


RESEARCH

Open Access



Universal prediction of cell-cycle position using transfer learning

Shijie C. Zheng¹, Genevieve Stein-O'Brien^{2,3,4,5}, Jonathan J. Augustin², Jared Slosberg², Giovanni A. Carosso², Briana Winer², Gloria Shin², Hans T. Bjornsson^{2,6,7,8}, Loyal A. Goff^{2,3,4*} and Kasper D. Hansen^{1,2*} 

*Correspondence:

loyalgoff@jhmi.edu;

khansen@jhsp.h.edu

¹Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

²Department of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, USA

³Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, USA

Full list of author information is available at the end of the article

Abstract

Background: The cell cycle is a highly conserved, continuous process which controls faithful replication and division of cells. Single-cell technologies have enabled increasingly precise measurements of the cell cycle both as a biological process of interest and as a possible confounding factor. Despite its importance and conservation, there is no universally applicable approach to infer position in the cell cycle with high-resolution from single-cell RNA-seq data.

Results: Here, we present tricycle, an R/Bioconductor package, to address this challenge by leveraging key features of the biology of the cell cycle, the mathematical properties of principal component analysis of periodic functions, and the use of transfer learning. We estimate a cell-cycle embedding using a fixed reference dataset and project new data into this reference embedding, an approach that overcomes key limitations of learning a dataset-dependent embedding. Tricycle then predicts a cell-specific position in the cell cycle based on the data projection. The accuracy of tricycle compares favorably to gold-standard experimental assays, which generally require specialized measurements in specifically constructed in vitro systems. Using internal controls which are available for any dataset, we show that tricycle predictions generalize to datasets with multiple cell types, across tissues, species, and even sequencing assays.

Conclusions: Tricycle generalizes across datasets and is highly scalable and applicable to atlas-level single-cell RNA-seq data.

Keywords: Cell cycle, Single-cell RNA-sequencing, Transfer learning

Background

The cell cycle is the biological process which controls faithful replication and division of cells across all species of life. Despite existing as a continuous process, cell cycle has historically been characterized as having four discrete stages during which the cell performs growth and maintenance (G1), replicates its DNA (S), increases further in size and prepares for mitosis (G2), and undergoes mitosis and cytokinesis (M). Cell cycle is a highly conserved mechanism with an integral role in generating the diversity of cell types within



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

multicellular organisms. As a result, maladaptive modifications of the cell cycle have devastating consequences in development and disease [1–3]. Despite its importance, many of the molecular mechanisms regulating and interacting with cell cycle remain poorly understood.

High-throughput expression data has been utilized for studying the cell cycle since the seminal work on the yeast cell cycle by Spellman et al. [4] and Cho et al. [5] at the dawn of the microarray era. This work used various approaches to synchronize cells in specific cell-cycle stages followed by assaying cells in bulk. The data from Spellman et al. [4] were later used by Alter et al. [6] to show that principal component analysis reveals a circular pattern which represents the cyclical nature of the cell cycle; widely cited as one of the first examples of the use of principal component analysis and singular value decomposition in analysis of high-throughput expression data. Subsequent work sought to systematically identify both periodically expressed genes and cell-cycle marker genes and deposited these into widely used databases [7, 8].

Single-cell technologies have enabled the ability to study the effects of cell cycle in multicellular organisms with a degree of sensitivity and accuracy only previously available in monocellular or clonal systems. Thus, cell cycle has been the subject of substantial interest, both as a biological variable of interest and as a possible confounding feature for other comparisons of interest [9]. A number of methods have been developed to estimate cell cycle state from single-cell expression data [10–15]; some of these approaches are related to more general methods for finding topological structure in single-cell data [16]. These methods differ broadly in the definition of cell-cycle state (discrete stages vs. continuous pseudotime) as well as the use of special training data. Most of these methods have been demonstrated to be effective on datasets consisting of a single cell type. Despite the conservation of the cell-cycle process, none of these methods have been shown to be applicable across single-cell technologies and mammalian tissues.

Results

Transfer learning

To develop a universal method for estimating a continuous cell-cycle pseudotime for a single-cell expression data set independent of technology, cell type, or species, we leverage transfer learning via dimensionality reduction [17, 18]. We define a reference cell-cycle embedding (or latent space) into which we project a new data set, an approach originally advocated for in Stein-O'Brien et al. [19]. After projection, we infer cell-cycle pseudotime as the polar angle around the origin. This pseudotime variable takes values in $[0, 2\pi]$ and is unrelated to wall time (time measured by a clock in SI units), but rather represents progression through the cell-cycle phases. We refer to this pseudotime variable as cell cycle position to avoid confusion with wall time and to emphasize its periodic nature.

To define a reference cell-cycle embedding, we leverage key features of principal component analysis of cell-cycle genes. Previous work has found that principal component analysis on expression data sometimes yields an ellipsoid pattern. This was first described by Alter et al. [6]; it has later been observed independently in multiple data sets [12, 15, 20]. Here, we draw attention to the fact that the ellipsoid pattern is a consequence of a link between Fourier analysis of periodic functions and principal component analysis which links the progression through the cell-cycle process with angular position on the ellipsoid. This is similar in spirit to previous observations of principal compo-

ment analysis of genotype data [21] and connected to mathematical results on circulant matrices [22].

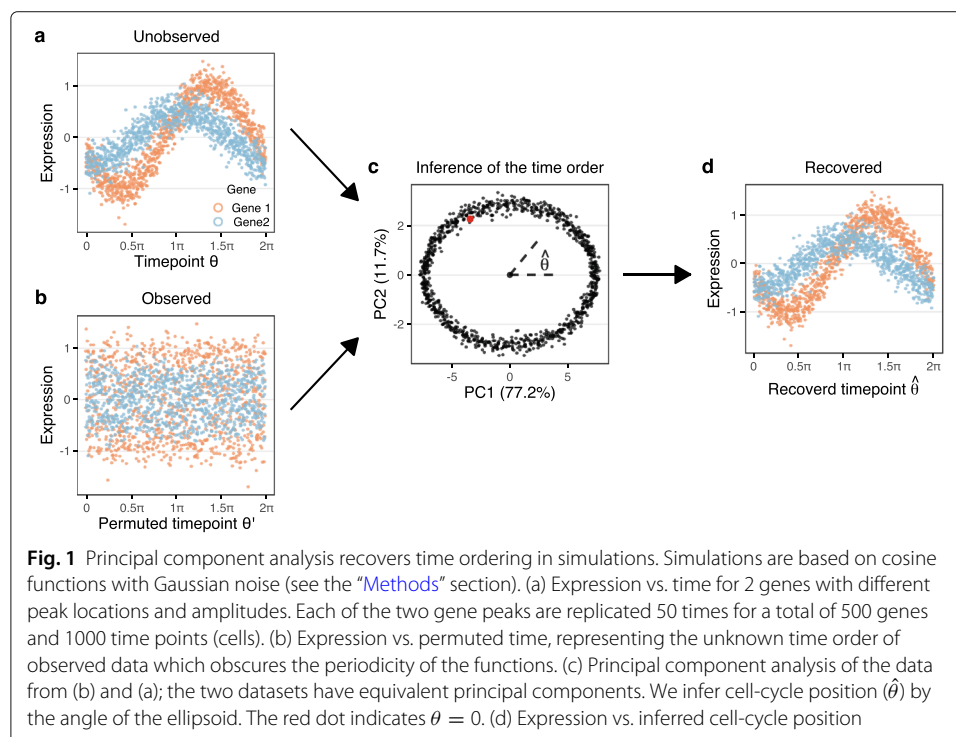
We use the first two principal components to define a reference embedding representing the cell cycle. Because this reference embedding is a low-dimensional linear space, we obtain an orthogonal projection operator allowing us to project any new data set into the reference embedding. We show that projecting new data into the reference cell-cycle embedding overcomes technical and biological challenges posed by data sets where substantial variation is explained by one or more factors different from cell cycle, such as cellular differentiation.

Principal component analysis and periodic functions

To gain insight into gene expression dynamics over the cell cycle, we start by analyzing principal component analysis of periodic functions. Our model is a collection of periodic functions with a single peak, taking the form

$$x_g(\theta) = A_g \cos(\theta - L_g)$$

with a gene-specific amplitude (A_g) and location of the peak (L_g) with $0 \leq \theta < 2\pi$ representing the unknown cell-cycle position. Figure 1 (a, b) depicts the unobserved (true) time ordering, observed on a discrete grid of time points, together with a random permutation of these time points; this represents the observed data which is not ordered by time. A key insight is the fact that the first two principal components are the same for the observed and the unobserved data (Fig. 1 (c)), when performed on a discrete set of observation times. The unknown time order can be inferred from the principal component plot as the angle of each point, making it possible to fully reconstruct the unobserved time order (Fig. 1 (d)), i.e., the first two principal components form an orthogonal projection



into a two-dimensional space representing the periodic time. Our “Methods” section contains a constructive proof in the case of single peak genes, which are particularly relevant for cell-cycle expression.

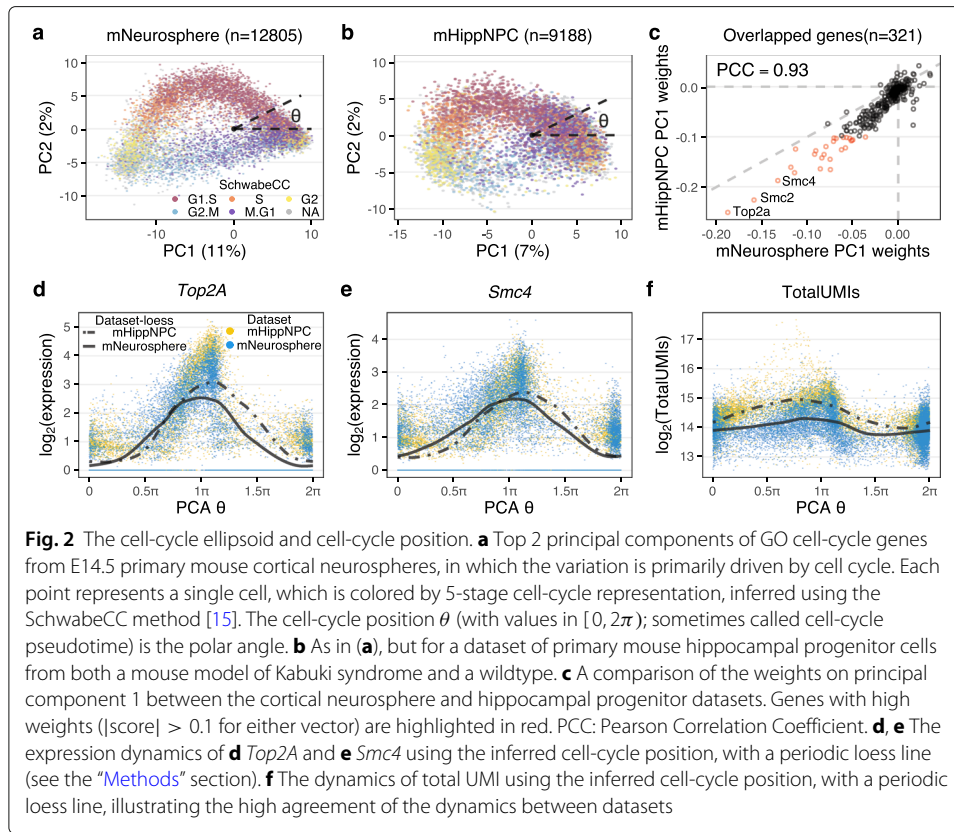
The simulated data depicted in Fig. 1 has Gaussian noise, but we have verified that the result holds for data generated using the negative binomial distribution with an associated mean-variance relationship. In addition to using the negative binomial distribution, we also mimic the library size normalization step usually performed in scRNA-seq data analysis, which imposes a constraint on the total counts. With the total counts constraint, we need at least 3 distinct peak locations to be stable (Additional file 1: Figs. S1, S2). For both distributions, this approach is robust to downsampling of the data similar to what is seen with the increased sparsity from droplet-based sequencing technology. In simulations, we can recover cell-cycle position with as little as 10 total counts per cell across 100 genes (depending on noise levels and heights of the peaks) (Additional file 1: Fig. S3).

Recovering cell-cycle position using principal component analysis on cell-cycle genes

We next show that PCA of cell-cycle genes form an ellipsoid as predicted by the cosine model presented in the previous section, and learn an embedding representing cell cycle. We use 10x Genomics Chromium single-cell RNA-sequencing (scRNA-seq) data on two replicate cultures of E14.5 mouse cortical neurospheres (see the “Methods” section), integrated using Seurat 3 and transformed to \log_2 -scale. The use of an alignment method (CCA in Seurat3) to integrate the two samples is important for the quality of the ellipsoid, by maximizing the correlation structure between the two samples. Since neurospheres are maintained in a proliferative state, we expect that cell-cycle phase is an important contributor to the variation in expression within this single-cell dataset. To confirm this expectation, we consider a UMAP representation of the data based on all variable genes (Additional file 1: Fig. S4) colored according to the predictions from two separate cell-cycle stage estimation utilities (cyclone and a modification of Schwabe et al. [15] we call SchwabeCC, see the “Methods” section); this analysis demonstrates that the cell cycle is a major source of transcriptional variation in the neurosphere dataset.

We then perform principal component analysis of the top 500 most variable genes amongst the roughly 1700 genes annotated with the Gene Ontology cell cycle term (*GO:0007049*; see the “Methods” section) [23]. The first two principal components form an ellipsoid with a sparse/empty interior (Fig. 2a); this reflects the behavior of our cosine model. Using the SchwabeCC cell-cycle stage predictor, we observe a strong relationship between polar angle on the ellipsoid and predicted cell-cycle stage.

The strong relationship between polar angle on the ellipsoid and predicted cell-cycle stage was also observed on an independent dataset on cultured primary mouse hippocampal progenitors from a wild-type mouse as well as from a *Kmt2d*^{+/ β geo} mouse, a previously described model of Kabuki syndrome [24]. The data were processed similarly to the neurosphere data. Again, we select the top 500 most variable cell-cycle genes and perform a principal component analysis (Fig. 2b) which reveal an ellipsoid pattern. The shape of the principal component plot differs between the two datasets, but the weights used to form the first two principal components are highly concordant (Fig. 2c, Additional file 1: Fig. S5 for PC2) for the 321 genes present in both cell-cycle embeddings. Almost all of the highly ranked genes (absolute weights > 0.1, highlighted in red and labeled with gene name) represent important regulators of, or participants in, the cell



cycle. For example, the highest ranked gene is *Top2A* (Topoisomerase 2A) which controls the topological state of DNA strands and catalyzes the breaking and rejoining of DNA to relieve supercoiling tension during DNA replication and transcription [25]. Also highly ranked are *Smc2* and *Smc4* which compose the core subunits of condensin, which regulates chromosome assembly and segregation [26, 27].

Given our mathematical analysis as well as the strong empirical relationship between polar angle on the ellipsoid and cell-cycle stage predictions, we define a method to learn cell-cycle position as the polar angle around the origin on the coordinate plane which we denote by θ . We center the coordinate plane on $(0, 0)$. Because we mean-center the expression data prior to performing PCA, this location corresponds to having mean (across the dataset) expression for all 500 variable cell-cycle genes. Linking to our idealized example with periodic functions, this will correspond to a timepoint (cell) where all genes have 0 expression after mean centering.

To demonstrate that cell-cycle position reflects the true biological cell-cycle progression, we consider expression dynamics of specific cell-cycle genes. For *Top2A* and *Smc4* the peak expressions are observed at G2 stage around π (Fig. 2e), consistent with their known increased expression through S phase and into G2 [27–29]. Furthermore, the dynamics are highly similar between the independently analyzed cortical neurosphere and hippocampal NPC datasets, which supports the observation that the two different embeddings yield concordant cell-cycle positions (despite each including dataset-specific genes). These observations hold for all genes with high weights (Additional file 1: Fig. S6).

This approach serves as an internal control in any single-cell RNA-seq data set and can be used to assess the quality of any continuous ordering.

Next, we directly relate estimated θ to the measured transcription values. Figure 2f shows the \log_2 transformed total UMI numbers against θ , with a periodic loess smoother for each dataset. In both datasets, the maximum level is reached around π and the minimum around 1.5π , which corresponds to the end of G2 and the middle of M stage respectively. We observe the total UMI number begins to increase at the beginning of G1/S phase and to decrease sharply as cells progress through M phase. The difference between the maximum and minimum of the periodic loess line is 1, corresponding to a twofold difference in total UMI, which is known to be proportional to cell size [30, 31]. This observation, and the timing with respect to cell-cycle position, is consistent with the approximate reduction in cellular volume by one half as a result of cytokinesis in M phase and the formation of two daughter cells of roughly equal size.

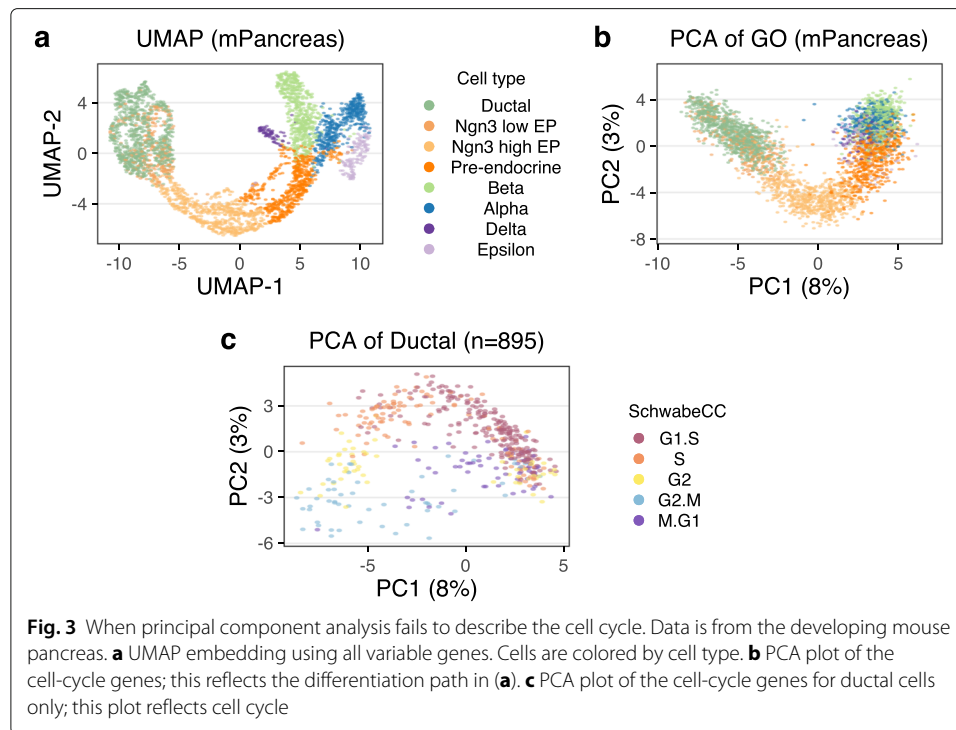
This approach of using expression dynamics and (if available) \log_2 -totalUMI can be used to evaluate whether *any* continuous position is related to the cell cycle. We name these “internal controls” and we note that they are available for *any* single-cell expression dataset. These internal controls will be used extensively throughout this manuscript.

Note that these principal component analyses are differentiating G2/M cells from G1/G0 cells on the first principal component. This contrasts with the mathematical analysis where the starting point ($\theta = 0$) can be any location (red point in Fig. 1) as there is no clear starting point for a periodic function. That the first principal component differentiates G2M from G1/G0 can be explained by the nature of principal component analysis. Before principal component analysis, we subtract each gene’s mean expression. However, genes marking G2/M usually have very high expression compared to other stages, with G0/G1 being the lowest (Additional file 1: Fig. S7), ensuring that this becomes the first principal component. A clustering analysis of the expression patterns provides further evidence that cell-cycle genes have a single peak pattern of expression (Additional file 1: Fig. S7). Thus, the observed behavior of the cell-cycle genes in these data sets satisfy the assumptions of the cosine model.

In summary, principal component analysis of the cell-cycle genes predicts cell-cycle progression for the mNeurosphere and mHippNPC datasets with a high degree of similarity between the cell-cycle position inferred independently in the two datasets. This observed behavior aligns with the results from the cosine model.

When principal component analysis fails to reflect cell-cycle position

A principal component analysis does not always yield an ellipsoid pattern; a requirement for this to work is for the first principal component to be dominated by cell cycle. To illustrate this, we used an existing mouse developing pancreas dataset, with cell type labels [32]. A major source of variation in this dataset is cellular differentiation as demonstrated by a standard UMAP embedding (based on all variable genes) illustrating the previously described [32] differentiation trajectories (Fig. 3a). When we perform principal component analysis using only the variable cell-cycle genes, the resulting PCA plot still reflects the differentiation trajectory and does not resemble the ellipsoid pattern observed in the previous section (Fig. 3b). Note that PC1 has some relationship with cell cycle since the differentiation path goes from cycling to non-cycling cells, but it also reflects the



progression from cycling multipotent cells to terminally differentiated cells. This result strongly suggests that some of the cell-cycle genes may participate in biological processes other than the cell cycle and demonstrates that PCA of cell-cycle genes does not always exclusively capture cell-cycle variance.

However, when we perform principal component analysis only on a subset of cells from a single, proliferating progenitor cell type, the ellipsoid pattern returns (Fig. 3c and Additional file 1: Fig. S8a,b). This highlights the challenge of inferring cell cycle for datasets that contain many different cell types, including postmitotic cells.

Transfer learning through projection

To overcome the challenges of inferring cell-cycle position in arbitrary datasets, we propose a simple, yet highly effective transfer learning approach we term tricycle (transferable representation and inference of cell cycle). In short, we first construct a reference embedding representing the cell-cycle process using a fixed dataset where cell cycle is the primary source of transcriptional variation. For the remainder of this manuscript, we will use the cortical neurosphere data as this reference, and we claim this is a useful reference regardless of dataset. We project new data into this embedding and infer cell-cycle position for each cell by the polar angle around the origin.

As a demonstration, we consider a diverse selection of single-cell RNA-seq datasets representing different species (mouse and human), cell types, and technologies (10x Chromium, SMARTer-Seq, Drop-seq, and Fluidigm C1) (Table 1). We project these datasets into the cell-cycle embedding learned from the neurosphere data (Fig. 4a, Additional file 1: Fig. S9a). To effectively visualize cell-cycle position defined as the polar angle, we use a circular color scale to account for the fact that position “wrap around” from 2π to 0. Although the shape of the projection varies from dataset to dataset, the cells of the

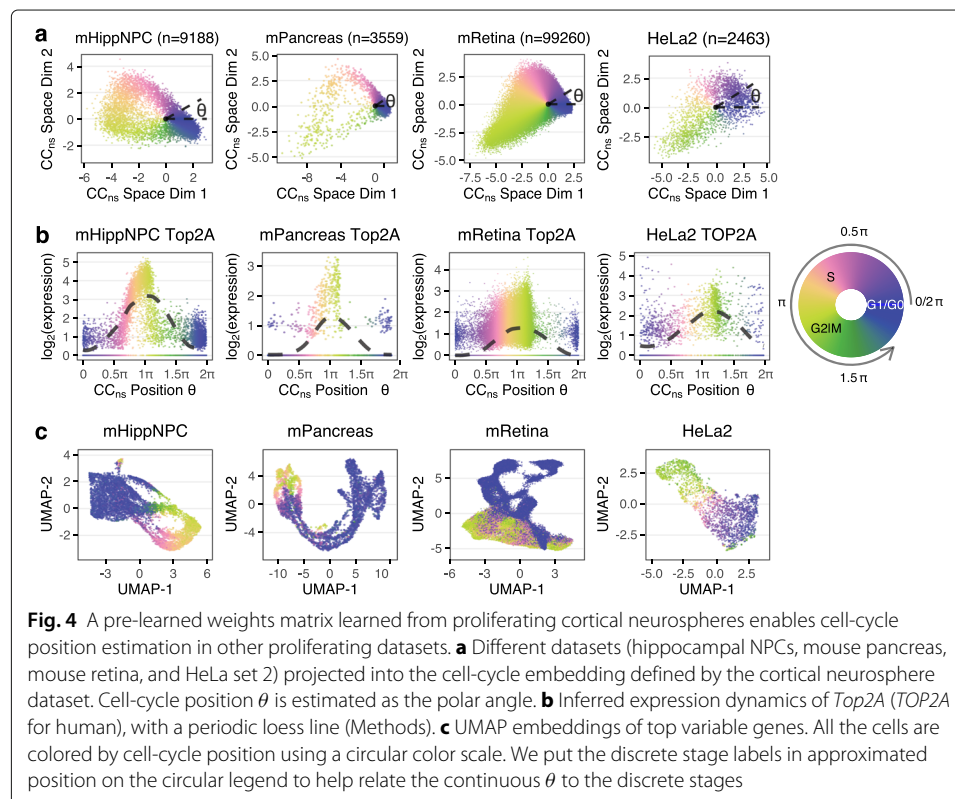
Table 1 Datasets

Dataset	Species	Platform	# cells	Note	Reference
mNeurosphere	Mouse	10x	12805		here
mHippNPC	Mouse	10x	9188		here
mPancreas	Mouse	10x	3559		[32]
mHSC	Mouse	SMARTer	1343		[33]
mRetina	Mouse	10x	99260		[34]
HeLa 1	Human	Drop-seq	1398		[15]
HeLa 2	Human	Drop-seq	2463		[15]
mESC	Mouse	Fluidigm C1	279	FACS	[9]
hESC	Human	Fluidigm C1	226	FACS	[10]
hU2OS	Human	SMART-seq2	1114	FUCCI	[20]
hiPSCs	Human	Fluidigm C1	888	FUCCI	[14]
Fetal tissue atlas	Human	sci-RNA-seq3	Varies		[35]

same stage always appear at a similar position of θ , such as cells at S stage centering at 0.75π .

To verify our cell-cycle ordering, we again use the internal controls as they exist in all datasets. Specifically, we show the expression dynamics of *Top2A* and *Smc4* as a function of θ (Fig. 4b, Additional file 1: Figs. S9b, S10). In contrast, PCA plots of the GO cell-cycle genes for each dataset illustrates the advantage of using a fixed embedding to represent cell cycle (Additional file 1: Fig. S11). Together, these results strongly support that tricycle generalizes across data modalities.

Having inferred cell-cycle position, we can visualize the cell-cycle dynamics on a UMAP plot representing the full transcriptional variation, as is standard in the scRNA-seq



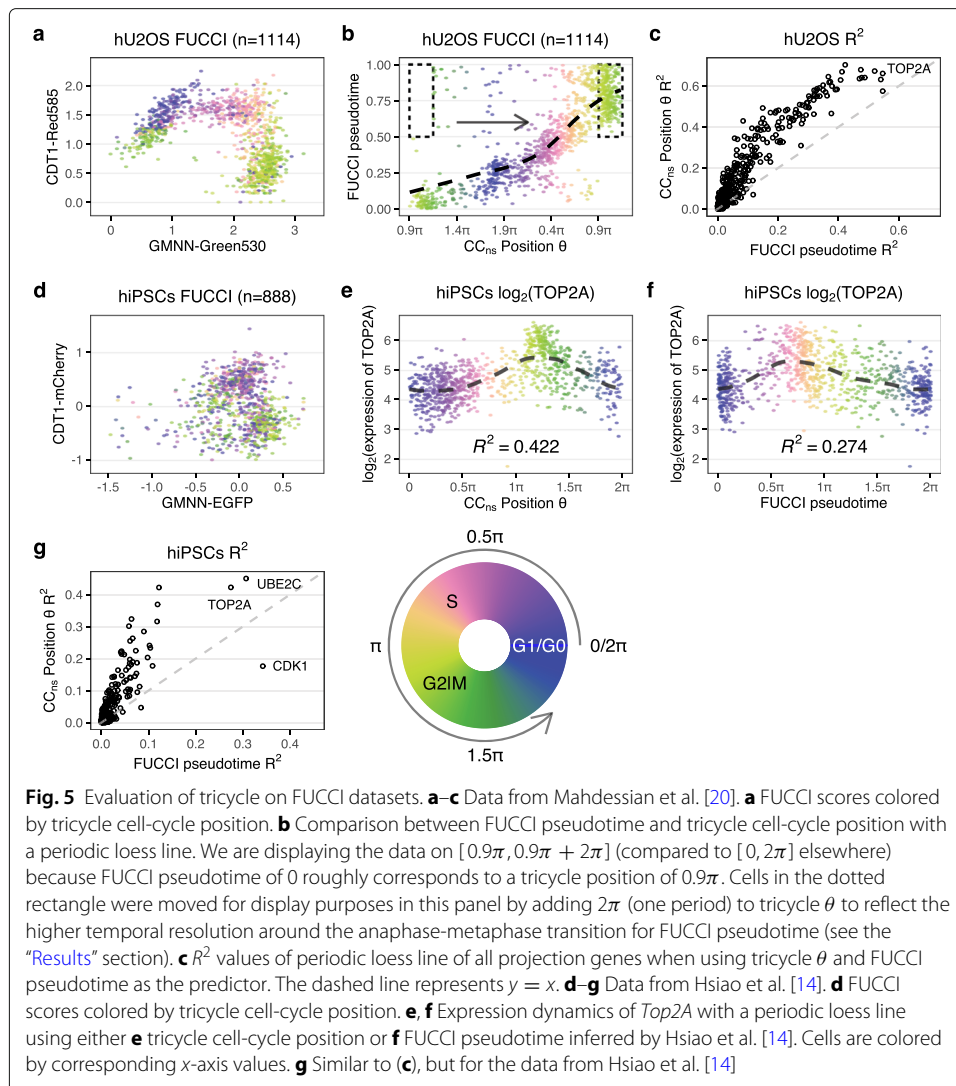
literature (Fig. 4c, Additional file 1: Fig. S9c). Doing so reveals the smooth behaviour of the tricycle predictions (despite not using smoothing or imputation) and argues for representing cell cycle in gene expression data as a continual progression rather than discrete states.

We draw attention to two specific datasets in these figures: the mPancreas and mRetina datasets (Fig. 4) and the HCA (human cell atlas) Pancreas data (Additional file 1: Fig. S9). These three datasets contain many different cell types and strong drivers of gene expression in addition to the cell cycle, such as differentiation. We have previously seen how principal component analysis fails to be ellipsoid on the mPancreas dataset (Fig. 3) and this example shows how tricycle—by projecting the data into a fixed reference embedding—overcomes the limitations of principal component analysis (Additional file 1: Fig. S12). Finally, note that the HCA Pancreas dataset is sparse with a median of 892 total UMIs per cell.

Cell-cycle position estimation on gold-standard datasets

We validated tricycle on multiple datasets containing “gold-standard” cell-cycle measurements, including measurements by proxy using the fluorescent ubiquitination-based cell-cycle indicator (FUCCI) system and by fluorescence-activated cell sorting (FACS) of cells into discrete cell-cycle stages. Both of these approaches allow for assignment to or selection of cells from discrete phases of the cell cycle. The FUCCI system uses a dual reporter assay in which the reporters are fused to two genes with dynamic and opposing regulation during the cell cycle [36], allowing for a quantitative assessment of whether cells are in G1 or S/G2/M phase. In contrast to FACS, FUCCI systems, combined with an appropriate quantification method, make it possible to continuously measure cell cycle progression by placing the 2 protein measurements in a 2-dimensional space. Cell-cycle pseudotime needs to be inferred from these 2-dimensional measurements, which is usually done by a variant of polar angle [14, 20].

Mahdessian et al. [20] measured human U-2 OS cells to derive a FUCCI-based pseudotime scoring. Their FUCCI measurements form a distinct horseshoe shape with the left side of the horseshoe representing time post-metaphase-anaphase transition with a continuous progression through G1, S, G2 and ending pre-metaphase-anaphase transition (Fig. 5; this depiction mirrors other data presentations [36, 37]). Cell cycle is a continuous process which is not immediately reflected in the horseshoe form because of the large gap (in the x -axis) between the two ends of the horseshoe. The x -axis reflects the protein levels of geminin (GMNN) which is degraded during the metaphase-anaphase transition [38] and the two “open” ends of the horseshoe are closely connected in time despite the visual gap in the scatterplot. This fact gives the FUCCI system the ability to assess whether a cell in M phase is before or after this transition, or said differently, a high temporal resolution around this transition despite the relatively short wall time compared to the rest of the cell cycle. We observe a close correspondence between tricycle cell-cycle position and FUCCI pseudotime (circular correlation coefficient $\rho = 0.70$). The only cells for which there is an apparent disagreement are placed in M phase by tricycle (cell-cycle position around 0.9π) and are split between pre-metaphase-anaphase transition and post-metaphase-anaphase transition by FUCCI pseudotime, for this particular transition the FUCCI system has higher temporal resolution than tricycle; adding a small offset to these cells results in a remarkable concordance between the two systems (Fig. 5). Elsewhere in



the cell cycle, there is no evidence of better temporal resolution with FUCCI; examining expression dynamics suggests that tricycle does at least as well as FUCCI at ordering key cell-cycle genes. We can use tricycle to examine the expression dynamics of *GMNN* and *CDT1* which reveals that *GMNN* expression is stable across the cell cycle (Additional file 1: Fig. S13), suggesting the protein is predominantly regulated post-transcriptionally during mitosis.

Hsiao et al. [14] used FUCCI on human induced pluripotent stem cells (iPSC) followed by scRNA sequencing using Fluidigm C1. While the Mahdessian et al. [20] FUCCI data look like a horseshoe, the Hsiao et al. [14] FUCCI data are more akin to a cloud (the data differ in quantification and normalization of the FUCCI scores). These data are used to estimate a continuous cell-cycle position (which we term “FUCCI pseudotime”) based on polar angle of the FUCCI scores. Compared with the data in Mahdessian et al. [20], there are larger differences between FUCCI pseudotime and tricycle cell-cycle position. However, we can directly compare the associated expression dynamics of key cell-cycle genes (Fig. 5 for *TOP2A*, Additional file 1: Fig. S14 for 8 additional genes). These results suggest that tricycle cell-cycle position is at least as good or better as the FUCCI

pseudotime at ordering the cells along the cell cycle; the R^2 for *TOP2A* is 0.42 for tricycle compared with 0.27 for peco.

In contrast to FUCCI measurements, FACS sorting and enrichment of cells yields groups of genes in (supposedly) distinct phases of the cell cycle. We consider 2 different datasets where FACS has been combined with single-cell RNA-seq. Buettner et al. [9] assays mouse embryonic stem cells (mESC) using Hoechst 33342-staining followed by cell isolation using the Fluidigm C1. They use very conservative gating for G1 and G2M at the cost of less conservative gating for S phase. Leng et al. [10] uses FACS on FUCCI-labeled H1 human embryonic stem cells (hESC) followed by cell isolation using the Fluidigm C1. In both experiments, cells largely appear as expected in the cell-cycle embedding defined by the cortical neurosphere reference embedding (Additional file 1: Fig. S15). For the mESC, we note that some cells labeled S (but not G1 or G2M) appear outside the position expected for this stage, consistent with the gating strategy used for these data.

Summarizing this evidence, we conclude that tricycle recapitulates and refines the cell cycle ordering consistent with current “state of the art” experimental methods. Tricycle cell-cycle position is competitive with FUCCI-based measurements, except for cells in the metaphase to anaphase transition during mitosis.

Comparison with existing tools for cell-cycle position inference

We next sought to compare tricycle cell-cycle position estimates with those obtained from other available methods. Existing methods for cell-cycle assessment can be divided into those which infer a continuous position and those which assign a discrete stage. We have evaluated the following methods: peco [14], Revelio [15], Oscope [10], reCAT [12], cyclone [11], Seurat [13], the original Schwabe [15], and the SchwabeCC 5 stage assignment method. Each method differs in which datasets it works well on and which issues it might have; a detailed comparison is available in the Supplementary Materials (Additional file 1: Supplementary Methods, and Figs. S16-S22).

Issues with existing methods include (a) the ability to work on datasets with multiple cell types, (b) the ability to scale to tens of thousands of cells or more, and (c) the ability to work on less information-rich datasets such as those generated by droplet-based or *in situ* scRNA-seq methods. Oscope requires data on many genes due to its use of pairwise correlations and therefore does not work on less information rich platforms (e.g. 10x Chromium or Drop-Seq). peco works better on less sparse and information-rich data (e.g. Fluidigm C1), but even on data from this platform, it is outperformed by tricycle. reCAT is critically dependent on the extent to which a principal component analysis of the cell-cycle genes reflects cell cycle and only infers a cell ordering; it is not straightforward to interpret the reCAT ordering, especially across datasets. Revelio is primarily a visualization tool, which appears to fail on datasets where substantial variation is driven by processes other than the cell cycle. Of the discrete predictors, Seurat agrees well with tricycle (and is very scalable) but is limited by only predicting a 3-stage cell-cycle representation (G1/S/G2M). Cyclone appears to do poorly in labeling cells in S phase and only predicts 3 stages. The SchwabeCC predictor assigns 5 stages, but has many missing labels and mis-assigns cells from G0/G1 to other stages.

Additionally, we benchmarked the computational speed and performance of tricycle against other cell-cycle estimation algorithms. We briefly compared the running time of several methods using subsets of the mRetina dataset (Additional file 1: Fig. S23). To

compute continuous estimates using tricycle takes a mean of about 0.58, 0.86, and 1.48 s when the number of cells is 5000, 10,000, and 50,000 respectively. In contrast, to compute finite discrete stages, Seurat takes a mean of about 1.10, 1.22, and 4.95 s for a three-stage estimation and cyclone takes a mean of about 7.96, 11.50, and 50.66 min for a three-stage estimation, when the number of cells is 5000, 10,000, and 50,000 respectively. Other methods (peco, Oscope, reCAT) are not capable of processing large (10k–100k+) datasets. All of the comparisons were run on Apple Mac mini (2018) with 3.2 GHz 6-Core Intel Core i7 CPU, 64GB RAM, and operating system macOS 11.2. Thus, tricycle is able to scale with the increasing size of datasets.

Application of tricycle to a single-cell RNA-seq atlas

To demonstrate the scalability and generalizability of tricycle we applied it to a recent dataset of ≈ 4 million cells from the developing human [35]. The data were generated using combinatorial indexing (sci-RNA-seq3) and are relatively lightly sequenced with a median of 429–892 total UMIs for 4 single-cell profiled tissues and 354–795 for 11 single-nuclei profiled tissues (Additional file 1: Fig. S24). Using tricycle, we are able to rapidly and robustly annotate cell-cycle position for each of the cells/nuclei in this atlas (Fig. 6a, Additional file 1: Fig. S25). Within a global UMAP embedding, tricycle annotations enable immediate visual identification of proliferating and/or progenitor cell populations for most cell types and tissues. The rapid annotation of cell-cycle position on this reference dataset further allowed us to examine the relative differences in the proportion of cells actively proliferating across different tissues and cell types in the developing human. To quantify this, we discretized all cells along θ into two bins corresponding to actively proliferating ($0.25\pi < \theta < 1.5\pi$; S/G2/M) or non-proliferating (G1/G0). We next ranked each tissue by the relative proportion of actively proliferating cells to identify the tissues and cell types with the highest proliferative index (Fig. 6b). To examine cell-type-specific differences in proliferation potential, we computed the cell-cycle embedding as well as the proliferative index for the 9 most abundant cell types within each tissue (Additional file 1: Figs. S26, S27).

Tissue-level proliferation indexes identified the thymus, cerebrum, and adrenal gland as having the highest overall proportions of dividing cells across the sampled fetal time-points. Within the thymus, thymocytes represent both the most abundant cell type and the most “prolific” cell types as a function of the proportion of mitotic cells. Thymocytes exhibit a circular embedding in UMAP space that effectively recapitulates the estimated cell-cycle position predictions from tricycle (Additional file 1: Fig. S27k). Within this circular embedding, there is a gap of cells with cell-cycle position estimates at $\approx \pi$, consistent with dropout of cells and lower information content in M-phase. Comparison of tricycle cell-cycle annotations to SchwabeCC cell-cycle phase calls in this embedding suggests that tricycle more accurately estimates cell-cycle position even on cell types with a mean total UMI of 354 (Additional file 1: Fig. S28).

Within tissues, lymphoid cells are often the cell type with the highest proliferation index (Additional file 1: Figs. S26, S27), often with a greater number of actively proliferating cells than not. Within the fetal liver and spleen — both sites of early embryonic erythropoiesis during human development [39] — erythroblasts represent the cell type with the highest fraction of proliferating cells. Across developmental time, most tissues maintain relatively monotonic proliferation indices; however, several (liver, placenta, intestine)

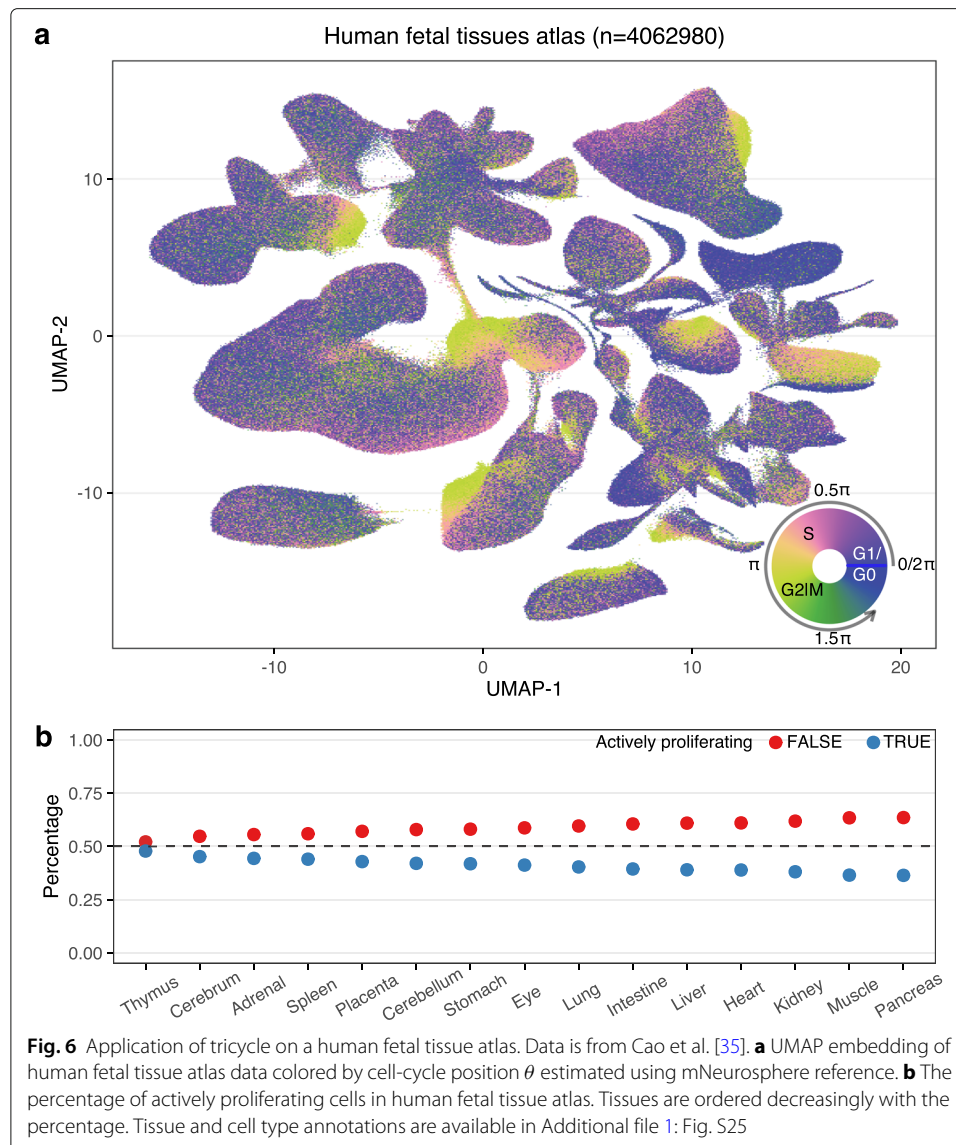


exhibit dynamic changes across the sampled timepoints. This application illustrates the utility of tricycle to atlas-level data.

Stability of the cell-cycle position assignments

To test the robustness of tricycle, we performed *in silico* experiments to determine the stability of cell-cycle position assignments. We evaluated three different types of stability wrt: (a) missing genes, (b) sequencing depth, and (c) data preprocessing.

When projecting new data into the cell-cycle reference embedding, it is common that the feature mapping between the two data sets contains only a subset of the 500 genes used in the embedding. The number of genes available for feature mapping has an impact on the shape of the resulting embedding; the mNeurosphere and mHippNPC datasets have almost the same shape when restricted to a set of common genes (Additional file 1: Fig. S29). To establish the stability of tricycle, we randomly removed genes from the neurosphere dataset and computed tricycle cell-cycle positions; we used the neurosphere

dataset as a positive control to ensure all genes are present. We used the circular correlation coefficient to assess the similarity between the tricycle cell-cycle position for the full dataset vs. the dataset with randomly pruned genes (Additional file 1: Figs. S30, S31). This reveals excellent stability (circular $\rho > 0.8$) using as little as 100 genes.

To examine the impact of sequencing depth, we downsampled the mHippNPC dataset (Additional file 1: Figs. S32, S33) and used the circular correlation coefficient to quantify the similarity to the cell-cycle position inferred using the full sample. Originally, the median of library sizes (total UMIs) is 10,000 for mHippNPC data. Downsampling to 20% of the original depth (approximate median of library sizes 2000) kept circular $\rho > 0.8$. This is congruent with the observed robustness of the method to the varying sequencing depth of the various datasets examined above.

Next, we examined the stability of tricycle regarding the choice of reference embedding. Above, we show a cell-cycle space estimated separately for the mNeurosphere and the mHippNPC datasets (Fig. 2). We observe that the inferred expression dynamics are more alike in the two datasets if we project the mHippNPC into the mNeurosphere embedding compare to using its own embedding. To quantify this, we pick key cell-cycle genes (previously examined in Additional file 1: Fig. S6) and compare the location of peak expression in the mNeurosphere dataset to that of the mHippNPC dataset with cell-cycle position estimated using these two approaches (Additional file 1: Fig. S34). For the vast majority of genes, the highest expression appear at a closer position when we estimate cell-cycle position by projecting the mHippNPC dataset into the mNeurosphere embedding.

To examine the impact of preprocessing data prior to projection, we compared cell-cycle position inferred using data processed with and without Seurat. Note that when we estimate the cell-cycle space, we use Seurat to align the different biological samples. But this is not done when we project new data using the prelearned reference. We observe negligible differences, whether or not Seurat is used (Additional file 1: Fig. S35). We also confirmed the direction of cell-cycle position θ is consistent with the direction of RNA velocity projections (Additional file 1: Fig. S36).

These results demonstrate the high sensitivity of tricycle to accurately estimate the cell cycle position across a high dynamic range of number of detectable genes within the feature map as well as depth of the information content in the target cells.

Discussion

Here, we have demonstrated the ability of tricycle to infer cell-cycle position in 26 datasets across species, cell types, and assay technologies. To do so — as is common in the field — we have made extensive use of gold standard datasets, with a particular emphasis on the FUCCI assay. We show tricycle compares favorably to FUCCI-based pseudotime, specifically the tricycle inferred cell-cycle position is a better predictor of expression dynamics of key cell-cycle genes compared to FUCCI-based pseudotime; however, FUCCI pseudotime has higher temporal resolution during the metaphase to anaphase transition, a very specific point during cell division.

An important limitation of existing gold standard datasets is that the measurements are done on cell lines where the main driver of expression can be assumed to be cell cycle. In contrast, many common applications of single-cell expression contain multiple cell types (e.g., tissue samples) and/or other strong drivers of expression such as differentiation. Predicting cell-cycle position in such datasets is much harder than predicting cell-cycle

position on homogeneous cell lines. For this reason, it is not enough to merely assess a method on gold standard datasets.

To address the limitation of gold-standard measurements on cell lines, we have made extensive use of internal controls. Specifically, we use inferred cell-cycle position to assess whether key cell-cycle genes (and log totalUMI if available) exhibit the expected expression dynamics across the learned θ progression. These internal controls are available in *any* single-cell expression dataset, including complex tissue samples. These internal controls do not by themselves give a clear answer to how precise the predictions are, but they undoubtedly carry some information on whether the inferred cell-cycle position is at all associated with cell-cycle phases. By using these internal controls, we overcome the limitations of the available gold standard datasets — which we can think of as having “external controls” — and show that tricycle performs well on differentiation datasets and datasets with multiple cell types. Additionally, we are able to ascertain the generalizability of the method. The cell-cycle genes we use for evaluation are also used to construct the reference embedding and for projection. However, at the projection stage, the weights are fixed without any dataset-dependent optimization. This removes — in our opinion — any circular reasoning. We note that internal controls are useful to assess any continuous prediction of cell-cycle position.

Here, we use a fixed reference embedding to represent cell cycle, defined using the mouse cortical neurosphere dataset. This raises the question: is there a single best embedding? One part of this question is whether there is a minimal best set of genes to construct the embedding? Empirically, when we project the neurosphere data into itself and remove genes, we get good performance with around 100 genes. However, this experiment does not measure generalizability, and we have anecdotally observed that reducing our gene list this much impacts performance in some datasets. Another part of this question is whether we can optimize the embedding to be as circular as possible. We observe that, despite different shapes, embeddings based on the cortical neurosphere and the primary hippocampal NPC datasets result in similar cell-cycle position estimates. One interpretation of these observations is that the robustness of the approach is derived from the structure created by the relationship of the genes to each other rather than the behavior of any individual marker gene, as described in Stein-O’Brien et al. [19].

In many single-cell experiments, cell cycle is often considered a confounding factor and as such, methods exist to remove this effect from the data prior to analysis. We caution against *removing* cell-cycle progression blindly as it can be intimately intertwined with other sources of variation of interest. Taking the mPancreas data as an example, there is a clear relationship between the number of cycling cells and differentiation as the multi-potent ductal cells advance to be terminally differentiated alpha and beta cells. If correction for cell-cycle progression is warranted, our analysis of the mPancreas data suggests that the common approach of regressing out principal components of cell-cycle genes may remove additional biological variation of interest.

It is currently unknown what will happen if tricycle is applied to a dataset without cycling cells. In this case, all cells belong to the G0/G1 cloud, and we hypothesize that the G0/G1 cloud will be centered on the (0, 0) origin and tricycle inferred cell-cycle position will be wrong. We expect to be able to diagnose this situation using internal controls, but we caution against using tricycle or other cell-cycle inference methods on a dataset without cycling cells.

Tricycle is a locked-down prediction procedure. There are no tuning parameters, neither explicitly set nor implicitly set through the use of cross-validation or alternatives. And in our applications, we have not aligned different samples to each other. That being said, we observe that different datasets exhibit small discrepancies. An example is that the precise location of peak expression in *Top2A* differs slightly from dataset to dataset. Possible sources of dataset-to-dataset variation include both biological and technical candidates such as technology and batch effects.

There are two specific sources of variation we want to highlight. First is the variation in nonzero expression. Specifically, we observe that 300–500 genes — out of a total of 500 genes — have non-zero expression in a given dataset, and we show that differences in these genes can cause changes in the shape of the embedding (Additional file 1: Figs. S29, S30, and S31). These differences in nonzero expression are also connected to the utility of using a larger set of genes to define the embedding space, as discussed above.

Second, the variation in the proportion of cells in G0 and G1, which is associated with the actual wall clock length of the cell cycle. Reflecting the biology of the system, this ratio also affects the placement of the origin of the projected data. However, this also potentially complicates across data set comparisons, as the only normalization currently performed in tricycle is the mean centering of each gene, which is susceptible to differences in this ratio. Lastly, we note that the peak expression location variations might be a consequence of different cell-cycle dynamics in different systems. We hope the biological implications will be examined closely in the future. Methods to expand tricycle to allow cross data comparisons are currently an active area of research.

We anticipate that the ability to model cell cycle as the continuous process that it is, will enable considerable advancements in the modeling of developmental and disease processes in which it plays a major role.

Conclusions

We have explained why the cell cycle — in datasets where the primary source of variation is cell cycle — is visible as an ellipsoid shape in the principal components of the data. We have shown that principal component analysis of cell-cycle genes sometimes reflects other processes such as differentiation. We have proposed to use projections into a reference embedding to isolate the specific cell-cycle signal in a dataset with many sources of variation, and we have shown that this approach allows us to isolate a specific, pre-specified signal.

We have shown that tricycle is capable of inferring continuous cell-cycle position and can be applied to datasets with multiple cell types, across species and a variety of single-cell technologies from relatively deeply sequenced plate based technologies to shallower sequenced droplet-based technologies, including very sparse data. As part of applying tricycle, a user can use internal controls to assess the validity of the inferred cell-cycle positions on their own dataset. Tricycle is highly scalable and available in an open-source implementation from the Bioconductor project.

Methods

Using principal component analysis to recover time ordering

We will consider the following statistical model. The mean expression of each gene is modeled as

$$f_g(\theta) = A_g \cos(\theta - d_g)$$

Here, A_g is a gene-specific amplitude and d_g is a mean-specific displacement (location of the peak). In this formulation, the mean function has a single peak and is periodic. We have G genes and each gene has its own (but not necessarily unique) (A_g, d_g) . We are assuming at least two peak locations not separated by π (i.e., two different d_g 's).

Basic trigonometry yields the identity

$$\begin{aligned} f_g(\theta) &= A_g \cos(\theta - d_g) \\ &= A_g \cos(d_g) \cos(\theta) + A_g \sin(d_g) \sin(\theta) \end{aligned}$$

which we can write as

$$\begin{aligned} f_g(\theta) &= (A_g \cos(d_g) \sqrt{\pi}) \phi_1(\theta) \\ &+ (A_g \sin(d_g) \sqrt{\pi}) \phi_2(\theta) = a_g^t \phi(\theta) \end{aligned}$$

using the orthonormal functions

$$\theta \mapsto \phi(\theta) = (\phi_1(\theta), \phi_2(\theta)) = \frac{1}{\sqrt{\pi}} (\cos(\theta), \sin(\theta))$$

Our derivation is based on Ramsey and Silverman [40] section 8.4. This section shows that the variance-covariance operator is given by

$$\nu(s, \theta) = \phi^t(s) (G^{-1} \mathbf{C}^t \mathbf{C}) \phi(\theta)$$

where the inner matrix (which turns out to determine the principal components) is a 2×2 matrix equal to

$$\begin{aligned} \frac{1}{G} \mathbf{C}^t \mathbf{C} &= \frac{1}{G} \sum_g a_g a_g^t = \frac{1}{G} \sum_g A_g^2 \pi \phi(d_g) \phi(d_g)^t \\ &= \frac{1}{G} \sum_g A_g^2 \begin{pmatrix} \cos^2(d_g) & \cos(d_g) \sin(d_g) \\ \cos(d_g) \sin(d_g) & \sin^2(d_g) \end{pmatrix} \end{aligned}$$

The principal component analysis is given by the Eigen-functions and -values of the variance-covariance operator. Such an Eigen-function and -value pair ξ, ρ takes the form

$$\xi(\theta) = b^t \phi(\theta)$$

for a vector b which satisfies

$$G^{-1} \mathbf{C}^t \mathbf{C} b = \rho b$$

i.e., b, ρ are Eigen-vectors and -values for the $G^{-1} \mathbf{C}^t \mathbf{C}$ matrix. Specifically, if $q_1, q_2, \lambda_1, \lambda_2$ are two such Eigen-vectors- and -values, then the two first principal components are given by

$$\theta \mapsto \xi(\theta) = (\cos(\theta), \sin(\theta))^t q_i \sqrt{G} \sqrt{\lambda_i}$$

Simulations

For Fig. 1 we performed the following simulation. Fifty realization of a cosine function with a location of 0.2 and an amplitude of 0.5 as well as 50 realizations of a cosine function with a location of 1.2 and an amplitude of 1. Each function was evaluated on an equidistant grid of 1000 points and independent Gaussian noise with a standard deviation of 0.2 was added. The depictions in Fig. 1a, b were each one of the realizations of the two different cosine functions.

For Additional file 1: Figs. S1, S2 and S3, we simulated data using the negative binomial distribution, inspired by the setup in Splatter [41]. In addition to a gene-specific amplitude (A_g) and location of the peak (L_g), we also consider different library size (l), which is an approximate as we still have some cell-to-cell variance. For a cell, we let $\lambda_g^0 = A_g \cos(\theta - L_g) + c$, with c a constant to ensure positivity of λ_g^0 . Then the cell mean is $\lambda_g' = l \cdot \lambda_g^0 / \sum^G \lambda_g^0$. The trended cell mean is simulated from a Gamma distribution as $\lambda_g \sim \text{Gamma}\left(1/B^2, \lambda_g' B^2\right)$, with B the biological coefficient of variation (we fix B as 0.1 in our simulations). Thus, the counts for gene g is given as $y_g \sim \text{Pois}(\lambda_g)$. We always simulate a 100 genes times 5000 cells count matrix, with cell timepoint θ uniformed distributed between 0 and 2π . We only vary one of L_g , A_g and l in Additional file 1: Figs. S1, S2, and S3. Specifically, in Additional file 1: Fig. S1, we used different number of distinct peak locations across 100 genes and fixed the amplitudes (across 100 genes) as 3 and library size as 2000. In Additional file 1: Fig. S2, we used different numbers of distinct amplitudes across 100 genes and fixed the number of distinct peak locations (across 100 genes) to 100 and library size to 2000. In Additional file 1: Fig. S3, we changed the library size l and fixed the number of distinct peak locations (across 100 genes) as 100 and the amplitudes (across 100 genes) as 3. PCA was performed on the library size normalized and \log_2 transformed matrix after we got the count matrix.

Generation of mouse primary hippocampal NPC scRNA-seq dataset

Hippocampal neural stem/progenitor cells (NPCs) were isolated by microdissection from E17 day embryos (offspring of male *Kmt2d*^{+/ β geo} and female C57Bl/6J) and cultured on Matrigel as described in [24]. Cells were maintained in an undifferentiated state via supplementation with growth factors (EGF, FGF2) in Neurobasal media. In a prior publication, we have demonstrated that the *Kmt2d*^{+/ β geo} cells exhibit defects in proliferation [24]. Replicate cultures from both genotypes were collected at the undifferentiated state (day 0) and then 2, 4, and 8 days after growth factor removal to induce neuronal differentiation. Cells were collected via trypsinization and pellets were washed and resuspended in Neurobasal media. scRNA-seq libraries were created with a Chromium Single-Cell 3' library & Gel Bead Kit v3 (10x Genomics) according to the manufacturer's protocol. Approximately 10,000 cells were targeted for each library. Only cells from day 0 are analyzed here.

Generation of mouse E14.5 Neurosphere scRNA-seq dataset

Cortical neurospheres were generated from the dissociated telencephalon of embryonic day 14.5 (E14.5) wild-type embryos. Embryos were harvested and the dorsal telencephalon was dissected away and collected in 1X HBSS at RT temperature. The dorsal telencephalon was gently triturated using p1000 pipette tips and the resultant cell suspension was spun at 500G for 5min and the media was aspirated off. The cell pellets were resuspended in complete neurosphere media 7ml (CNM) and plated in ultra-low adherence T25 flasks. CNM is made by combining 480ml DMEM-F12 with glutamine, 1.45g of glucose, 1X N2 supplement, 1X B27 supplement without retinoic acid, 1x penicillium/streptomycin, and 10ng/ml of both epidermal growth factor (EGF) and basic fibroblast growth factor (bFGF). The cell pellets were cultured for 3–5 days, or until spheroids have formed. The neurospheres were then collected and spun at 100G for 5min and the supernatant was removed. Neurospheres were resuspended in 5ml TrypLE and

incubated for a maximum of 5min at 37 °C with gentle trituration every 1.5min with a p1000 until the neurospheres are mostly a single-cell suspension. The cells were spun down at 500 G for 5 min and the supernatant was removed. The cells were resuspended in 15 ml of CNM and gently passed through a 40- μ M filter to remove large cell clumps. The resultant cell suspension was then plated in T75 flasks for another 2–5 days or until spheres began to have dark centers. This process was repeated two more times before cells were collected for 10X Genomics single-cell library prep. Before single-cell library preparation, the neurospheres were dissociated as described above and passed through a 40- μ M filter to ensure a single-cell suspension. Approx. 7000 cells were selected from each sample for input to the scRNA-seq library prep. scRNA-seq libraries were created using the Chromium Single-Cell 3' library & Gel Bead Kit v3 (10x Genomics) according to manufacturer protocol.

Reference genome and mapping index building

For mouse, GRCm38 reference genome fasta file and primary gene annotation GTF file (v25) were downloaded from GENCODE (<https://www.genencodegenes.org>). Similarly, GRCh38 reference genome fasta file and primary gene annotation GTF file(v35) were downloaded for human. We built a reference index for use by alevin as described in [42] using R package eisaR(v1.2.0), which we use to quantify both spliced and unspliced counts of annotated genes.

scRNA-seq preprocessing

Mouse Neurosphere (mNeurosphere) dataset

fastqs files were used to quantify both spliced and unspliced counts by Alevin (Salmon v1.3.0) with default settings as described in [42]. Abundances matrices were read in by R package tximeta (v1.8.1). The spliced counts were treated as the expression counts. We removed cells with less than 200 expressed genes, and cells flagged as outliers (deviating more than triple median absolute deviations(MAD) from the median of $\log_2(\text{TotalUMIs})$, $\log_2(\text{number of expressed genes})$, percentage of mitochondrial gene counts, or $\log_{10}(\text{doublet scores})$). The doublet scores were computed using doubletCells function in R package scran (v1.18.1). All mitochondrial genes and any genes which were expressed in less than 20 cells were further excluded from all subsequent analyses. Expression abundances were then library size normalized and \log_2 transformed by function normalizeCounts in R package scuttle (v1.0.2). The biological samples were integrated by Seurat (v3.2.2). We then run PCA on the top 2000 highly variable genes of the integrated $\log_2(\text{expression})$ using the runPCA function with default parameters, followed by running the runUMAP function on the resulting top 30 principal components with default parameters. Note that we did not restrict genes to cell-cycle genes in this step, as we would like to see the overall variation of the data. Cell types were inferred by SingleR package v1.4.0 using built-in MouseRNAseqData dataset as the reference.

Mouse primary hippocampal NPC (mHippNPC) dataset

All preprocessing are the same as for the mouse Neurosphere (mNeurosphere) dataset.

Mouse developing pancreas (mPancreas) dataset

We obtained the spliced and unspliced count matrices of the Mouse developing pancreas dataset from the python package scvelo (v0.2.1). The spliced counts were treated as the

expression counts. We removed cells with less than 200 expressed genes, and any cells flagged as outliers (deviating more than triple median absolute deviations (MAD) from the median of $\log_2(\text{TotalUMIs})$, $\log_2(\text{number of expressed genes})$, percentage of mitochondrial gene counts, or $\log_{10}(\text{doublet scores})$). Here, the doublet scores were computed using `doubletCells` function in R package `scrn` (v1.18.1). All mitochondrial genes and any genes which were expressed in less than 20 cells were further excluded from all subsequent analyses. Expression abundances were then library size normalized and \log_2 transformed by function `normalizeCounts` in R package `scuttle` (v1.0.2). We run PCA on the top 500 highly variable genes using the `runPCA` function with default parameters, followed by running the `runUMAP` function on the resulting top 30 principal components. When running the UMAP, we set `min_dist` to 0.5 instead of default value 0.01 to replicate the UMAP figure shown in [43] with other parameters default. Of note, the single-cell libraries of the data was generated using 10x Genomics' Chromium v2 system.

Mouse Hematopoietic Stem Cell (mHSC) dataset

We downloaded processed \log_2 transform TPM matrix directly from GEO under accession number GSE59114 [33]. We only used the cells from C57BL/6 strain, of which contains more cells, as the number of overlapped genes between xlsx file of C57BL/6 strain and DBA/2 strain is too small. Because the data was already processed and filtered, we did not perform any other processing. Unlike the abovementioned dataset, the SMARTer protocol was applied during library preparation.

Mouse Retina (mRetina) dataset

This dataset is available at https://github.com/gofflab/developing_mouse_retina_scRNASeq [34]. We removed cells flagged as outliers (deviating more than triple median absolute deviations (MAD) from the median of $\log_2(\text{TotalUMIs})$, $\log_2(\text{number of expressed genes})$, percentage of mitochondrial gene counts, or $\log_{10}(\text{doublet scores})$). As the total UMIs depend on cell type, we filtered the cells by blocking for each cell type. The doublet scores were computed using `doubletCells` function in R package `scrn` (v1.18.1). All mitochondrial genes and any genes which were expressed in less than 20 cells were further excluded from all subsequent analyses. Expression abundances were then library size normalized and \log_2 transformed by function `normalizeCounts`. We used the cell type annotations as the `new_CellType` column in the provided phenotype file. The single-cell libraries of the data was generated using 10x Genomics' Chromium v2 system.

HeLa cell lines datasets

The spliced and unspliced count matrices of HeLa Set 1 (HeLa1) and HeLa Set 2 (HeLa2) were downloaded from GEO website with accession number GSE142277 and GSE142356 [15]. Both datasets were generated by the same lab under the same protocol, while the sequencing depth of Set 2 is only about half that of Set 1. For each dataset, we only used the genes existing in both spliced and unspliced count matrices. The spliced counts were treated as the expression counts. We removed cells with less than 200 expressed genes, and cells flagged as outliers (deviating more than triple median absolute deviations (MAD) from the median of $\log_2(\text{TotalUMIs})$, $\log_2(\text{number of expressed genes})$, percentage of mitochondrial gene counts, or $\log_{10}(\text{doublet scores})$). All mitochondrial genes and any genes which were expressed in less than 20 cells were further excluded from all

subsequent analyses. Expression abundances were then library size normalized and \log_2 transformed by the function `normalizeCounts`. The single-cell libraries of the data were generated using Drop-seq system.

Mouse embryonic stem cell (mESC) dataset

The processed count matrix was downloaded from ArrayExpress website under accession number E-MTAB-2805 (<https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2805/>) [9]. We only retained 279 cells with $\log_2(\text{counts})$ greater than 15. The count matrix was library size normalized across cells and \log_2 transformed by function `normalizeCounts`. The RNA-seq data was generated using Fluidigm C1 system in this dataset.

Human embryonic stem cells (hESC) dataset

The processed count matrix was downloaded from GEO under accession number GSE64016 [10]. We only retained FACS sorted cells. The count matrix were library size normalized across cells and \log_2 transformed by function `normalizeCounts`. The RNA-seq data was generated using Fluidigm C1 system in this dataset.

Human U-2 OS cells (hU2OS) dataset

The TPM matrix was downloaded from GEO under accession number GSE146773 [20]. We only retained FACS sorted cells with $\log_2(\text{counts})$ greater than the 3 times MAD range. Genes which were expressed in less than 20 cell were removed. The left TPM matrix were library size normalized across cells and \log_2 transformed by function `normalizeCounts`. The RNA-seq data was generated using SMART-seq2 chemistry in this dataset. We got the FUCCI coordinates and FUCCI pseudotime directly from the authors of [20] (version 1.2).

Human induced pluripotent stem cells (hiPSCs) dataset

The processed FUCCI intensity and RNA-seq data was downloaded from <https://github.com/jdblischak/fucci-seq/blob/master/data/eset-final.rds?raw=true>. The preprocessing was described in [14]. The count matrix were library size normalized across cells and \log_2 transformed by function `normalizeCounts`. The RNA-seq data was generated using Fluidigm C1 system in this dataset.

Fetal tissue dataset

We got the loom file containing gene counts of all tissue from GEO under accession number GSE156793 [35]. We then processed and analyzed each tissue separately. For each tissue type, cells of which $\log_2(\text{TotalUMIs})$ is lower than median $- 3 \times \text{MAD}$, and genes expressed in less than 20 cells were excluded from further analyses. The count matrix was library size normalized across cells and \log_2 transformed by function `normalizeCounts`. All 4 tissues profiled using single-cell and 9 tissues profiled using single-nuclei were generated on sci-RNA-seq3 system.

Five-stage cell-cycle assignments

The 5-stage (G1S, S, G2, G2M, and MG1) cell-cycle assignments were adapted from [15] with some modifications. Briefly, the assignments use the high expression genes list for each stage, curated by [7]. Let k represent one of the 5 stages, and $l_k = \{g_k^1, g_k^2, \dots, g_k^{p_k}\}$

represent the gene list with p_k genes. For each stage k , we could calculate the mean expression across genes in the gene list l_k for the j th cell as $m_{j,k} = \frac{1}{p_k} \sum_{g_k^i \in l_k} E_{g_k^i,j}$ with $E_{g_k^i,j}$ as the \log_2 transformed expression value of gene g_k^i and cell j . Then we assess how well a gene in a gene list correlates to the mean expression level of that gene list as $c_{g_k^i} = \text{cor}(E_{g_k^i}, m_k)$. For each stage, the gene list is pruned to genes with $c_{g_k^i} > 0.2$. (For the fetal tissues dataset, we used $c_{g_k^i} > 0.15$ since the extremely shallowly sequenced data shows less co-expression patterns and the threshold 0.2 could leave us with no genes.) We label this pruned new gene list as $L_k = \{g_k^1, g_k^2, \dots, g_k^{q_k}\}$ with q_k the number of genes. The stage assignment score for cell j and stage k is given as

$$A_{k,j} = \frac{1}{q_k} \sum_{g_k^i \in L_k} E_{g_k^i,j}$$

The 5-by-n matrix \mathbf{A} , of which the number of columns equals to the number of cells, follows z-score transformations w.r.t. first rows and then columns, resulting the 5-by-n matrix $\tilde{\mathbf{A}} = (\tilde{A}_{k,j})$. For each cell, we compute the preliminary stage assignment as $s_j = \arg \max_k \{\tilde{A}_{k,j}\}$.

As in the [15], we also apply two filtering steps. The first filtering, which is unchanged from the original method, is as follows. We require s_j , the stage with the second largest assignment score, to be the neighboring stage to s_j . This requirement corresponds to the 5 stages being continuously cyclic processes.

As for the second filtering step, the original method discards all cells with the second largest assignment score $\tilde{A}_{s_j,j} > 0.75$. We found the threshold of 0.75 to some extent not applicable, as in some datasets it leads to losing 90% of cells. Therefore, we use a more adaptive threshold by requiring $\tilde{A}_{s_j,j} - \tilde{A}_{s_j,j} > 0.3$.

If the cell passes two filtering steps, it will be assigned to a stage s_j . Otherwise, it would be assigned as *NA* w.r.t. 5 stages of the cell cycle. To mitigate the batch effect on the 5 stage assignments, the assigning procedures are done for each sample/batch separately within each dataset, as recommended in Revelio package [15].

PCA of GO cell-cycle genes

For each dataset, we subsetting the preprocessed \log_2 transformed expression matrix to genes in the GO term cell cycle (GO:0007049). If there are clear batches defined in the dataset, such as sample or batch, we use Seurat3 to remove batch effect. In the case of using Seurat3, we used a library size normalized count matrix as input instead of \log_2 transformed values. The integration anchors were searched in the space of the top 30 PCs. The output integrated matrix is a \log_2 transformed matrix of the top 500 most variable genes. We then performed principal component analysis on the gene-wise mean centered expression matrix. In the case of no batch existing, we also restricting to the top 500 variable genes among GO cell-cycle genes.

Projection of new data to cell-cycle embedding and calculation of cell-cycle position θ

The projection using pre-learned weights matrix during PCA of GO cell-cycle genes is straight forward, given by

$$\mathbf{P} = \tilde{\mathbf{E}}^t \cdot \mathbf{R}$$

where \mathbf{R} represents the o -by-2 reference matrix ($o \leq 500$), contains the weights of top 2 PCs learned from PCA of GO cell-cycle genes; $\tilde{\mathbf{E}}$ is a o -by- n matrix, subsetting from \mathbf{E} (the \log_2 transformed expression matrix) with genes in the weights matrix and row-means centered. The resulting n -by-2 \mathbf{P} is the cell-cycle embedding projected by the reference. The calculation of the cell-cycle position θ is given by

$$\theta = \arctan\left(\frac{P_2}{P_1}\right)$$

where P_i is the i th column of matrix \mathbf{P} . When mapping the genes between weights matrix and the data that we want to project, the Ensemble ID is given higher priority than the gene symbol for mouse. For across species projection, we only consider the homologous genes of the same gene symbols.

Periodic loess

As θ is a circular variable bound between 0 to 2π , fitting a traditional loess model $y \sim \theta$, with y as any response variable, such as the gene expression of gene, or $\log_2(\text{TotalUMIs})$, has problems around the boundaries 0 and 2π . Hence, we concatenate triple y and triple θ with one period shift to form $[y, y, y]$ and $[\theta - 2\pi, \theta, \theta + 2\pi]$, on which the loess line is fitted. We then only use the fitted value \hat{y} when θ is between 0 and 2π for visualization purpose.

The calculation of the coefficient of determination R^2 of the fitted loess model is given by

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{total}}}$$

Here $SS_{\text{res}} = \sum_i^n (y_i - \hat{y}_i)^2$ and $SS_{\text{total}} = \sum_i^n (y_i - \bar{y})^2$. Note that instead of using all three copies of data points, we restrict the calculation of SS_{res} and SS_{total} on the original data points (the middle copy). The residuals are not the same for the three copies, especially at the beginning and end of $[-2\pi, 2\pi]$.

The circular correlation coefficient ρ

We use the circular correlation coefficient ρ defined by [44] to evaluate concordance between two polar vectors θ_1 and θ_2 . It is defined as follows

$$\rho = \frac{\Sigma[\sin(\theta_1 - \mu_1) \cdot \sin(\theta_2 - \mu_2)]}{\sqrt{\Sigma[\sin^2(\theta_1 - \mu_1)] \cdot \Sigma[\sin^2(\theta_2 - \mu_2)]}}$$

μ_1 and μ_2 represent the mean of θ_1 and θ_2 respectively, and are estimated by maximum likelihood estimation under von Mises distribution assumption.

Running other methods

For other cell-cycle inference methods, we use all default parameters and its built-in reference (if needed) in the following packages: cyclone in scran (v1.18.5), CellCycleScoring in Seurat (v4.0.0.9015), Revelio (v0.1.0), peco (v1.1.21), and reCAT (v1.1.0).

Silhouette index on angular separation distance of tricycle cell-cycle position θ

For cyclone and Seurat, we could use Silhouette index to describe consistency between discretized cell-cycle stage and tricycle cell-cycle position θ . We use angular separation

distance metric to quantify the distance between cell i and cell j as

$$d(i, j) = 1 - \cos(\theta_i - \theta_j)$$

For a cell $i \in S_{k^{(i)}} \ni k^{(i)} \in \{G1, S, G2M\}$. The mean distance between cell i and all other cells assigned to the same stage

$$a(i) = \frac{1}{|S_{k^{(i)}}| - 1} \sum_{j \in S_{k^{(i)}}, i \neq j} d(i, j)$$

with $|S_{k^{(i)}}|$ the cardinality of $S_{k^{(i)}}$. Specially, $a(i) = 0$ if $|S_{k^{(i)}}| = 1$. The mean distance from cell i to all cells assigned to other stage k' such that $k' \neq k^{(i)} \wedge k' \in \{G1, S, G2M\}$ is

$$b(i) = \min_{k' \neq k^{(i)}} \frac{1}{|S_{k'}|} \sum_{j \in S_{k'}} d(i, j)$$

The Silhouette index for cell i is given as

$$s(i) = \begin{cases} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} & \text{if } |S_{k^{(i)}}| > 1 \\ 0 & \text{if } |S_{k^{(i)}}| = 1 \end{cases}$$

For any cell i , the Silhouette index $s(i)$ is bound between -1 and 1 ($-1 \leq s(i) \leq 1$). An $s(i)$ close to 1 means the cell is consistently assigned to its neighbors w.r.t. its cell-cycle position θ_i . An $s(i)$ close to -1 means the cell is closer to the other stage. An $s(i)$ equals to 0 means the cell is on the border of two stages. The mean Silhouette index on all cells measures how tight the stage assignments are. In this context, this value must be interpreted carefully as it is different from traditional clustering which might put hard boundaries and gaps between clusters. As the cell-cycle process is continuous in nature, there must be cells assigned on the boundaries and ambiguous to either stage, and no gap should appear between stages. Thus, the mean silhouette index greater than 0 might be appropriate to conclude the agreement between tricycle cell-cycle position θ and discretized cell cycle stages.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-021-02581-y>.

Additional file 1: Supplementary Materials. Supplementary methods about comparison with existing cell cycle tools and supplementary figures (Fig. S1-S36).

Additional file 2: Review history.

Review history

The review history is available as Additional file 2.

Peer review information

Barbara Cheifet was the primary editor of this article and managed its editorial process in collaboration with the rest of the editorial team.

Disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or National Science Foundation.

Authors' contributions

Conceptualization: SCZ, GS-OŠB, LAG and KDH; Methodology: SCZ, GS-OŠB, JS, LAG and KDH; Software: SCZ; Formal analysis: SCZ; Investigation: JJA, JS, GAC, BW and GS; Resources: SCZ, JJA, GAC, BW, GS, HTB, LAG and KDH; Data Curation: SCZ, JJA, JS and LAG; Writing - Original Draft: SCZ, GS-OŠB, LAG and KDH; Writing - Review & Editing: all co-authors; Visualization: SCZ; Supervision: LAG and KDH; Project administration: LAG and KDH; Funding acquisition: HTB, LAG and KDH. All authors read and approved the final manuscript.

Funding

This project has been made possible in part by grant number CZF2019-002443 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award R01GM121459. This work was additionally supported by awards from the National Science Foundation (IOS-1665692), the National Institute of Aging (R01AG066768), and the Maryland Stem Cell Research Foundation (2016-MSCRFI-2805). GSO is supported by postdoctoral fellowship awards from the Kavli NeuroDiscovery Institute, the Johns Hopkins Provost Award Program, and the BRAIN Initiative in partnership with the National Institute of Neurological Disorders (K99NS122085).

Availability of data and materials

The mouse E14.5 Neurosphere data is available under GEO accession number GSE171636 [45]. The mouse primary hippocampal NPC data is available under GEO accession number GSE190514 [46]. Other public datasets used from GEO or the corresponding webpage: mouse developing pancreas (mPancreas) dataset (<https://scvelo.readthedocs.io/scvelo.datasets.pancreas/#scvelo.datasets.pancreas> [32, 43]), mouse hematopoietic stem cell (mHSC) dataset (GSE59114 [33]), mouse retina (mRetina) dataset (https://github.com/gofflab/developing_mouse_retina_scRNASeq [34]), Hela cell lines datasets (GSE142277 and GSE14235 [15]), mouse embryonic stem cell (mESC) dataset (<https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2805/> [9]), human embryonic stem cells (hESC) dataset (GSE64016 [10]), human U-2 OS cells (hU2OS) dataset (GSE146773 [20]), human induced pluripotent stem cells (hiPSCs) dataset (<https://github.com/jdblichak/fucci-seq/blob/master/data/eset-final.rds?raw=true> [14]), and fetal tissue dataset (GSE156793 [35]). All code to analyze the data and generate figures is available at https://github.com/hansenlab/tricycle_paper_figs [47]. The processed data is available at <https://doi.org/10.5281/zenodo.5519841> [48]. The tricycle method is implemented in the R package tricycle containing the mNeurosphere reference, which is available on <https://github.com/hansenlab/tricycle> [49] and on Bioconductor under the GPL-3 license (<https://www.bioconductor.org/packages/tricycle> [50]).

Declarations

Ethics approval and consent to participate

All mouse experiments were conducted in accordance with the National Institutes of Health Guidance for the Care and Use of Laboratory Animals. All procedures and protocols were approved by the Animal Care and Use Committee of the Johns Hopkins University (protocol numbers: MO21M54 for mouse cortical neurosphere generation; and MO21M132 for mouse primary hippocampal NPC generation).

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA. ²Department of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, USA. ³Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, USA. ⁴Kavli NeuroDiscovery Institute, Johns Hopkins University, Baltimore, USA. ⁵Division of Biostatistics and Bioinformatics, Department of Oncology, Johns Hopkins School of Medicine, Baltimore, USA. ⁶Department of Pediatrics, Johns Hopkins School of Medicine, Baltimore, USA. ⁷Faculty of Medicine, University of Iceland, Reykjavik, Iceland. ⁸Landspítali University Hospital, Reykjavik, Iceland.

Received: 7 June 2021 Accepted: 17 December 2021

Published online: 31 January 2022

References

- McConnell S, Kaznowski C. Cell cycle dependence of laminar determination in developing neocortex. *Science*. 1991;254(5029):282–5. <https://doi.org/10.1126/science.1925583>.
- Ambros V. Cell cycle-dependent sequencing of cell fate decisions in *Caenorhabditis elegans* vulva precursor cells. *Development*. 1999;126(9):1947–56.
- Ohnuma S-I, Harris WA. Neurogenesis and the cell cycle. *Neuron*. 2003;40(2):199–208. [https://doi.org/10.1016/S0896-6273\(03\)00632-9](https://doi.org/10.1016/S0896-6273(03)00632-9).
- Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*. 1998;9(12):3273–97. <https://doi.org/10.1091/mbc.9.12.3273>.
- Cho RJ, Campbell MJ, Winzler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*. 1998;2(1):65–73. [https://doi.org/10.1016/S1097-2765\(00\)80114-8](https://doi.org/10.1016/S1097-2765(00)80114-8).
- Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*. 2000;97(18):10101–6. <https://doi.org/10.1073/pnas.97.18.10101>.
- Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell*. 2002;13(6):1977–2000. <https://doi.org/10.1091/mbc.02-02-0030>.
- Gauthier NP, Larsen ME, Wernersson R, de Lichtenberg U, Jensen LJ, Brunak S, Jensen TS. Cyclebase.org—a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Res*. 2008;36(Database issue):854–9. <https://doi.org/10.1093/nar/gkm729>.

9. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol.* 2015;33(2):155–60. <https://doi.org/10.1038/nbt.3102>.
10. Leng N, Chu L-F, Barry C, Li Y, Choi J, Li X, Jiang P, Stewart RM, Thomson JA, Kendziorski C. Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat Methods.* 2015;12(10):947–50. <https://doi.org/10.1038/nmeth.3549>.
11. Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann SA, Stegle O, Marioni JC, Buettner F. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods.* 2015;85:54–61. <https://doi.org/10.1016/j.jymeth.2015.06.021>.
12. Liu Z, Lou H, Xie K, Wang H, Chen N, Aparicio OM, Zhang MQ, Jiang R, Chen T. Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat Commun.* 2017;8(1):22. <https://doi.org/10.1038/s41467-017-00039-z>.
13. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck 3rd WM, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. *Cell.* 2019;177(7):1888–190221. <https://doi.org/10.1016/j.cell.2019.05.031>.
14. Hsiao CJ, Tung P, Blischak JD, Burnett JE, Barr KA, Dey KK, Stephens M, Gilad Y. Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq data analysis. *Genome Res.* 2020;30(4):611–21. <https://doi.org/10.1101/gr.247759.118>.
15. Schwabe D, Formichetti S, Junker JP, Falcke M, Rajewsky N. The transcriptome dynamics of single cells during the cell cycle. *Mol Syst Biol.* 2020;16(11):9946. <https://doi.org/10.15252/msb.20209946>.
16. Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, Rabadan R. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nat Biotechnol.* 2017;35(6):551–60. <https://doi.org/10.1038/nbt.3854>.
17. Pan SJ, Kwok JT, Yang Q. Transfer learning via dimensionality reduction. In: AAAI. Association for the Advancement of Artificial Intelligence; 2008. p. 677–82.
18. Pan SJ, Yang Q. A Survey on Transfer Learning. *IEEE Trans Knowl Data Eng.* 2009;22(10):1345–59. <https://doi.org/10.1109/tkde.2009.191>.
19. Stein-O'Brien GL, Clark BS, Sherman T, Zibetti C, Hu Q, Sealton R, Liu S, Qian J, Colantuoni C, Blackshaw S, Goff LA, Fertig EJ. Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. *Cell Syst.* 2019;8(5):395–4118. <https://doi.org/10.1016/j.cels.2019.04.004>.
20. Mahdessian D, Cesnik AJ, Gnann C, Danielsson F, Stenström L, Arif M, Zhang C, Le T, Johansson F, Shuttan R, Bäckström A, Axelsson U, Thul P, Cho NH, Carja O, Uhlén M, Mardinoglu A, Stadler C, Lindskog C, Ayoglu B, Leonetti MD, Pontén F, Sullivan DP, Lundberg E. Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature.* 2021;590(7847):649–54. <https://doi.org/10.1038/s41586-021-03232-9>.
21. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet.* 2008;40(5):646–9. <https://doi.org/10.1038/ng.139>.
22. Gray RM. Toeplitz and Circulant Matrices: A Review. *Found Trends Commun Inf Theory.* 2005;2(3):155–239. <https://doi.org/10.1561/0100000006>.
23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Michael Cherry J, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: tool for the unification of biology. *Nat Genet.* 2000;25(1):25–29. <https://doi.org/10.1038/75556>.
24. Carosso GA, Boukas L, Augustin JJ, Nguyen HN, Winer BL, Cannon GH, Robertson JD, Zhang L, Hansen KD, Goff LA, Bjornsson HT. Precocious neuronal differentiation and disrupted oxygen responses in kabuki syndrome. *JCI Insight.* 2019;4(20):. <https://doi.org/10.1172/jci.insight.129375>.
25. Lodish, Berk H, Kaiser A, Kaiser CA, Krieger C, Scott M, Bretscher MP, Ploegh Ay, Matsudaira H, others P. Section 12.3 the role of topoisomerases in DNA replication. In: *Molecular Cell Biology*, 4th ed. New York: W. H. Freeman; 2008.
26. Ono T, Losada A, Hirano M, Myers MP, Neuwald AF, Hirano T. Differential contributions of condensin I and condensin II to mitotic chromosome architecture in vertebrate cells. *Cell.* 2003;115(1):109–21. [https://doi.org/10.1016/s0092-8674\(03\)00724-4](https://doi.org/10.1016/s0092-8674(03)00724-4).
27. Wei-Shan H, Amit VC, Clarke DJ. Cell cycle regulation of condensin Smc4. *Oncotarget.* 2019;10(3):263–76. <https://doi.org/10.18632/oncotarget.26467>.
28. Heck MM, Hittelman WN, Earnshaw WC. Differential expression of dna topoisomerases I and II during the eukaryotic cell cycle. *Proc Natl Acad Sci.* 1988;85(4):1086–90. <https://doi.org/10.1073/pnas.85.4.1086>.
29. Belluti S, Basile V, Benatti P, Ferrari E, Marverti G, Imbriano C. Concurrent inhibition of enzymatic activity and NF- κ B-mediated transcription of Topoisomerase- α by bis-DemethoxyCurcumin in cancer cells. *Cell Death Dis.* 2013;4(8):756. <https://doi.org/10.1038/cddis.2013.287>.
30. Marguerat S, Bähler J. Coordinating genome expression with cell size. *Trends Genet.* 2012;28(11):560–5. <https://doi.org/10.1016/j.tig.2012.07.003>.
31. Padovan-Merhar O, Nair GP, Biaesch AG, Mayer A, Scarfone S, Foley SW, Wu AR, Churchman LS, Singh A, Raj A. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol Cell.* 2015;58(2):339–52. <https://doi.org/10.1016/j.molcel.2015.03.005>.
32. Bastidas-Ponce A, Tritschler S, Dony L, Scheibner K, Tarquis-Medina M, Salinno C, Schirge S, Burtscher I, Böttcher A, Theis FJ, Lickert H, Bakhti M. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development.* 2019;146(12):173849. <https://doi.org/10.1242/dev.173849>.
33. Kowalczyk MS, Tirosh I, Heckl D, Rao TN, Dixit A, Haas BJ, Schneider RK, Wagers AJ, Ebert BL, Regev A. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* 2015;25(12):1860–72. <https://doi.org/10.1101/gr.192237.115>.
34. Clark BS, Stein-O'Brien GL, Shiau F, Cannon GH, Davis-Marcisak E, Sherman T, Santiago CP, Hoang TV, Rajaii F, James-Espósito RE, Gronostajski RM, Fertig EJ, Goff LA, Blackshaw S. Single-cell RNA-Seq analysis of retinal development identifies NFI factors as regulating mitotic exit and late-born cell specification. *Neuron.* 2019;102(6):1111–11265. <https://doi.org/10.1016/j.neuron.2019.04.010>.

35. Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, Zager MA, Aldinger KA, Blecher-Gonen R, Zhang F, Spielmann M, Palis J, Doherty D, Steemers FJ, Glass IA, Trapnell C, Shendure J. A human cell atlas of fetal gene expression. *Science*. 2020;370(6518):7721. <https://doi.org/10.1126/science.aba7721>.
36. Sakaue-Sawano A, Kurokawa H, Morimura T, Hanyu A, Hama H, Osawa H, Kashiwagi S, Fukami K, Miyata T, Miyoshi H, Imamura T, Ogawa M, Masai H, Miyawaki A. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell*. 2008;132(3):487–98. <https://doi.org/10.1016/j.cell.2007.12.033>.
37. Sakaue-Sawano A, Yo M, Komatsu N, Hiratsuka T, Kogure T, Hoshida T, Goshima N, Matsuda M, Miyoshi H, Miyawaki A. Genetically encoded tools for optical dissection of the mammalian cell cycle. *Mol Cell*. 2017;68(3):626–6405. <https://doi.org/10.1016/j.molcel.2017.10.001>.
38. McGarry TJ, Kirschner MW. Geminin, an inhibitor of DNA replication, is degraded during mitosis. *Cell*. 1998;93(6):1043–53. [https://doi.org/10.1016/s0092-8674\(00\)81209-x](https://doi.org/10.1016/s0092-8674(00)81209-x).
39. Cumano A, Godin I. Ontogeny of the hematopoietic system. *Ann Rev Immunol*. 2007;25:745–85. <https://doi.org/10.1146/annurev.immunol.25.022106.141538>.
40. Ramsay H, Silverman BW. *Functional Data Analysis*, 2nd Ed: Springer; 2005.
41. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol*. 2017;18(1):174. <https://doi.org/10.1186/s13059-017-1305-0>.
42. Soneson C. RNA velocity with alevin. <https://combine-lab.github.io/alevin-tutorial/2020/alevin-velocity/>. Accessed 1 Aug 2020.
43. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol*. 2020;1–7. <https://doi.org/10.1038/s41587-020-0591-3>.
44. Jammalamadaka SR, Sarma Y. A correlation coefficient for angular variables. In: *Statistical theory and data analysis II*. Proceedings of the Second Pacific Area Statistical Conference. Amsterdam North Holland; 1988. p. 349–364.
45. Augustin JJ, Winer B, Goff LA. Universal prediction of cell cycle position using transfer learning. *Datasets*. 2021. Gene Expression Omnibus. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE171636>. Accessed 7 Apr 2021.
46. Giovanni CA, Gloria S, Zheng SC, Stein-O'Brien G, Augustin JJ, Jared S, Winer B, Hans BT, Goff LA, Hansen KD. Mouse primary hippocampal neural progenitor cell scRNA-seq from wild type and a Kmt2d-haploinsufficient model of Kabuki syndrome. *Datasets*. 2021. Gene Expression Omnibus. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE190514>. Accessed 13 Dec 2021.
47. Zheng SC. Universal prediction of cell cycle position using transfer learning. GitHub. 2021. https://github.com/hansenlab/tricycle_paper_figs. Accessed 5 Jan 2022.
48. Zheng SC. Universal prediction of cell cycle position using transfer learning. Zendo. 2021. <https://doi.org/10.5281/zenodo.5519841>.
49. Zheng SC. Tricycle: Transferable Representation and Inference of cell cycle. GitHub. 2021. <https://github.com/hansenlab/tricycle>. Accessed 1 Dec 2021.
50. Zheng SC. Tricycle: Transferable Representation and Inference of cell cycle. Bioconductor. 2021. <https://doi.org/10.18129/B9.bioc.tricycle>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

