and red corresponds to 293T cells (preferentially expressing XIST, as 293T is a female cell line, while Jurkat is a male cell line).

the recommended thresholds in our experiments. In Sect. 3 we present results comparing the accuracy of cell cycle labels inferred by Cyclone to those generated by SC1CC using datasets with both known and unknown cell cycle phase labels.

reCAT. The reCAT method [12] takes a different approach to cell cycle analysis. Rather than labeling the cells with an inferred cell cycle phase, reCAT attempts to *order* the cells in a manner consistent to their position along the cell cycle. The cell ordering problem is computationally modeled as a traveling salesman problem (TSP). First, reCAT performs normalization of the data followed by clustering of the cells. It then orders the identified clusters by finding a traveling salesman cycle. It also computes for each cell two scores (a Bayes score and a mean score) that differentiate between the cell cycle phases. Finally, a hidden Markov model (HMM) and a Kalman smoother are used to estimate the underlying gene expression levels of the ordered single cells. The results of experiments comparing the order reconstructed by reCAT to the order identified by SC1CC are presented in Sect. 3.

2   Methods

2.1   Datasets

In addition to the Jurkat-239T dataset described in Sect. 1 we used four other datasets to further evaluate the performance of SC1CC and existing cell cycle

analysis tools. These datasets were selected to span a broad range of cell cycle related modalities. For example, all cells in the Human Embryonic Stem Cells (hESC) dataset are expected to be proliferating, whereas the Peripheral blood mononuclear cells (PBMC) dataset is expected to consist solely of quiescent cells [19]. The immune cells from anti-CTLA-4 treated mice (α-CTLA-4) dataset and the mouse Hematopoietic Stem Cells (mHSC) are both expected to contain a mix of quiescent and proliferating cells.

The cells of the hESC dataset have labeled cell cycle phase annotations, while for the α-CTLA-4 and mHSC datasets only the percentage of proliferating cells was established in the original publications.

Basic quality control (QC) was uniformly applied to each of these datasets, whereby cells expressing less than 500 genes as well as genes detected in less than 10 cells were filtered out. Pre-processed versions of all datasets are accessible as example datasets for the SC1 web-based scRNA-Seq analysis pipeline [13], publicly available at https://sc1.engr.uconn.edu/.

Human Embryonic Stem Cells (hESC, Cycling Cells). There are very few scRNA-Seq datasets where the cell-cycle phase of each cell is known *a priori*. For this work, we used a labeled dataset of undifferentiated H1 human embryonic stem cells (hESCs) from [11]. Fluorescent ubiquitination-based cell-cycle indicator H1 (H1-Fucci) human embryonic stem cells were sorted according to the G1, S, and G2/M cell cycle phases by fluorescence activated cell sorting (FACS). Full-length scRNA-Seq data was generated for a total of 247 H1-Fucci cells (91 G1, 80 S, and 76 G2/M cells, respectively) captured using the Fluidigm C1 microfluidic platform.

Peripheral Blood Mononuclear Cells (PBMC, Non-cycling Cells). The PBMC dataset is comprised of a mixture of mature FACS-sorted dendritic cells, natural killer, B and T cells from a healthy donor from [19] and further analyzed in [14]. This dataset consists of 2,882 cells randomly sampled from seven PBMC sub-populations independently sorted by FACS. scRNA-Seq data for these cells was generated using the 10x Genomics droplet-based platform and the 3′-end v1 protocol, as described in [19]. Figure 2a shows a 3-dimensional *t-Distributed Stochastic Neighbor Embedding (t-SNE)* plot of the PBMC dataset and the breakdown into the seven cell types. Since PBMCs typically differentiate in the thymus or lymph nodes, this dataset is expected to contain only non-cycling cells.

Tumor Infiltrating Immune Cells from Anti-CTLA-4 Treated Mice (α-CTLA-4, Mixture of Cycling and Non-cycling Cells). This dataset (publicly available in the NCBI GEO database under accession GSM3371686) was also generated using the 3′-end v1 scRNA-Seq protocol on the 10x Genomics platform. CD45+ cells were sorted by FACS from cell suspensions of dissociated tumors excised from mice treated with 9D9, an anti-CTLA-4 antibody, as
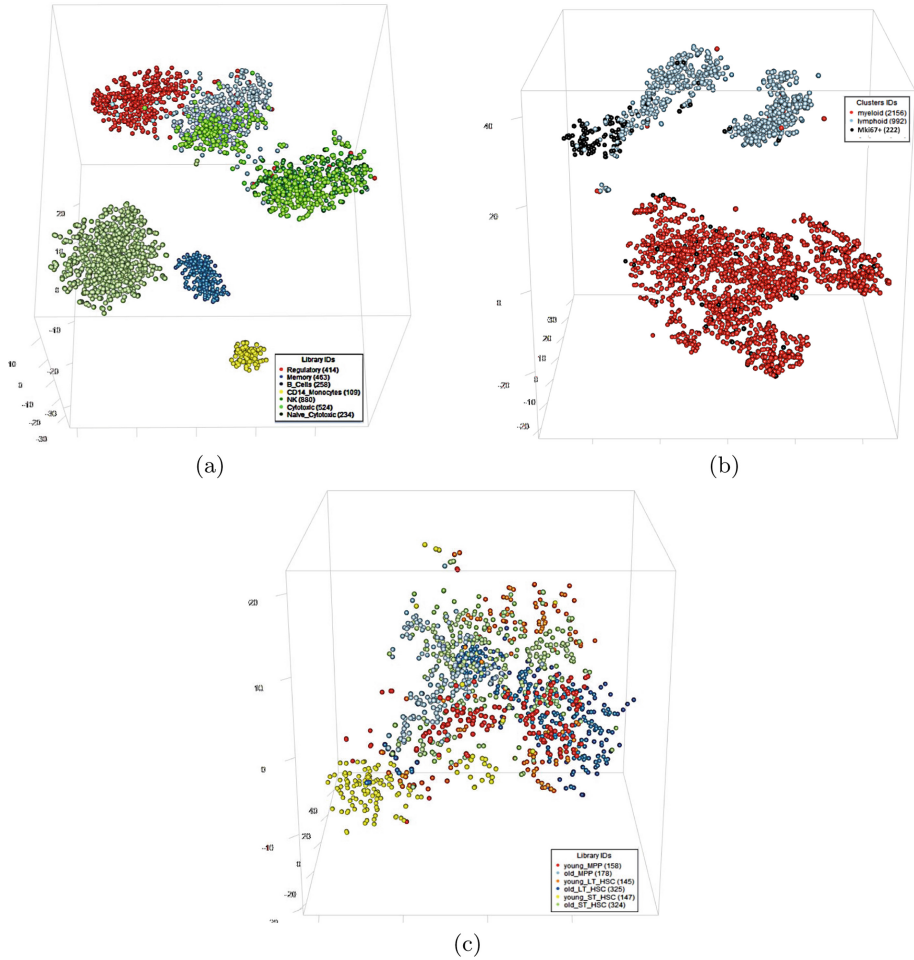
Fig. 2. 3-dimensional t-SNE plots of the PBMC, α-CTLA-4 and mHSC datasets. (a) 3D t-SNE plot of the 10x Genomics PBMC dataset consisting of 2,882 cells randomly sampled from seven PBMC sub-populations independently sorted by FACS. (b) 3D t-SNE plot of the α-CTLA-4 dataset consisting of 992 lymphoid (blue) and 2156 myeloid cells (red). The 'ki67-Hi' cells (black) are a mixture of proliferating CD4+ T cells, CD8+ T cells, Tregs, and NK cells ($\approx 17.5\%$ of the lymphoid cells). (c) 3D t-SNE plot of the mHSC dataset consisting of a total of 1,277 MPP, ST-HSC, and LT-HSC cells, further grouped by the age of the mice (young and old).

described in [7]. According to the analysis in [7], this dataset, henceforth referred to as α-CTLA-4, consists of 992 lymphoid and 2,156 myeloid cells. Notably, the unsupervised clustering analysis of the α-CTLA-4 dataset in [7] has identified a cluster, labeled 'Mki67-Hi', comprised of a mixture of proliferatingCD4+ T

cells, CD8+ T cells, Tregs, and NK cells ($\approx$17.5% of the lymphoid cells, see Fig. 2b). Thus, this dataset is well suited for assessing the ability of our method to correctly differentiate between quiescent and proliferating cells.

Short- and Long-Term Mouse Hematopoietic Stem Cells from Young and Old Mice (mHSC, Mixture of Cycling and Non-cycling Cells). This scRNA-seq dataset (1,277 cells after applying QC) is publicly available in the NCBI GEO database under accession GSE59114. The dataset was used in [9] to dissect the variability in hematopoietic stem cell (HSC) and hematopoietic progenitor cell populations from young and old mice. A 3D t-SNE projection of the mHSC dataset is shown in Fig. 2c. Based on the analysis in [9], this dataset is comprised of cells of three different types – Multipotent Progenitor Cells (MPP), short-term hematopoietic stem cell (ST-HSC), and long-term hematopoietic stem cell (LT-HSC) – that are further grouped by the age of the mice (young and old). The six cell populations are thoroughly analyzed in [9] with regards to cell cycle effect on differentiation while aging. We use the findings of this analysis as the ground truth for evaluating the performance of our approach. Specifically, the computational and biological analysis in [9] identifies 65% of all cells analyzed as non-dividing and estimates an equal percentage of proliferating cells in young and old mice for MPP and ST-HSC cells but not for LT-HSCs (of which old mice have fewer dividing cells). The analysis in [9] also estimates the percentages of cells in G1, S and G2M phases as 20%, 6% and 9% of the total, respectively.

2.2   The SC1 Cell Cycle (SC1CC) Analysis Tool

A repeated observation in single cell RNA-Seq data analysis is that a bias can be introduced by cell cycle effects. Indeed, such effects result in significant factor loadings of annotated cell cycle genes to the first few principal components for many scRNA-Seq datasets. Furthermore, it has been shown that the first few principal components obtained by using expression levels of annotated cell cycle genes are sufficient for capturing cell to cell similarities and the covariance due to cell cycle effects [2,3,5,11,16]. We leverage this observation in SC1CC and start by computing the first few principal components (PCs) for the sub-matrix of normalized scRNA-Seq counts comprised of cell cycle genes only.

The SC1CC implementation available at https://sc1.engr.uconn.edu/ allows users to select one of three different gene lists: the genes annotated with the "cell cycle" term (GO:0007049) in the Gene Ontology database [6], genes included in the Cyclebase 3.0 database of cell cycle related genes [15], and finally the list of periodic genes identified from single cell data in [5]. All results in this paper are based on the GO-annotated cell cycle gene unless otherwise indicated. The selected list of cell cycle genes is further filtered based on the gene expression values in the current dataset in order to keep only those expressed genes that have a correlation higher than α to at least one other cell cycle gene. The purpose of this step is to remove genes that – although annotated as cell cycle genes – do not have expression levels correlated with that of other cell cycle genes, and hence

might represent outliers. All experiments reported in Sect. 3 were generated using the default value of 0.25 for α.

Since using a large number of PCs can add unnecessary noise to subsequent analysis steps, by default SC1CC automatically determines the number of relevant PCs by assessing the drop in variance explained for each pair of consecutive principal components. The online SC1CC implementation allows users to manually specify the number of PCs if desired. The principal component analysis is followed in SC1CC by a 3-dimensional t-SNE projection using the identified principal components. Performing t-SNE based dimensionality reduction using the main PCs aims to capture the local similarity of the cells without sacrificing the global variation already captured by the PC analysis. Next, the cells – now identified by their representation in t-SNE space – are clustered into a hierarchical structure (dendrogram) based on their Cosine similarity. Unless otherwise indicated all results reported in the paper are based on using hierarchical clustering with average linkage; the online SC1CC implementation also allows users to select between average linkage and Ward's method.

Since the cell cycle is typically divided into 6 distinct phases (G1, G1/S, S, G2, G2/M, and M, see, e.g., [4]), by default SC1CC attempts to extract up to 7 clusters from the hierarchical clustering dendogram – corresponding to the 6 cell cycle phases plus at least one potential cluster of non-cycling cells – with a minimum cluster size threshold of 25 cells. The maximum number of clusters can be modified by the user in the online implementation of SC1CC, which also includes an 'auto' option for determining the optimal number of clusters based on the Gap Statistics Analysis algorithm from [17].

Finally, to generate an order of cells consistent to their position along the cell cycle, SC1CC reorders the leaves of the hierarchical clustering dendogram (corresponding to the individual cells) by using the *Optimal Leaf Ordering (OLO)* algorithm [1] as implemented in [8]. Performing additional leaf-node reordering is equivalent to minimizing the length of a Hamiltonian path [1]. For n cells, the dendrogram produced by the hierarchical clustering algorithm (essentially a rooted binary tree) has $n - 1$ internal nodes and 2^{n-1} possible leaf orderings. That is, at each internal node the left and right subtrees can be independently flipped or not. The OLO algorithm produces a leaf ordering that minimizes the sum of distances between adjacent leaves. The time complexity of the implementation in [8] is $O(n^3)$, and its practical performance as part of SC1CC is further improved since the pairwise distances between cells are already available from the distance based hierarchical clustering step.

Cluster Mean-Scores. Six groups of genes (G1, G1/S, S, G2, G2/M, and M genes, respectively) are formed by including cell cycle genes that are known to reach their peak expression in the corresponding cell cycle phases [12]. For each of these genes and each cell, a 'z-score' is computed by subtracting the gene's mean expression level from the expression level of the gene in the cell and then dividing by the gene's standard deviation. For each group of genes and each cluster identified during the hierarchical clustering step we compute a mean-score by averaging

over cells in the cluster and genes in the group. The maximum mean-score of a cluster is used to determine its cell cycle phase. Note that with this procedure multiple clusters can be labeled with the same cell cycle phase, and some cell cycle phases may not be assigned as labels to any of the clusters. Also, since a mean-score of each gene group corresponding to each of the cell cycle phases can be calculated for each identified cluster, the maximum mean-score is relative between gene groups of different cell cycle phases and can only indicate a potential cell cycle phase designation. We therefore introduce in next sub-section an independent metric that can be used to distinguish dividing from non-dividing cells.

Gene-Smoothness Score (GSS). Normalized gene scores computed as above or as defined by reCAT [12] or Cyclone [16] are relative between cell cycle phases and cannot distinguish clearly, if at all, between cycling vs. non-cycling cells or provide a useful metric for assessing cell orderings. We therefore propose a novel metric, referred to as Gene-Smoothness Score (GSS), based on serial correlation, i.e., the correlation between a given variable and a lagged version of itself. The GSS can be computed for any ordered group of cells and can help to directly assess the suggested cell order. Strengths of this metric include the fact that the cells do not need to have known cell-cycle labels and that no specific model assumptions are required for the marker gene expression (whether binary, bimodal, sinusoidal, etc.). Our experiments also indicate that the GSS results are relatively insensitive to the choice of annotated cell cycle genes, hence the GSS can be useful even when a "perfect" annotation is not available.

The GSS of an ordered cluster/group c of cells is defined as

$$GSS(c) = Median \left\{ \left| SC_{ord}(g_i) - \frac{1}{R} \sum_{j=1}^{R} SC_{rand_j}(g_i) \right| \ : \ i = 1, \ldots, N \right\} \quad (1)$$

where N is the number of annotated cell cycle genes, $SC_{ord}(g_i)$ denotes the first-order serial correlation of gene i with respect to the given cell order, and $SC_{rand_j}(g_i)$, $j = 1, \ldots, R$, denote the first-order serial correlation of gene i with respect to R randomized cell orders (we use $R = 50$ in all experiments). The first-order serial (or auto-) correlation is the correlation value between a given gene expression vector and a version of itself shifted by one position. Serial correlation is a value between -1 and 1. First-order serial correlation near 0 implies that there is no overall correlation between adjacent data points. On the other hand, a first-order serial correlation near 1 suggests a smoothly varying series, while a first-order serial correlations near -1 indicates a series that alternates between high and low values. Because individual cell cycle genes can be expressed in different patterns throughout the cell cycle phase transitions, and even abruptly switch direction when the assessed cluster includes mostly cells in one of the transient cell cycle phases (G1/S or G2/M), we define GSS as the median (over all cell cycle genes) of the absolute differences between the serial correlation of a gene's expression values ordered according to the given cell ordering and the average serial correlation computed over R randomized orders.

A cluster/group of cells is considered to be cycling/dividing when its GSS is greater than an error margin ε (default 0.05), i.e., when at least 50% of the genes have an absolute difference in serial correlations between randomized order and identified cell cycle order of at least 0.05. The value of the error margin is set to 0.05 by default but can be adjusted by the user in the online SC1CC implementation. The GSS score is more robust with a higher number of cells per cell cycle cluster, as the chance of a random order producing spurious auto-correlation and therefore high GSS scores is lower when more data points are included in the series.

Figure 3 provides examples of cell cycle genes that contribute positive values to the GSS score in the hESC dataset and illustrates their expression values for both SC1CC and randomly ordered cells. The online implementation of SC1CC allows the user to select any cell cycle gene of interest and examine its normalized expression levels along the inferred order. In Figure 3, gray dots represent normalized gene expression values for individual cells, while the red and blue curves represent the fitted local polynomial regression of these values for the SC1CC and a random cell order, respectively. As expected, the fitted expression lines under random ordering of the cells convey no recognizable pattern and stay nearly flat close to an altitude of 0. In contrast, the SC1CC cell order results in fitted curves that appear to peak at different positions, consistent with these gene's involvement in different cell cycle phases.

3   Results and Discussion

3.1   Results on the hESC Dataset

The cell order inferred by SC1CC's OLO algorithm and the cell cycle order reconstructed by reCAT are shown in Fig. 4a. SC1CC groups together almost
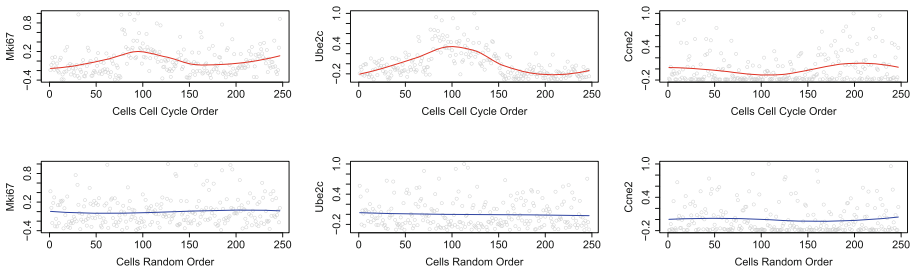

Fig. 3. Example cell cycle genes in the hESC dataset. Normalized expression levels for select cell cycle genes and cells ordered by SC1CC (red) vs. a shuffled cells order (blue). Different cell cycle genes follow different patterns of expression along the cell cycle phases. Given the SC1CC inferred cell order, which reflects the cells' progression through the cell cycle, different patterns for individual cell cycle genes can be seen for different genes associated with the cell cycle, including Mki67, Ube2c, Ccne2
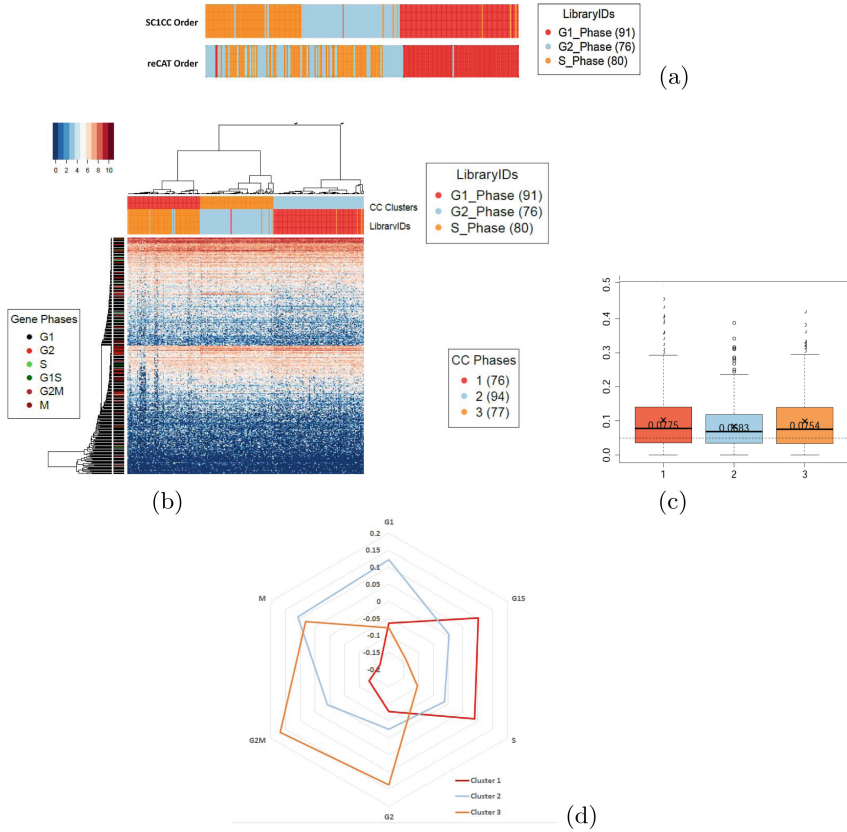
Fig. 4. Cell cycle analysis of the hESC dataset. (a) hESC cells orders inferred by SC1CC and reCAT. Experimentally determined cell labels are color coded as library IDs. SC1CC groups the majority of cells from each phase together (G1 in red, G2 in blue, and S in orange), whereas only G1 cells are grouped together in the reCAT order. (b) Heatmap of $log_2(x+1)$ expression values of cell cycle genes for the hESC cells using 'GO:0007049' gene list ordered according to SC1CC. Colors in the top barlabeled 'CC Clusters' represent the identified cell cycle clusters according to SC1CC, whereas the colors in the 'Library IDs' bar of the heat map indicate the cell cycle phases determined by FACS. (c) GSS for the three clusters identified by running SC1CC on the hESC cells. (d) Mean-scores for each of the three clusters identified by SC1CC and each of the six considered cell cycle phases.

all cells labeled with the same phase. Although the reCAT order maintains the grouping of G1 cells, cells from S and G2 phases are highly interleaved in this order. The SC1CC order also has a higher GSS score of 0.0632 compared to 0.0519 for the reCAT order.

Figure 4b displays the heat map of $log_2(x+1)$ expression values of cell cycle genes for the hESC cells ordered according to SC1CC. Colors in the top bar

labeled 'CC Clusters' represent the identified cell cycle clusters according to SC1CC, whereas the colors in the 'Library IDs' bar of the heat map indicate in this case the cell cycle phases determined by FACS. Note that the colors for library IDs and inferred cell cycle clusters are assigned independently from the same color palette in our online implementation and therefore are not necessarily in one-to-one correspondence. The heat map in Fig. 4b was generated by running SC1CC using the set of genes associated with the GO term "Cell Cycle" (GO:0007049). The hierarchical clustering algorithm implemented by SC1CC identifies three clusters. The GSS scores (Fig. 4c) for the three clusters were 0.0775, 0.0683, and 0.0754, respectively, indicating that all clusters consist of dividing cells, as expected.

Figure 4d gives the mean scores for each of the three clusters identified by SC1CC and each of the six considered cell cycle phases. Based on majority matching of cell labels determined by FACS, the three clusters are comprised of cells in the S, G1, and G2 phases, respectively. Albeit not perfect, the cluster assignments based on peak mean scores have good agreement. Specifically, cluster 1 (consisting of S phase cells according to the FACS labels) has very close highest mean scores for the G1S and S phases, with the G1S score slightly higher. Cluster 2 (G1 according to FACS) has two close highest scores for G1 and M phases, with the G1 score slightly higher. Finally, cluster 3 (G2 according to FACS) has two close highest mean scores for the G2 and G2M phases, with the G2M score slightly higher. The relatively low number of cells as well as the limited resolution of the ground truth labels are both likely contributing factors to the near-ties in peak score assignments for the three clusters.

In Table 1 we compare the clusters (cell labels) generated by Cyclone with the clusters inferred by SC1CC using different cell cycle gene sets for the hESC dataset. We assess clustering accuracy using the macro and micro-accuracy measures from [10] and [18], defined as:

$$Micro\ Accuracy = \sum_{i=1}^{K} C_i / \sum_{i=1}^{K} N_i \tag{2}$$

$$Macro\ Accuracy = \frac{1}{K} \sum_{i=1}^{K} \frac{C_i}{N_i} \tag{3}$$

where K is the number of classes, N_i is the size of class i, and C_i is the number of correctly labeled samples in class i relative to the ground truth.

Both Cyclone and SC1CC cluster the cells with high accuracy, with Cyclone scoring slightly higher. As detailed in Sect. 2.2, SC1CC gives the user the choice to use three different lists: genes included in the Cyclebase 3.0 database [15], genes annotated with the "cell cycle" term (GO:0007049) in the Gene Ontology database [6], and the list of periodic genes identified from single cell data in [5]. As can be seen in Table 1, the genes associated to the term "Cell Cycle" (GO:0007049) from the The Gene Ontology (GO) database achieve slightly higher clustering micro- and macro-accuracy for the hESC dataset compared to the other two gene sets.

Table 1. Clustering accuracy for Cyclone and SC1CC run with three different gene lists on the hESC dataset.

| | Cyclone | SC1CC | | |
		Cyclebase 3.0	GO	Periodic Genes
G1-Phase	**1.000**	0.9890	0.9890	**1.000**
G2-Phase	0.9342	0.7895	**0.9605**	0.8290
S-Phase	**1.000**	**1.000**	0.9125	0.9875
Micro accuracy	**0.9798**	0.9312	0.9555	0.9433
Macro accuracy	**0.9781**	0.9262	0.9540	0.9388

3.2 Results on the PBMC Dataset

As described in Sect. 2.1, the PBMC dataset is expected to include mostly non-dividing cells, which is confirmed by the results of the SC1CC analysis. Figure 5a shows the heat map of the PBMC cells featuring the GO cell cycle related genes that are expressed in the dataset (using $log_2(x+1)$ expression) and the clustering obtained by SC1CC. The majority of the genes have low expression levels in most cells. Furthermore, the GSS scores of all clusters fall below the 0.05 cutoff and hence they are all labeled as non-dividing by SC1CC (Fig. 5b), as expected. Cyclone labels 2,192 of the 2,882 cells in the PBMC dataset as G1, 398 as G2M, and 292 as S phase cells, underscoring the need for a separate analysis step to determine if the cells are actually cycling.

3.3 Results on the α-CTLA-4 Dataset

As discussed in Sect. 2.1, the α-CTLA-4 dataset is expected to include a mix of dividing and non-dividing cells. This is the most likely scenario for many scRNA-Seq datasets where no knowledge of the cell cycle effect within the data is available *a priori*. We reasoned that the best analysis approach for such data is to perform a two stage analysis, where we first separate the dividing from the non-dividing cells, followed by a detailed cell cycle analysis of the potentially dividing cells identified in the first step. Indeed, after clustering and ordering the cells using SC1CC, we are able to distinguish the potentially dividing cells by their GSS score. Figure 6a shows the $log_2(x + 1)$ expression heat map of the 3,148 α-CTLA-4 cells passing the default QC described in Sect. 2.1 based on the GO cell cycle genes and using the first 4 principal components. One cluster (cluster 7 in light green color) consists of 193 cells that show markedly higher expression levels for the cell cycle genes. Independent clustering analysis based on full gene expression profiles performed using the SC1 pipeline shows that cluster 7 is comprised mostly of lymphoid cells (light blue in the horizontal bar labeled "Clusters" in the heat map). This cluster has the highest GSS score, exceeding the SC1CC detection threshold for dividing cells, as shown in Fig. 6b. Indeed, this cluster closely matches the "Mki67-Hi" cluster identified in [7] as
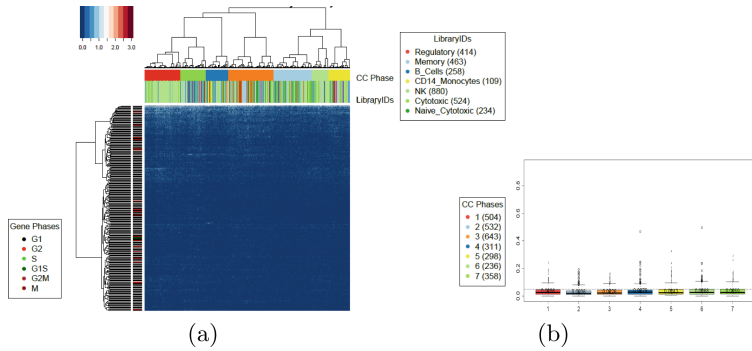
Fig. 5. Cell cycle analysis of the PBMC dataset. Heat map and cell order (a) along with GSS scores for the clusters inferred by SC1CC (b) on the PBMC dataset. GSS scores for all clusters fall below a cutoff of 0.05 and are labeled as non-dividing by SC1CC.

consisting of highly proliferative lymphoid cells. Further SC1CC analysis of the 193 cells in this cluster based on the Cyclebase 3.0 gene list reveals three sub-clusters (Fig. 6c), all of which are found to be actively dividing according to GSS scores (Fig. 6d). Cell cycle phase assignments based on maximum mean-scores suggests that the three sub-clusters consist of cells in the M, S, and G1S phases, respectively (Fig. 6e). For the sake of completeness we also tested Cyclone classification method on the of the 3,148 α-CTLA-4 cells, and 2,957 cells were labeled as G1, 149 were labeled as G2M, and 42 were labeled as S phase cells.

3.4   Results on the mHSC Dataset

As discussed in Sect. 2.1, this dataset also includes a mix of dividing and non-dividing cells. As with the α-CTLA-4 dataset analysis, we followed a two stage SC1CC analysis approach, where we first separate the dividing from the non-dividing cells, followed by a detailed cell cycle analysis of the potentially dividing cells identified in the first step. In excellent agreement with the percentages reported in [9], the first analysis stage (Fig. 7a–b) places 472 of the 1,277 mHSC cells (36.96%) in a dividing cluster with GSS score of 0.3075, and the remaining cells in a non-dividing cluster with GSS score of 0.0285. Furthermore, as shown in Table 2, the percentage of dividing cells identified by SC1CC among the three cell types identified in [9] are indeed approximately equal in young and old mice, with the exception of long term HSC, only 13% of which are dividing in old mice compared to 35% in young mice.

The analysis in [9] 'roughly' estimates the percentage of cells in G1, G1/S and G2M phases as 20%, 6% and 9% respectively. Following the second stage of analysis, where cycling cells identified in first stage are further clustered and assigned cell cycle phases based on peak mean scores (Fig. 7c–e), SC1CC identifies 217 cells (17% of total) as G1, 68 cells as G1/S (5.3% of total), and 187 cells as G2M (14.6% of total).
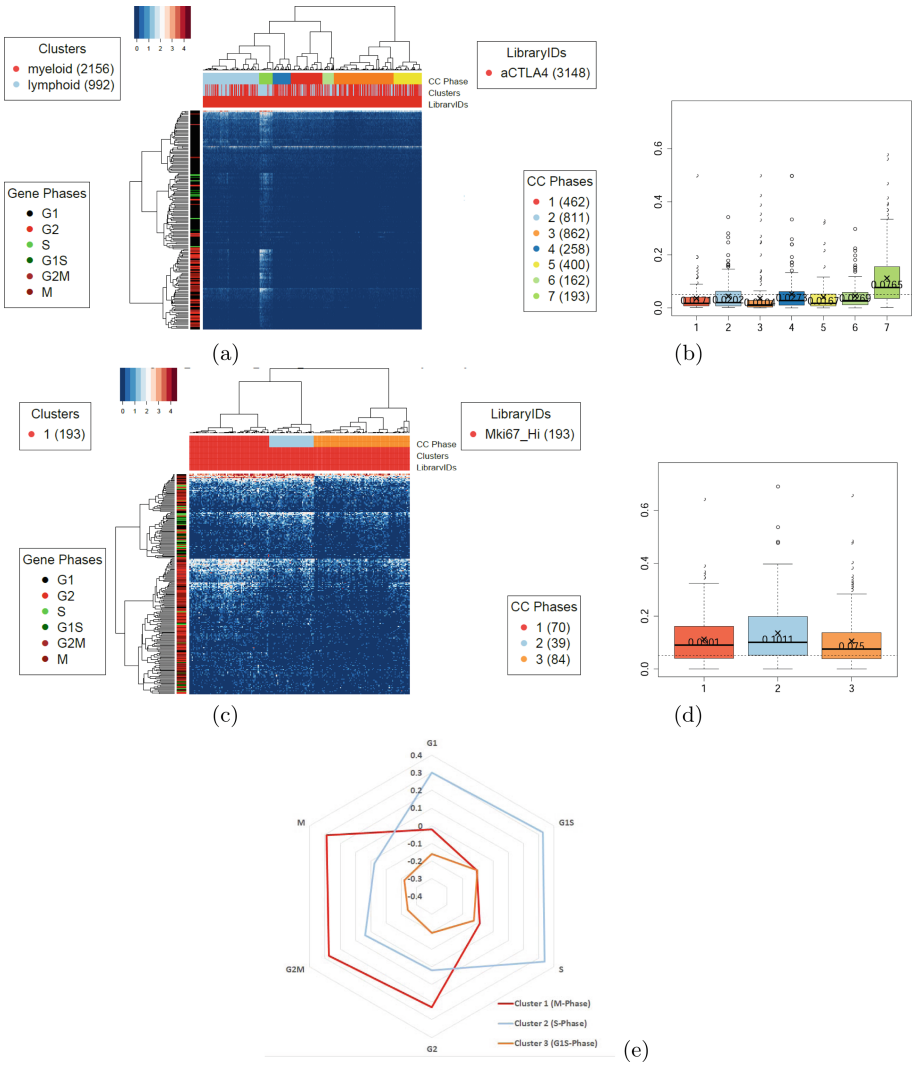
(a)

(b)

(c)

(d)

(e)

Fig. 6. Cell cycle analysis of the α-CTLA-4 dataset. Heat map and cell order (a) along with GSS scores for the clusters inferred by SC1CC (b) on the α-CTLA-4 dataset. The 193 cells in cluster 7 are further partitioned by SC1CC into 3 sub-clusters (c), all of which are marked as actively dividing based on GSS scores (d). Mean-scores for each of the three sub-clusters of dividing cells and each of the six considered cell cycle phases are given in (e). Mean-scores for each of the three sub-clusters of dividing cells and each of the six considered cell cycle phases are given in (e). The maximum mean-scores of sub-clusters 1 (red), 2 (blue), and 3 (orange) are achieved for the M, S, and G1/S phases, respectively.
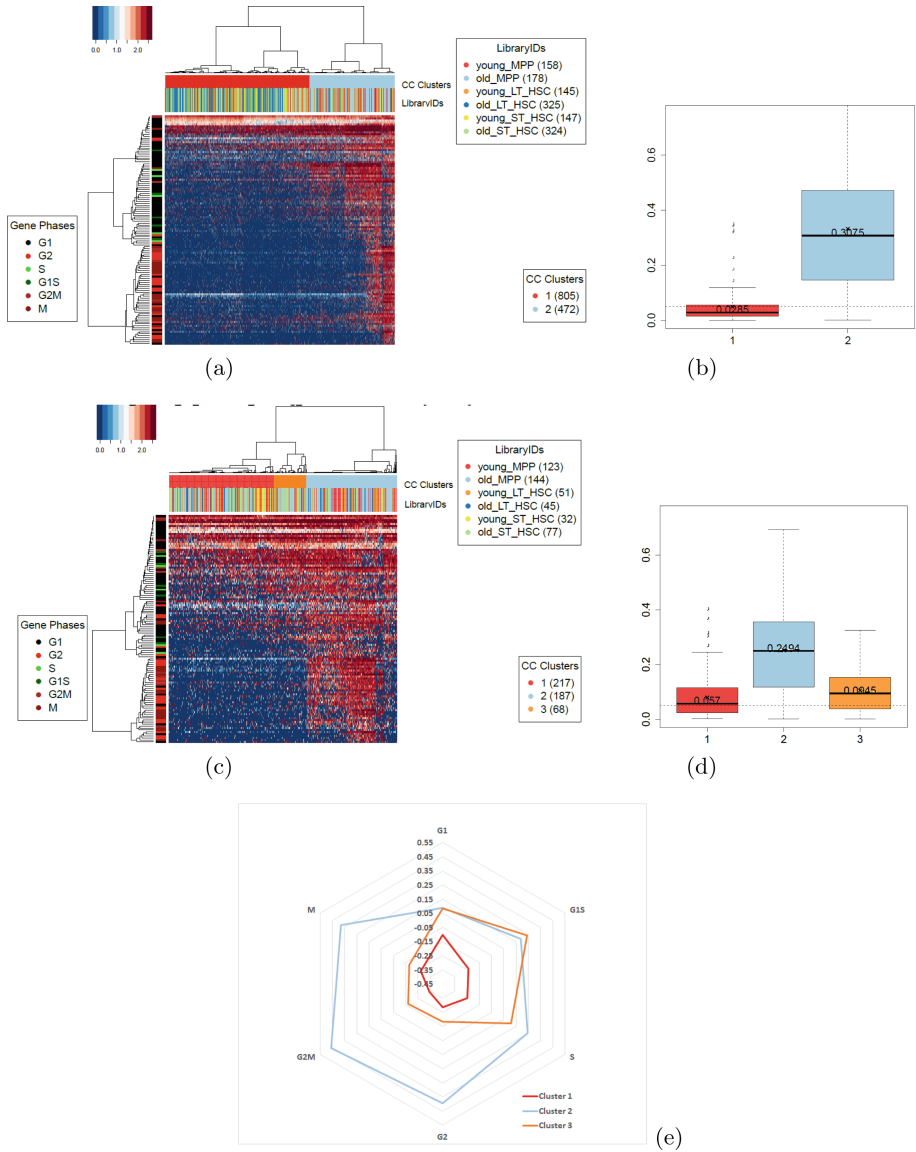
Fig. 7. Cell cycle analysis of the mHSC dataset. Heat map and cell order (a) along with GSS scores for the clusters inferred by SC1CC (b) on the mHSC dataset. The 472 cells in cluster 2 are further partitioned by SC1CC into 3 sub-clusters (c), all of which are marked as actively dividing based on GSS scores (d). Mean-scores for each of the three sub-clusters of dividing cells and each of the six considered cell cycle phases are given in (e). The maximum mean-scores of sub-clusters 1 (red), 2 (blue), and 3 (orange) are achieved for the G1, G2M, and G1/S phases, respectively.

Table 2. The inferred numbers of dividing vs. non-dividing cells in the six cell populations of the mHSC dataset.

	Non-Dividing	Dividing	% Dividing
Old mice MPP	34	144	80%
Young mice MPP	35	123	78%
Old mice ST-HSC	247	77	24%
Young mice ST-HSC	115	32	22%
Old mice LT-HSC	280	45	13%
Young mice LT-HSC	94	51	35%

4   Conclusion

In this paper we introduce SC1CC, a novel method for clustering and ordering single cell transcriptional profiles according to their cell cycle phase. The main contributions include a novel technique for ordering cells based on hierarchical clustering and optimal leaf ordering, and a new GSS metric based on serial correlation for assessing gene expression change smoothness along a reconstructed cell order as well as differentiating between cycling and non-cycling groups of cells. While many of the existing methods focus on a specific aspect of scRNA-Seq cell cycle analysis (e.g., assigning phase labels, ordering the cells, or removing the cell cycle contribution to gene expression), SC1CC is, to our best knowledge, the first method that enables a comprehensive analysis of the cell cycle effects, addressing four complementary analysis aspects. SC1CC differentiates between dividing and non-dividing cells, clusters the cells based on cell cycle effects independently from cell type effects, while also assigning cell cycle phases to the resulting clusters and ordering the cells based on their progression along the cell cycle phases. SC1CC has been implemented in R and deployed via a user-friendly interactive interface as part of the SC1 scRNA-Seq analysis pipeline [13], freely accessible at https://sc1.engr.uconn.edu/.

Empirical evaluation experiments on a diverse set of real scRNA-Seq datasets show that the GSS robust evaluation metric which allows distinguishing with high accuracy between dividing and non-dividing cells based on minimal assumptions about the underlying cell cycle gene expression changes. In direct comparisons with the existing specialized tools, SC1CC also achieves similar or better accuracy for clustering the cells according to cell cycle phases or ordering them according to the progression along the cell cycle phases. Importantly, SC1CC analysis is performed orthogonally to cell type identification, avoiding potentially artifacts of sequential analysis advocated in [2].

References

1. Bar-Joseph, Z., Gifford, D.K., Jaakkola, T.S.: Fast optimal leaf ordering for hierarchical clustering. Bioinformatics **17**(suppl_1), S22–S29 (2001)
2. Barron, M., Li, J.: Identifying and removing the cell-cycle effect from single-cell RNA-sequencing data. Sci. Rep. **6** (2016)
3. 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**(2), 155–160 (2015)
4. Cooper, G.M., Hausman, R.E., Hausman, R.E.: The Cell: A Molecular Approach, vol. 10. ASM Press, Washington DC (2000)
5. Dominguez, D., Tsai, Y.H., Gomez, N., Jha, D.K., Davis, I., Wang, Z.: A high-resolution transcriptome map of cell cycle reveals novel connections between periodic genes and cancer. Cell Res. **26**(8), 946 (2016)
6. Gene Ontology Consortium: the gene ontology (GO) database and informatics resource. Nucleic Acids Res. **32**(suppl_1), D258–D261 (2004)
7. Gubin, M.M., Alspach, E., et al.: High-dimensional analysis delineates myeloid and lymphoid compartment remodeling during successful immune-checkpoint cancer therapy. Cell **175**(4), 1014–1030 (2018)
8. Hahsler, M., Hornik, K., Buchta, C.: Getting things in order: an introduction to the R package seriation. J. Stat. Softw. **25**(3), 1–34 (2008)
9. Kowalczyk, M.S.: Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res. **25**(12), 1860–1872 (2015)
10. Lee, C., Măndoiu, I.I., Nelson, C.E.: Inferring ethnicity from mitochondrial DNA sequence. In: BMC Proceedings. vol. 5, p. S11. BioMed Central (2011)
11. Leng, N., et al.: Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. Nat. Methods **12**(10), 947 (2015)
12. Liu, Z., et al.: Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. Nat. Commun. **8**(1), 22 (2017)
13. Moussa, M., Măndoiu, I.I.: SC1: a tool for interactive web-based single cell RNA-seq data analysis. In: Cai, Z., Mandoiu, I., Narasimhan, G., Skums, P., Guo, X. (eds.) ISBRA 2020. LNCS, vol. 12304, pp. 389–397. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-57821-3_39
14. Moussa, M., Mandoiu, I.: Single cell RNA-seq data clustering using TF-IDF based methods. BMC-Genomics **19**(Suppl 6), 569 (2018)
15. Santos, A., Wernersson, R., Jensen, L.J.: Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. Nucleic Acids Res., gku1092 (2014)
16. Scialdone, A., et al.: Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods **85**, 54–61 (2015)
17. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. J. Royal Stat. Society Ser. B (Stat. Methodol.) **63**(2), 411–423 (2001)
18. Van Asch, V.: Macro-and micro-averaged evaluation measures. Technical report (2013)
19. Zheng, G.X.Y., et al.: Massively parallel digital transcriptional profiling of single cells. Nat. Commun. **8**, 14049 (2017)