Trajectories of cell-cycle progression from fixed cell populations

5 authors, including:

Some of the authors of this publication are also working on these related projects:

Machine learning for biology View project

Methods for mass cytometry View project

Trajectories of cell-cycle progression from fixed cell populations

Gabriele Gut[1,2,5], Michelle D Tadmor[3,5], Dana Pe'er[3,6], Lucas Pelkmans[1,6] & Prisca Liberali[1,4,6]

An accurate dissection of sources of cell-to-cell variability is crucial for quantitative biology at the single-cell level but has been challenging for the cell cycle. We present Cycler, a robust method that constructs a continuous trajectory of cell-cycle progression from images of fixed cells. Cycler handles heterogeneous microenvironments and does not require perturbations or genetic markers, making it generally applicable to quantifying multiple sources of cell-to-cell variability in mammalian cells.

The cell cycle controls fundamental activities of eukaryotic cells and is a major source of cell-to-cell variability in unsynchronized cell populations. Despite a deep understanding of the molecular machinery regulating cell-cycle progression, study of the variability of cellular activities determined by the cell cycle remains experimentally challenging. Time-lapse imaging allows tracking of the full cell cycle in individual cells, but doing this in parallel with measurements of a protein's abundance—and specifically its activated form—relies on fluorescent live-cell reporters whose development is often difficult and time-consuming[1,2]. Other established methods that infer cell-cycle state are more easily combined with additional single-cell measurements, but these focus on specific sub-steps (typically mitosis or M phase)[1,3], lack temporal accuracy[4] or require perturbations[5,6]. A recent approach that allows the inference of cell-cycle progression rates has the disadvantage that it requires genetic modifications and homogenous growth conditions[7]. Thus, we found a need for a versatile approach to infer cell-cycle state in additional experimental scenarios.

Here we describe Cycler, a method that constructs a trajectory of cell-cycle progression from fixed images of unperturbed cells growing in heterogeneous microenvironments. Cycler achieves this by inferring a trajectory within a multivariate feature space, which orders single cells according to their relative position in the cell cycle and quantifies single-cell activities along this trajectory. First, nuclei are imaged and segmented. Then, single-cell measurements of DNA content, DNA replication and pattern, nuclear area

and local cell crowding[8] are combined in a multivariate feature vector (**Fig. 1a** and **Supplementary Fig. 1a**). Given the nonlinear nature of the feature space (**Fig. 1a** and **Supplementary Fig. 1b**), Cycler, a new version of Wanderlust[9], performs a k-nearest neighbor graph–based embedding of the multivariate feature space into a single dimension, the cell-cycle trajectory (CCT) (**Fig. 1a** and Online Methods). By taking advantage of changes in cell-cycle features (**Supplementary Fig. 1c,e**), the algorithm recovers the temporal order of progression, revealing the expected behavior of changes along the cell cycle (**Fig. 1b** and **Supplementary Fig. 1c**). A support vector machine (SVM)-based classification of single cells is used to subdivide the CCT into regions belonging to each of the discrete cell-cycle phases. (**Fig. 1a**, **Supplementary Fig. 1d** and Online Methods). This allows identifying the cell-cycle phase transition points along the CCT, as well as the 'starting cells', a subset of G1 cells close to M cells in multivariate space (Online Methods and **Supplementary Fig. 2a**). Cycler's accuracy in inferring the CCT is highly robust to variations in starting cell selection (Online Methods and **Supplementary Fig. 2b**).

To validate the CCT, we first quantified single-cell levels of proliferating cell nuclear antigen (PCNA) and cyclin A in five cell lines using immunofluorescence. As expected, PCNA and cyclin A levels showed variation over the cell cycle, peaking in S and G2, respectively (**Fig. 1c** and **Supplementary Fig. 2c**). Additionally, we synchronized cells using mitotic arrest and examined how they moved along the CCT at different time points following release (Online Methods). Two hours after release, single cells were primarily positioned early in the CCT, and over time the bulk of cells progressed along the CCT as expected (**Fig. 1d**). Finally, we tracked 100 single cells for 72 h to quantify growth dynamics along the cell cycle. The nuclear area growth curve along the CCT derived from a fixed cell population was nearly identical to the average nuclear area growth curve of the tracked cells[7] (median $r = 0.91 \pm 0.013$, s.e.m.) (**Fig. 1e**). Moreover, single-cell tracks show that individual cells temporally transitioned through the CCT (**Fig. 1e**). Thus, Cycler achieves highly accurate trajectories that reflect order in cell-cycle progression and reveals dynamic details that correspond to high temporal resolution.

We found that taking local cell crowding into account was essential for Cycler's high performance. Although the nuclear area of adherent mammalian cells is influenced by cell-cycle progression, it is also determined by microenvironmental influences such as local cell crowding (**Fig. 2a,b**) that act independently of the cell cycle, as shown in the partial correlation network (**Supplementary Fig. 3a**). For example, a particular nuclear size (**Fig. 2b**, dashed line) can belong to G1 phase cells growing in areas of low crowding, as well as to S cells growing in areas of high

Figure 1 | Cycler infers trajectories of cell-cycle progression from image-based data sets. (**a**) Outline of experimental and computational workflow. Cell and nuclei are segmented, and features are extracted for the construction of CCT (1); Cycler uses the single-cell multivariate feature space to generate the CCT. The 3D scatter plot represents the nonlinear relations between nuclear area, DNA content and DNA replication and is color-coded for CCT progression (2). Cells are ordered along the CCT and fractions of cells in the cell-cycle stage phases (SVM-based) are overlaid (3). (**b**) Selected single-cell features used to build the CCT are plotted along the CCT. Lines, weighted local mean; shading, weighted local s.d. (Online Methods). Fraction of cells in G1, S and G2 are indicated. (**c**) Behavior of cell-cycle markers along the CCT. Lines, weighted local mean; shading, weighted local s.d. (Online Methods). Cell-cycle marker data was not used for CCT construction. (**d**) Cells arrested in M phase are released for various amounts of time and fixed. The violin plots show the distribution of cells along the CCT at increasing release intervals. Colors show the discrete classification into G1, S and G2. (**e**) Nuclear area corrected for local cell crowding along the CCT (black curve), nuclear area corrected for local cell crowding of 100 single cells tracked between two mitotic events (blue curve) and single-cell traces on tracked nuclear area (green). Solid lines, median; shading, interquartile range LCC, local cell crowding.

crowding. Cycler's ability to take microenvironmental effects into account allows accurate CCT retrieval from five cell lines with different population characteristics (**Supplementary Fig. 2d**). It was also important for Cycler's robustness and reproducibility between CCTs inferred from two independent populations of the same cell line. Improvement was primarily seen for cells in G1 (**Fig. 2c,d** and Online Methods), as nuclear size is the dominant feature used to infer progression in this part of the CCT (**Supplementary Fig. 3b**).
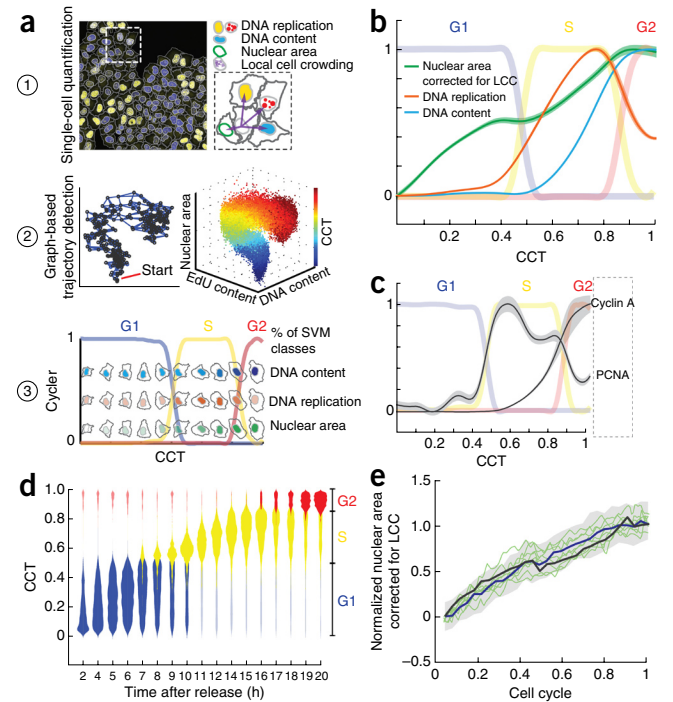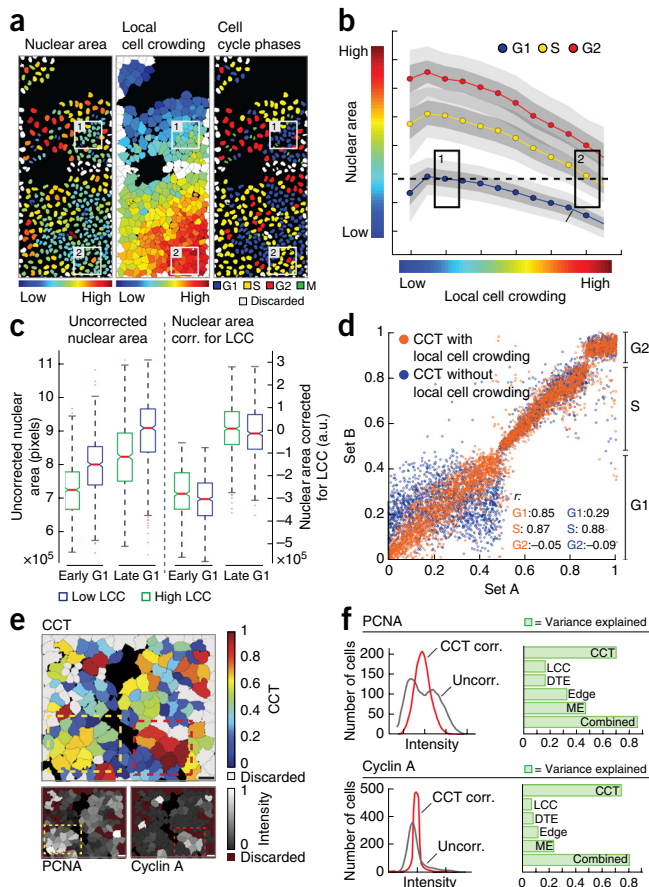
To delineate the minimal information necessary to build an accurate CCT, we compared trajectories using combinations of the features measured in each single cell. We found that three features—nuclear area, DNA content and local cell crowding—were sufficient to infer a good-quality CCT (**Supplementary Fig. 3c**). However, to obtain sharp transition boundaries between G1/S and S/G2, features of DNA replication must be included (**Supplementary Fig. 3c**). To further evaluate robustness, we quantified the minimal number of cells necessary to construct reliable CCTs. A CCT constructed from 1,200 randomly picked cells is already highly reproducible ($r_{ALL} = 0.98$; $r_{G1} = 0.85$; $r_S = 0.99$; $r_{G2} = 0.76$) (**Supplementary Fig. 3d,e**).

Figure 2 | Features of the single-cell microenvironment are important for accurate CCTs. (**a**) Overview of a cell population growing in heterogeneous environment. Left, nuclei color-coded for nuclear area. Middle, cells color-coded for local cell crowding; right, nuclei color-coded for cell-cycle phases. Region 1 marks G1 cells that grow at low local cell crowding and have the same nuclear area as S phase cells, which grow at high local cell crowding (region 2). (**b**) Nuclear area of G1, S and G2 phase decreases as local cell crowding increases. G1 cells growing at low crowding (box 1) have the same nuclear area (dashed line) as S cells growing at high local cell crowding (box 2). Points represent the median value in each of 12 bins based on degree of cell crowding; dark gray, 40th to 60th percentile; light gray, interquartile range. (**c**) Box plots comparing the distribution of nuclear area in crowded (green) or sparse (blue) areas, corrected (right) and uncorrected (left) for local cell crowding. a.u., arbitrary units. Box boundaries, 25th and 75th percentile; center lines, median; whiskers, 0.7th and 99.3rd percentile. (**d**) Correlation between CCTs built from two independent populations of cells (sets A and B), with or without taking into account local cell crowding. (**e**) Cells are color-coded for their position along the CCT (top) and on the abundance of PCNA and cyclin A (bottom). Scale bars, 50 μm. (**f**) Distribution of PCNA and cyclin A abundance, CCT corrected (red) and uncorrected (gray). Bar plots show the variance explained by the CCT and microenvironment features. LCC, local cell crowding; DTE, distance to edge; edge, position at the edge of population or not; ME, combination of local cell crowding, distance to edge and edge; combined, all features combined; corr., corrected; uncorr., uncorrected.
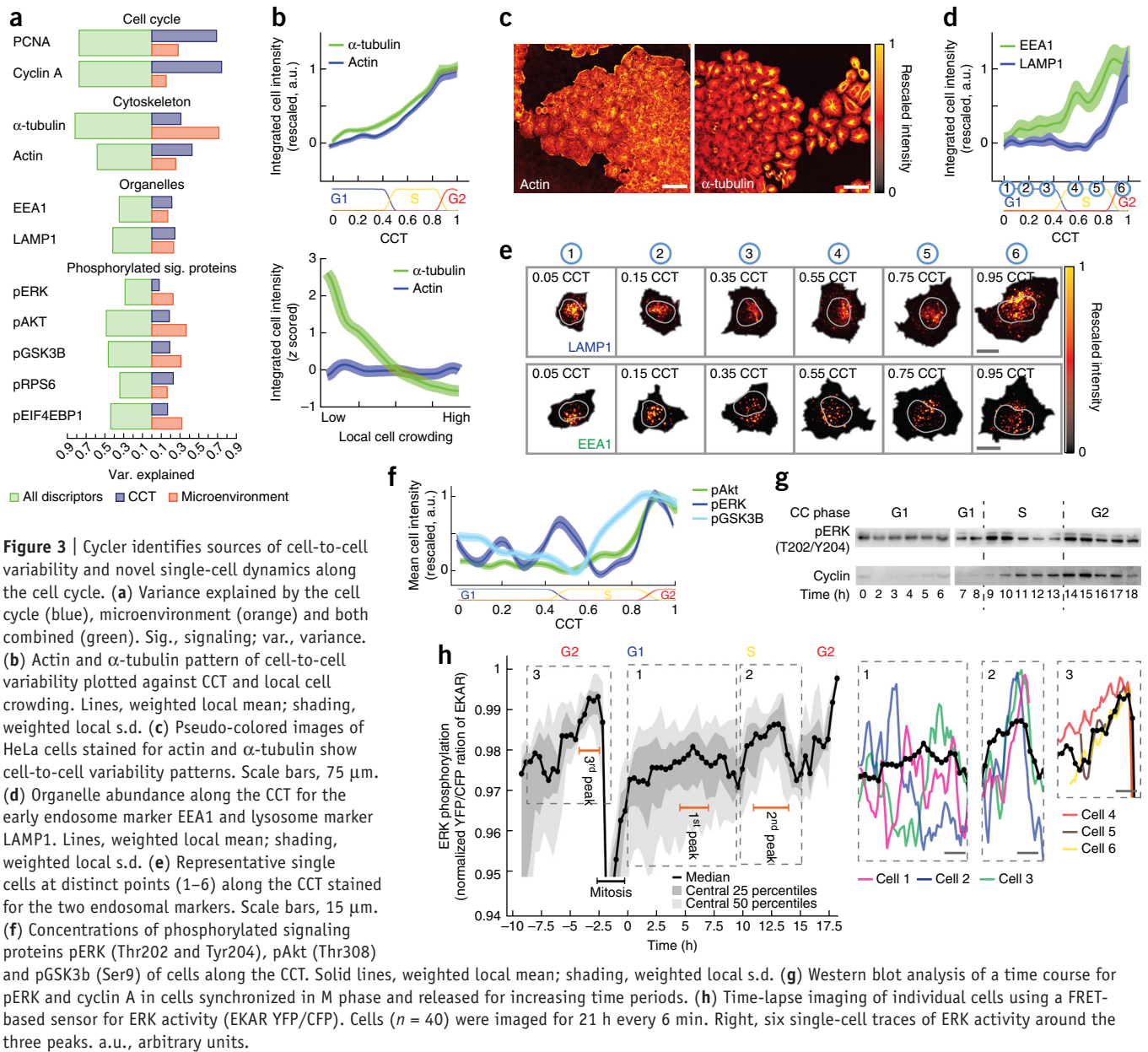
Figure 3 | Cycler identifies sources of cell-to-cell variability and novel single-cell dynamics along the cell cycle. (**a**) Variance explained by the cell cycle (blue), microenvironment (orange) and both combined (green). Sig., signaling; var., variance. (**b**) Actin and α-tubulin pattern of cell-to-cell variability plotted against CCT and local cell crowding. Lines, weighted local mean; shading, weighted local s.d. (**c**) Pseudo-colored images of HeLa cells stained for actin and α-tubulin show cell-to-cell variability patterns. Scale bars, 75 μm. (**d**) Organelle abundance along the CCT for the early endosome marker EEA1 and lysosome marker LAMP1. Lines, weighted local mean; shading, weighted local s.d. (**e**) Representative single cells at distinct points (1–6) along the CCT stained for the two endosomal markers. Scale bars, 15 μm. (**f**) Concentrations of phosphorylated signaling proteins pERK (Thr202 and Tyr204), pAkt (Thr308) and pGSK3b (Ser9) of cells along the CCT. Solid lines, weighted local mean; shading, weighted local s.d. (**g**) Western blot analysis of a time course for pERK and cyclin A in cells synchronized in M phase and released for increasing time periods. (**h**) Time-lapse imaging of individual cells using a FRET-based sensor for ERK activity (EKAR YFP/CFP). Cells (*n* = 40) were imaged for 21 h every 6 min. Right, six single-cell traces of ERK activity around the three peaks. a.u., arbitrary units.

Furthermore, we built a CCT with features of DNA replication obtained with PCNA immunostaining (**Supplementary Fig. 4a–c**), yielding identical results. This demonstrates that Cycler is robust to various markers of DNA replication. We could further improve the cellular staining protocol by combining the stains for cell segmentation and DNA replication in the same channel (**Supplementary Fig. 4d,e** and Online Methods). Finally, we compared Cycler with a previous approach to infer a CCT from images of fixed cells[7] and found that Cycler has higher accuracy, particularly in populations of cells with heterogeneous microenvironments (**Supplementary Fig. 5a** and Online Methods).

Cycler was designed to distinguish the contribution of the cell cycle to cell-to-cell variability. This variability, which may result in bimodal distributions of cellular activities, is often attributed to random fluctuations and noise in gene expression[10] and is therefore considered unpredictable at the single-cell level. Unsynchronized cell culture results in patterns of cell-to-cell variability in cell-cycle progression (**Fig. 2e** and **Supplementary Fig. 6a**). For instance, comparing the levels of PCNA and cyclin A protein abundance relative to their position along the CCT reveals specific correlations (**Fig. 2e**). Moreover, correcting these single-cell intensity distributions for cell-cycle progression (Online Methods) turns the bimodal distributions into narrow, unimodal distributions (56% and 86.1% interquartile range reduction for PCNA and cyclin A, respectively). As expected for cell cycle–related markers, the CCT explains the majority (~72%) of the observed cell-to-cell variability, whereas microenvironment explains only ~12% (**Fig. 2f**).

We next expanded this analysis to additional markers, including cytoskeleton components, markers of intracellular organelles and the phosphorylated state of kinases central to different signal transduction pathways (**Fig. 3a**). Collectively, the cell cycle and microenvironment explain 65–85% of cell-to-cell variability in

fixed cell populations. Notably, by correcting for cell cycle and microenvironment, we reduce the often complex, broad and bimodal single-cell distributions to simple unimodal distributions (**Supplementary Fig. 6b**).

Specifically, we found that actin and α-tubulin abundance increase similarly as cells progress through the cell cycle (**Fig. 3b**), reflecting a scaling with cell size (**Fig. 3a**). However, α-tubulin abundance also showed a strong adaptation to local cell crowding (**Fig. 3a,b**), whereas actin showed such adaptation mainly at the level of subcellular distribution and not abundance (**Fig. 3b,c**). Early endosome abundance increased linearly along the CCT, probably to accommodate an increase in cell–surface area turnover as cell size increases (**Supplementary Fig. 6c**). In contrast, lysosome abundance stayed stationary for almost 70% of the CCT and specifically peaked in G2 (**Fig. 3d,e**).

For glycogen synthase kinase-3β (GSK3B), we found that the concentration of its phosphorylated state, which reflects its inhibition, was lowest at the beginning of S, coinciding with its known peak activity in the degradation of cyclin D1 (**Fig. 3f**). After that point, phosphorylation of GSK3B increased toward the end of the cell cycle, suggesting the existence of a mechanism to inhibit GSK3B-induced protein degradation[11] as cells progress through S and G2. For phosphorylated AKT (pAKT), we found that its concentration along the CCT remained low throughout G1 and S, but strongly increased at 75% of cell-cycle progression with a phosphorylation peak in G2 (**Fig. 3f**)[12]. This may reflect AKT's role in stimulating cell growth during G2. Finally, Cycler revealed three peaks in the concentration of pERK—one at the beginning of G1, one early in S and one in G2 (ref. 13). This may reflect ERK's roles during early G1 progression and in activating the expression of cyclin D1 (ref. 14) (early G1), in cell proliferation (starting S phase)[15] and in cell division[13,15,16]. To validate this, we performed western blot analysis of a time course after mitotic release, as well as time-lapse imaging of individual cells using a fluorescence resonance energy transfer (FRET)-based sensor for ERK activity[17]. The western blots confirmed two of the three peaks (early S and end of G2) but did not reveal the first peak (early G1) (**Fig. 3g**). The FRET-based approach revealed all three peaks (**Fig. 3h**). However, because single cells progress with different rates through G1, they reach the position where ERK phosphorylation increases at different time-points, resulting in a relatively broad and unpronounced G1 peak in a cell-averaged time trace. This suggests another advantage of Cycler: because it uses single-cell ordering and not timing to infer a trajectory, it is less hampered by temporal heterogeneity in cell-cycle progression to detect transiently occurring events.

Cycler is a versatile method to infer accurate CCTs from fixed cell populations without transgenic markers and is generally applicable to different cell lines grown in heterogeneous microenvironments, outperforming previous methods[7]. Cycler creates a trajectory that reflects high-resolution order in cell-cycle progression. It can be used with a simple cellular staining protocol, allowing the simultaneous quantification of single cells' cell-cycle position with multiple cellular activities. In combination with features of the microenvironment[18], Cycler enables a comprehensive deconvolution of the dominant sources of cell-to-cell variability in a population of mammalian cells. Furthermore, Cycler allows a fast, robust and scalable analysis of a wide variety of single-cell activities along the cell cycle, which can be extended to the analysis of tissues and other single-cell methods. Finally, given its insensitivity to heterogeneous microenvironments[8,19], Cycler may be applied in large-scale genetic perturbation screens in human cells, allowing the identification of direct cell cycle–dependent perturbations of cellular activities. In summary, Cycler provides a powerful, accessible and versatile tool for the analysis of cellular heterogeneity.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS

L.P., P.L. and D.P. initiated the study. G.G., M.D.T., L.P., P.L. and D.P. designed and analyzed the experiments and wrote the manuscript. G.G. performed the experiments.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Held, M. *et al. Nat. Methods* **7**, 747–754 (2010).
2. Farkash-Amar, S. *et al. PLoS ONE* **7**, e48722 (2012).
3. Neumann, B. *et al. Nat. Methods* **3**, 385–390 (2006).
4. Roukos, V., Pegoraro, G., Voss, T.C. & Misteli, T. *Nat. Protoc.* **10**, 334–348 (2015).
5. Kutsuna, N. *et al. Nat. Commun.* **3**, 1032 (2012).
6. Shedden, K. & Cooper, S. *Proc. Natl. Acad. Sci. USA* **99**, 4379–4384 (2002).
7. Kafri, R. *et al. Nature* **494**, 480–483 (2013).
8. Liberali, P., Snijder, B. & Pelkmans, L. *Cell* **157**, 1473–1487 (2014).
9. Bendall, S.C. *et al. Cell* **157**, 714–725 (2014).
10. Eldar, A. & Elowitz, M.B. *Nature* **467**, 167–173 (2010).
11. Acebron, S.P., Karaulanov, E., Berger, B.S., Huang, Y.L. & Niehrs, C. *Mol. Cell* **54**, 663–674 (2014).
12. Liu, P. *et al. Nature* **508**, 541–545 (2014).
13. Roberts, E.C. *et al. Mol. Cell. Biol.* **22**, 7226–7241 (2002).
14. Lavoie, J.N., L'Allemain, G., Brunet, A., Müller, R. & Pouysségur, J. *J. Biol. Chem.* **271**, 20608–20616 (1996).
15. Tamemoto, H. *et al. J. Biol. Chem.* **267**, 20293–20297 (1992).
16. Shapiro, P.S. *et al. J. Cell Biol.* **142**, 1533–1545 (1998).
17. Harvey, C.D. *et al. Proc. Natl. Acad. Sci. USA* **105**, 19264–19269 (2008).
18. Snijder, B. *et al. Nature* **461**, 520–523 (2009).
19. Snijder, B. *et al. Mol. Syst. Biol.* **8**, 579 (2012).

ONLINE METHODS

Cell culture. HeLa Kyoto was used as main cell line for this study. The cells were tested for identity by karyotyping[20] and tested for the absence of mycoplasma before use. Further cell lines used were A431, RPEI, Hek293 and COS-7. All cell lines were cultured in DMEM (Gibco) supplemented with 10% FCS and glutamine (complete medium) at 37 °C, 95% humidity and 5% CO_2. For experiments cells were seeded at a density of 700 cells per well in 384-well plates and 1,500 in 96-well plates (Greiner) and grown for 3 d at 37 °C, 95% humidity and 5% CO_2.

Microscopy. Imaging was performed with an automated spinning disk microscope from Yokogawa (CellVoyager 7000) with an enhanced CSU-W1 spinning disk (Microlens-enhanced dual Nipkow disk confocal scanner, wide view type), a 40× Olympus objective of 0.95 NA, and Neo sCMOS cameras (Andor, 2,560 × 2,160 pixels), acquiring 20 z planes per site spanning 10 µm. The 20 z planes were then maximum intensity projected. Image analysis was performed on the resulting image. UV (406 nm) and far-red (647 nm) signals were acquired in dual camera mode; green (488 nm) signal and red (546 nm) signal were acquired separately. Cycler is also able to infer accurate CCTs from a wide-field microscope at 40× magnification (using an ImageXpress microscope from Molecular Devices), but also here, multiple z sections and maximum intensity projection are recommended to avoid problems due to non-flat bottoms of multiwell plates.

Immunofluorescence. All steps were performed at room temperature unless stated otherwise. Cells were fixed in 4% paraformaldehyde (Electron Microscopy Sciences) for 15 min. Cells were then permeabilized with 0.5% Triton X-100 (Roth) for 15 min, then washed once with wash solution (3% BSA from Sigma Aldrich in PBS). Click chemistry reaction for fluorescent labeling of incorporated 5-ethynyl-2′-deoxyuridine (EdU, Life Technologies) was then performed as described below. Subsequently to click chemistry, cells were incubated with blocking solution (5% FCS in PBS) for 1 h. Cells were subsequently incubated with primary antibodies (1:300 in blocking solution) overnight at 4 °C. Cells were then washed with PBS and incubated for 2 h with secondary antibodies (1:500 in blocking solution). Nuclei of cells were then stained with DAPI (Life Technologies), for 10 min at a concentration of 0.4 µl/ml. As a last step, cell outlines were stained with Alexa Fluor 647 NHS Ester (SE-af647, Life Technologies). The following primary antibodies were used: cyclin A (Abcam, ab7956), phospho-GSK-3β (Ser9)(Cell Signaling, #9323), α-tubulin (BioConcept,11H10), pan-actin (Cell Signaling, #8456), EEA1 (BD Biosciences), LAMP1 (BD Biosciences), phospho-S6 ribosomal protein (Ser235/236) (Cell Signaling, #4858), pEIF4BP1 (Thr37/46) (Cell Signaling, #2855), phospho-p44/42 MAPK (Erk1/2) (Thr202/Tyr204) (Cell Signaling, #9101) and phospho-Akt (Thr308) (Cell Signaling, #13038), PCNA (Cell Signaling, #13110). Secondary antibodies used were as follows: anti-mouse Alexa Fluor 488 (Life Technologies), anti-mouse Alexa Fluor 546 (Life Technologies), anti-rabbit Alexa Fluor 488 (Life Technologies) and anti-rabbit Alexa Fluor 546 (Life Technologies).

Costaining of DNA replication and cell outlines. The fluorescence signals from DNA replication (Click-iT EdU Alexa Fluor 647 Imaging Kit, Life Technologies) and the cell outline staining (Alexa Fluor 647 NHS Ester, Life Technologies) were both measured in the same acquisition channel (far red). In order to ensure robust cell segmentation and a trustworthy CCT, both staining protocols have been optimized: 15 min before fixation cell were incubated in a 100 µM EdU solution at 37 °C and 5% CO_2. The manufacturer's protocol for fluorescence labeling of incorporated EdU via click chemistry was adapted to perform the click reaction with residual volume in a multiwell plate by adding the reaction components at double the concentration suggested by the manufacturer's protocol. Cell outlines were stained with SE-af647 at a concentration of 0.3 µg/ml for 5 min in carbonate buffer.

Image analysis. All images were analyzed with the image analysis software CellProfiler[21] (http://www.cellcycler.org provides a full pipeline for image analysis). Nuclei were segmented using images from the DAPI staining. The cell outlines were detected using the watershed algorithm of the SE-af647 signal. Standard CellProfiler features for intensity, size, and texture of objects were then extracted for nuclei and cells. Data cleanup was performed using supervised machine learning with CellClassifier[3], by training an SVM to classify wrongly segmented objects and cells with immunofluorescence artifacts. These cells are then discarded from further analysis. Segmentation of cells and nuclei as well as measurement of single cell features were performed by standard modules of CellProfiler using the image analysis platform called IBRAIN[19]. We then provided Cycler with a multivariate feature vector consisting of five features obtained from CellProfiler modules (MeasureObjectIntensity, MeasureTexture, and MeasureObjectAreashape): nuclei area (corrected for local cell crowding, see below), nuclear DAPI content (integrated intensity), nuclear EdU content (integrated intensity), nuclear EdU texture feature 5 and 12 (InverseDifferenceMoment and InfoMeas1), which it uses to construct a cell-cycle trajectory (CCT).

Correction of nuclear area for local cell crowding. Nuclear area of cells was corrected for effects of local cell crowding by first sorting cells according to their nuclear area and then by applying a moving average filter with a size of 15 over the sorted cells. The resulting locally averaged nuclear area was then subtracted from the original nuclear area. These methods have been extensively used previously[8,18,19]. Nuclear area of cells was corrected for local cell crowding in order to use it as a feature in Cycler to construct CCT in population of cells growing in heterogeneous microenvironment, as local cell crowding has an effect on nuclear area (**Fig. 2a–c** and **Supplementary Fig. 3a**). Nuclear area was not corrected for local cell crowding to compensate for errors in nuclear segmentation in regions of high crowding, as nuclei with high local cell crowding were segmented with high accuracy (**Fig. 2a**).

Cycler. Cycler receives as input multivariate vectors of single-cell features extracted from microscopy images and uses a graph-based approach to embed these into one dimension representing the CCT, which is an assignment of a temporal position along cell-cycle progression for each cell. The earliest point has a score of 0, which corresponds to cells at the beginning of G1, and the last has a score of 1, corresponding to cells at the end of G2; the rest of the points fall between these two. Cycler is based on a previously published algorithm, Wanderlust[9], developed to recover

the developmental timeline from high-dimensional single-cell mass cytometry data of cells progressing through a cell lineage. Cycler is an implementation that changes the Wanderlust code in a number of ways to recover cell-cycle progression from features extracted from microscopy images.

Cycler constructs the CCT by taking advantage of gradual changes in cells' features, such as nuclear area and DNA content, as they transition from one cell-cycle phase to another. These gradual changes are captured in a *k*-nearest-neighbor graph, where each cell is represented as a node and is connected to its *k* neighbors, the cells most similar to it. Embedding the cells in such a graph structure introduces a new metric of graph geodesic distances between the cells: a cell's distance to another cell is the shortest-path distance between them in terms of 'walks', that is, transitions through neighbors, where each transition constitutes a gradual 'step' between similar cells. Cells that are near each other in their order along the cycle will have similar features and thus will be separated by a small number of close neighbors and the shortest-path distance between them will be low. This metric frees us from considering direct cell-to-cell distances, as those can be misleading in the setting of nonlinear relationships between features. Moreover, this metric is resilient to the technical noise that is ubiquitous in biological measurements.

The algorithm starts by constructing an ensemble *l*-out-of-*k*–nearest neighbor graphs (L-NNGs). First, a *k*-nearest-neighbor graph (K-NNG) is constructed by connecting each cell to its $k > l$ nearest cells by Euclidean distance. Edges connecting cells to their nearest neighbors are weighted by the Euclidean distance between them. Next, an ensemble of *ng* L-NNG graphs is constructed by iterating over each node in the K-NNG, randomly keeping only *l* of its *k*-nearest neighbors to produce an *l*-out-of-*k*–nearest-neighbor graph. A proposed positioning of the cells is computed from each graph, and the final CCT averages the positions for each cell from each of these graphs.

Cycler proceeds by iteratively calculating the trajectory in each graph separately: the position of each cell or target is first set to the shortest-path distance from *s*, a user-defined 'early-cell'. The farther away a cell is from the starting point, the farther it is along the trajectory. This initial positioning encapsulates the start point's perspective of the other cells' progression on the basis of their computed shortest-path distance from *s*. This initial estimate loses resolution and accuracy as the distances from the start point get larger. To increase the accuracy of the cell ordering along the entire trajectory, we randomly flagged a set of cells as waypoints to serve as reinforcements to the early cell: the position of each cell will now be calculated as the average of its distance from all of the waypoints. The target's position was further refined by averaging that initial estimate derived from the starting cell alone, averaged along with the perspective of the *nl* – 1 additional waypoints. The distances were weighed (described below) so that waypoints closer to the target contributed more to the calculation (as they are less susceptible to the noise inherent in the shortest-path distance). However, the waypoints are themselves cells. Therefore, their positions will change at each iteration after the refinement. Because all cell positions depend on waypoint positions, the refinement step was repeated with the new waypoint positions until the positions of all cells converge.

Tracing cell cycle from microscopy images presented new challenges, most notably the relatively high variance in the measured

dimensions at each point along the cell-cycle progression. To overcome these challenges, we implemented into Cycler three key changes to the original Wanderlust implementation, focusing on how waypoints are used to ensure an accurate trajectory; specifically, we changed how waypoints are selected and how they weigh in to determine cell positioning.

First we found a dense population of cells in the G1 phase, which can account for 60% of the cells. A uniform selection of waypoints based on the data results in a heavy bias of cells at earlier parts of the cell cycle, leaving later stages underrepresented. To ensure an even representation of waypoints along the entire cell cycle, cells were binned into ten bands on the basis of graph-geodesic-distances from the start point, and an equal number of waypoints was evenly sampled from each band and pooled. A user-specified number of waypoints was then sampled uniformly from this subset of points. An even representation of cells in S and G2 recovered the separation of those phases to a greater resolution.

Second, trajectory detection is more sensitive to noise and outliers in the reduced dimensionality of microscopy images. To filter measurement noise and reduce the risk of randomly selecting an outlier as a waypoint, the waypoints were refined using a median filter. For each waypoint, we selected its *k* nearest neighbors and computed their median; the waypoint was then replaced with the cell closest to this median. This procedure pseudo-code can be described as *K* nearest points = find-nearest-neighbors (data, waypoint, *k*); median of *k* nearest points = median(*K*-nearest-points); new waypoint = find nearest neighbor (data, median of *k* nearest points).

Median filter is known to de-noise data while preserving its structure and boundaries. A data point at the boundary can be thought of as a cell whose features are only marginally representative of our phenotypes (for example, G1 or S phase cell), and an outlier beyond this boundary is even more dissimilar. However, these outlier and borderline cells are few, and we therefore expected most of their neighbors to exhibit more typical behavior within the boundaries. The data or structure of the graph is not changed, but rather our waypoint is replaced with another data point, which better represents the data. If a randomly selected waypoint is an average representative cell, replacing it with one of its average neighbors will not have any effect. If an outlier or boundary cell is selected, it is replaced with one exhibiting a more typical behavior that is within the boundaries of the data. This method is often used in image and signal processing to remove signal spikes or salt-and-pepper noise while avoiding blur.

Last, we needed to determine the means and formula by which the waypoints refined the placement of all cells along the CCT. In Cycler, Gaussian weights are used to compute the weighted average. For a given target cell *t*, the weight for each waypoint is computed by the formula

$$w_{l,t} = \frac{e^{-d(l,t)^2 f\sigma}}{\sum_m e^{-d(l,m)^2/\sigma}}$$

where *l* is the waypoint, *d* is their graph geodesic distance, and σ is the s.d. of all distances for that waypoint. The CCT score for *t* is the weighted average over all waypoint distances

$$\mathrm{CCT}_t = \sum_l \frac{d(l,t)}{nl} w_{l,t}$$

In Cycler we used a Gaussian-weighting scheme, whereas in Wanderlust, the average is linearly weighted by the distance of the waypoint from the cell. In both cases, the farther away a waypoint is, the less weight it has on the positioning of the target cell t. This follows from the logic that distances computed on the basis of the shortest path are more accurate between cells in closer proximity. In Cycler, the contribution of each waypoint drops exponentially with its distance. This is to ensure that the positioning of each cell is most strongly influenced by cells in the same cell-cycle stage. For example points in G1 do not vote as strongly for the positioning of points in G2, and vice versa.

Supplementary Figure 5a demonstrates the impact of these changes and the improvement of Cycler trajectory detection over that of the original Wanderlust implementation. We saw improvement primarily for later stages of the cell cycle; for example, Wanderlust typically misplaces cells on the higher end of EdU intensity, relative to their cell-cycle stage, much later in the CCT (**Supplementary Fig. 5a**). However, the best indication for Cycler's improvement over Wanderlust was the former's increased robustness and reproducibility to replicates and down-sampling (data not shown).

How Cycler builds a cell-cycle trajectory. Cycler uses five single-cell features: integrated nuclear DAPI intensity, integrated nuclear EdU intensity, nuclear EdU texture feature #5, nuclear EdU texture feature #12 and nuclear area corrected for local cell crowding. The features were normalized and standardized by subtracting the second percentile and dividing by ninety-eighth percentile minus the second percentile. In Cycler, M-phase cells were excluded before the construction of the CCT, as they link early G1 cells to late G2 cells and decrease robustness in the building of the CCT. Cycler was run using the following parameters: $nl = 100$, $dist$ = euclidean, $ng = 5$, $l = 15$, $k = 8$. The user-chosen early cell was randomly selected from the population below the one-halfth percentile of DAPI intensity and nuclear size. The output cell-cycle trajectory was first ranked, placing the cells in consecutive order, then normalized, giving each cell a position in the [0 1] range, by dividing the cell position by the total number of cells.

Robustness of Cycler to the selection of the start population. To test the influence of the start-cell parameter on the Cycler algorithm output, we used Cycler output from **Figure 1b** as a baseline trajectory and re-ran Cycler ten times. Each time, the input start cell was shifted by 0.1 across the baseline trajectory. We plotted the output of each rerun against the baseline trajectory and computed their correlations (**Supplementary Fig. 2b**). As long as the start cell remained in the first 50% of the baseline trajectory, the two trajectories coincide (Pearson's $\rho > 0.99$). In the context of cell cycle, the first half of the cycle is typically all of G1 phase and beginning of S phase, implying that any G1 cell can be used to recover the cell-cycle trajectory with confidence. Moreover, when starting from a G2 cell, Cycler can recover an accurate reverse trajectory going from G2 toward G1 (**Supplementary Fig. 2b**, start = 0.9 CCT).

Code availability. Custom code used for this study is available at http://www.cellcycler.org.

Feature traces along the cell-cycle trajectory. First, 100 positions were marked and uniformly distributed along the CCT. For each point a Gaussian filter, centered at that point, was used to compute the weighted average of each feature. This resulted in smooth traces of the features along the cell-cycle trajectory. A weighted s.d. at each location along the trajectory was computed using the same Gaussian weights and represents the error along the feature traces.

Classification of discrete cell-cycle phases: G1, S, G2 and M. Classification into the four cell-cycle phases was achieved through a combination of two SVM classifiers to identify S phase and M phase cells and a Gaussian mixture model to distinguish between G1 and G2 among the remaining cells. CellClassifier[22] was used to train an SVM, which classified cells as S phase or non–S phase cells. The feature set used by the S phase SVM consisted of nuclear EdU intensity and nuclear EdU texture features. For the classification of mitotic cells, a second SVM was trained using nuclear DAPI intensity and nuclear DAPI texture features. For the classification of G1 and G2, cells classified in S phase or M phase were excluded from the data set. The resulting histogram of integrated nuclear DAPI intensities consisted of two distinct and separated Gaussians, with the median of second Gaussian at the double of the intensity of the median of the first Gaussian. Cells were then sorted into two classes by k-means clustering. Cells comprised in the first, low-intensity Gaussian, were classified as G1 cells; cells that were part of the second, high-intensity Gaussian, were classified as G2 cells.

The Matlab-based graphical (http://www.cellcycler.org) user interface (Cycler GUI) accompanying this publication provides a different cell-cycle phase classification utility to simplify the usage of our method by encompassing more of the pipeline within a single tool. The alternative classification is done over five features: integrated nuclear DAPI intensity, integrated nuclear EdU intensity, nuclear EdU texture feature #5 and #7 and nuclear EdU texture feature #12. First, the cells are clustered into four groups by expectation maximization algorithm for Gaussian mixtures (EMGM). S cells are identified by selecting the cluster with the highest mean of EdU. M cells are then identified as the cluster with the highest mean of EdU texture feature #5. G1 and G2 phases are delineated from the remaining two clusters over DAPI intensity The cluster with higher DAPI intensity is labeled G2. 2D kernel density estimation is computed for each cell over the DAPI and EdU intensities, and the lower fifth percentile of the cells are discarded. This approach was tested against the SVM approach on 12 wells that were classified using both methods and achieved an average rand index of 0.89.

Correction of cell activities for effects of cell-cycle progression or microenvironment. The average of each feature of interest (for example, nuclear area or integrated intensity) was calculated along the CCT or the local cell crowding for overlapping bins, each consisting of 15 cells. For each single cell, the corrected measurement consists of the raw measurement minus the mean of all bins of which the cell is part.

Live cell imaging and tracking. Cells were seeded in a 384-well plate and grown for ~24 h. Then Hoechst was added at a

concentration of 800 nM and incubated for 1 h. No additional Hoechst was added during live imaging. The cells were then imaged at 37 °C and 5% CO_2 for ~72 h every 20 min with a 405-nm laser, acquiring a single image per site using automated spinning disk microscope from Yokogawa (CellVoyager 7000) with an enhanced CSU-W1 spinning disk (Microlens-enhanced dual Nipkow disk confocal scanner, wide view type), a 40× Olympus objective of 0.95 NA, and Neo sCMOS cameras (Andor, 2,560 × 2,160 pixels). Nuclei were segmented using the Hoechst signal and tracked by a custom Matlab function based on the nuclear segmentation. Because single tracks can be different in length, interpolation was applied to individual tracks to generate a set of equally long tracks using the interp1 function of Matlab. To compare how similar nuclear area growth computed by Cycler was to actual nuclear area growth of individual cells, both the tracked nuclear area along the cell cycle (dynamic data) and nuclear area along the CCT (static data) were binned into 50 equally spaced 'pseudo–time points'. The median nuclear area of those bins was calculated, and the two resulting time series were scored for their correlation. Our comparison resulted in a correlation of 0.91 ± 0.013 (median ± s.e.m.), demonstrating congruency between the ordering imputed by Cycler and the dynamics of live cell imaging.

Building cell-cycle trajectories for multiple cell lines. A431, HEK293, RPEI and COS-7 cells were seeded in 96-well plates and grown for ~72 h. EdU incorporation, immunofluorescence against cyclin A and PCNA as well as DAPI and SE-af647 staining were performed (see above). The four cell lines were imaged using an automated spinning-disk microscope (specifications described above). Classification of the discrete cell-cycle phases (SVM) and the construction of the cell-cycle trajectories were performed separately for each cell line and according to the descriptions above. Parameters to construct the CCTs were unchanged between HeLa (main cell line used in this study) and A431, HEK293, RPEI and COS-7 cells. The start population for each cell line was gated individually, but each consisted of G1 cells with the smallest nuclear area and the lowest DNA content.

M phase synchronization experiment. 800 cells per well were seeded in a 384-well plate and grown for 24 h at 37 °C. To synchronize all cells of a well in M phase and minimize cytotoxic effects, we first incubated cells in 2 mM thymidine (Sigma-Aldrich) for 24 h, blocking cells at the onset of S phase. Complete release of the population of cells from S phase block was achieved by extensive washing with complete medium. 3 h after release, cells were incubated in 25 ng/ml nocodazole (Sigma-Aldrich) in complete medium and 1% penicillin-streptomycin for 12 h, blocking cells in M phase. Cells were then released for 0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 or 19 h and fixed.

Reproducibility analysis. For two independent sets (A and B) of 20,000 randomly selected cells, we constructed two cell-cycle trajectories using two different feature sets (X and Y). Both feature sets X and Y consisted of integrated nuclear DAPI intensity, integrated nuclear EdU intensity, nuclear EdU texture feature #5 and nuclear EdU texture feature #12. Feature set X included nuclear area, whereas Y included nuclear area corrected for local cell crowding. This resulted in four cell-cycle trajectories: xa (CCT constructed with set A by feature set X); xb (CCT constructed

with set B by feature set X) ya (CCT constructed with set A by feature set Y) and yb (CCT constructed with set B by feature set Y). For every cell in set A, the five most similar cells in B were found by proximity search using feature set X or Y and Euclidean distance as distance metric. The positions along their corresponding CCTs (xb or yb) of the five found cells were averaged and saved in xm and ym. This was repeated for every cell in A, resulting in two vectors, xm and ym, with the same length as xa and ya. To quantify the reproducibility between two independent CCTs, Pearson's correlations were computed between xm and xa and ym and ya.

Minimal number of cells required for robust CCTs. To quantify the minimal cell number required to construct a robust CCT over the whole interphase (G1, S and G2), we first constructed a CCT with 24,000 cells. Then we measured the correlation between the position of cells in the CCT constructed with 24,000 cells (full populations) and their position in CCTs constructed with lower cell numbers (20–20,000 cells, CN). At each CN (for example, 400), we randomly subsampled the corresponding number of cells (for example, 400) from the full population of 24,000 cells and constructed a CCT. The CCT was constructed with the = input parameters $l = 15$, $k = 8$, #waypoints = 100 (if #waypoints < 1,000, #waypoints = cell number / 10). We then measured the correlation between the CCT value for every cell in the generated CCT from low CN (for example, CCT of 400 cells) and the CCT value of the same cell in the CCT generated from the full population. This was repeated 50 times per CN.

To quantify correlations per cell-cycle phase, we followed the same procedure, as for the quantification of correlation over interphase (see above). However, instead of quantifying correlation over G1, S and G2 combined, we calculated it for each cell-cycle phase separately. As we randomly subsampled cells for every CCT, the proportion of G1, S and G2 may vary between each subsampling. We calculated the mean cell number per phase at any CN and used it as the x-axis label for the representation of the correlation of a CCT in individual cell-cycle phases (**Supplementary Fig. 3e**).

Construction of the CCT with PCNA staining. To build the cell-cycle trajectory with the PCNA antibody staining instead of the fluorescently labeled EdU staining as S phase marker, we provided Cycler with the following features: integrated nuclear DAPI intensity, integrated nuclear PCNA intensity, nuclear PCNA texture feature #5, nuclear PCNA texture feature #12 and nuclear area corrected for local cell crowding.

Comparison between Cycler and density-based detection. A popular method to detect trajectories is based on tracking peak density. The approach requires two features that change smoothly over the course of the cell cycle. The 2D density plot of these two features creates an arc-shaped ridge (**Supplementary Fig. 5a**), where the cell cycle progresses through this ark. Density-based methods aim to estimate the peak density along this arc, under the assumption that this path of peak density represents the average cell behavior along the cell cycle. This method was successfully used previously in Kafri *et al.*[7] to create a cell cycle-based ordering of the cells, which was needed as input for their ERA (ergodic rate analysis) procedure. In Kafri *et al.*[7] a two-dimensional

density-based method is used to extract cell-cycle ordering of homogenously plated cells (for example, no cell crowding) using two features: nuclear integrated DAPI and Geminin (mAG-hGem) intensity measurements. The authors note[7] that homogenous conditions are important for success of the method. We compared this density-based method to Cycler over two data sets for which we measured two features—integrated nuclear DAPI intensity and integrated nuclear EdU intensity—each containing approximately 30,000 cells with heterogeneous microenvironments (densely and sparsely growing cells). We ran Cycler using the parameters described above (details for density-based trajectory detection are described below). The 'early point' was selected to the lowest intensity over both features.

In our comparison, Cycler outperformed density-based methods in the conditions tested. As previously reported[8], density-based methods cannot account for heterogeneous microenvironments (crowding of cells), whereas Cycler robustly dealt with such images (**Fig. 2a**). We note that in this comparison we limited Cycler to only two dimensions, whereas Cycler may aggregate as many features as desired, thus achieving higher and higher resolution (we used five features). As dimensionality increases, density-based methods become prohibitively computationally expensive. Additionally, as dimensionality increases, density detection's sensitivity to noise and minor fluctuations in the data further intensifies. Cycler, on the other hand, is robust to increased dimensionality and microenvironment, making it a versatile method that works well on imaging data of all sorts, such as images from tissue, 2D and 3D cultures both at high and low magnification. Finally, where density detection starts at the point of highest density, which is already in the middle of G1, Cycler can begin from the earliest cell, revealing greater detail during the start and end of the trajectory.

In **Supplementary Figure 5a** are presented two cases where density detection computed a CCT. For HeLa cells, density detection failed to exit G1 and the computed ridge (black line) was an infinite loop within the peak density of G1, thus failing to find the cell-cycle position for the majority of cells. For A431 cells, density detection looped back in the middle of S phase (black line) failing to follow the cell cycle through S and G2.

Computing trajectory based on density based detection. We used Matlab code provided in the supplement of Kafri et al.[7], which accepts as input a list of cells with their measured integrated nuclear DAPI intensity and integrated nuclear EdU intensity and outputs a one-dimensional vector of points, the ridge *l* that represents the path of an 'average' cell. The ridge attempts to trace the space following the path of highest density, but is susceptible to being misled by local maxima in the population density. Following the application in Kafri et al.[8], to order the input cells along the cell cycle, each cell is then associated with the point on *l* to which it is closest by based on Euclidean distance. We note that this is not the ERA method provided in that paper, but rather an earlier preprocessing step that provided a good implementation of density-based trajectory detection.

We made two changes to the code. First, we supplied a standard implementation of bin_raw, as this method was not supplied in the previously published code[7]. Given the data and the requested bin indices, bin_raw counts the number of points per bin. Second, multiple places in the code called a function

switchcoordinates(y, x,……), yet the function header was defined with switched parameter order switchcoordinates(x, y,……). The function header was reordered. To run the method we selected a start and end from the computed 2D density plot of the data, as directed by the accompanying GUI. The start area was selected at the point of low integrated nuclear DAPI intensity and low integrated nuclear EdU intensity—the start of G1—and the end area was selected as a point of high integrated nuclear DAPI intensity and low integrated nuclear EdU intensity (end of G2). The code does not require any further parameterization. At times the code would infinitely loop in G1 or would locate the center of G1 and trace to the beginning of G2, loop back and fail to converge. We manually halted the algorithm after either a complete loop in G1 or when the results started looping back after S.

Variance explained by the cell-cycle trajectory and microenvironment. The variance of the uncorrected population of cells for a cellular activity of interest (var1) was measured. The cell activity was corrected for effects of the cell cycle or the microenvironment (see above) and its variance was calculated (var2). Variance explained by the cell cycle or the microenvironment was then calculated as follows:

$$1 - \frac{var\,2}{var\,1}$$

Western blots of synchronized HeLa cells. HeLa cells were synchronized in M phase and released for 0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17 or 18 h 10-cm dishes. Sample preparation at every time point was performed as follows: cells were washed twice with PBS, then all liquid was aspirated from the dish and cell plates was shock frosted and at −20 °C. Lysates were prepared on ice by adding 100 μl of lysis buffer (150 mM NaCl, 50 mM HEPES, 1% Triton X-100, 0.1% SDS (Sigma-Aldrich) in miliQ H_2O. Just before use,10 ml lysis buffer was supplemented with 1 tab protease inhibitor cocktail (cOmplete Roche), 3 tabs phosphatase inhibitor cocktail (PhosSTOP Roche), 2 mM DTT and 5 mM EDTA and added to one 10-cm dish. All cells were scraped from the plate and incubated 30 min on ice (with 15 s sonication at 15 min and 30 min of incubation). Lysates were then centrifuged at 21,000 RCF for 15 min at 1 °C and the supernatant transferred in to a new tube and stored at −80 °C. Proteins in lysate were denatured by the addition sample buffer and boiling for 10 min. Proteins in lysates were resolved by 10% SDS-PAGE and analyzed by immunoblotting. After wet transfer of the protein onto a PDVF membrane (Immobilon), membranes were incubated in 5% lowfat milk for 1 h at room temperature. Membranes were sequentially probed for ppERK1/2 (1/1,000, over night at 4 °C, Cell Signaling, #9101) or cyclinA2 (1/500 overnight at 4 °C, Abcam, ab7956) in 5% lowfat milk. Secondary antibodies were incubated for 2 h at room temperature at 1/5,000 in 5% lowfat milk.

Long-term FRET imaging of the EKAR (YFP/CFP). EKAR Cerulean-Venus with nuclear localization was obtained from Addgene[17] (Plasmid #18681). HeLa cells were reverse transected 24 h before imaging with Lipofectamine 2000 (Life Technologies) in 8-well Lab-Tek chambers at 13,000 cells per well. Cells were imaged every 6 min for 21 h at 37 °C and 5% CO_2 using an automated spinning-disk microscope from Nikon (Eclipse Ti) with

an enhanced CSU-W1 spinning disk (Microlens-enhanced dual Nipkow disk confocal scanner, wide view type), a 40× Nikon objective of 0.95 NA (CFI Plan Apochromat λ 40X), and pco.edge sCMOS cameras (pco.edge 5.5, 2,560 × 2,160 pixels), acquiring 1 z plane per site. CFP (406 nm) and YFP (561 nm) signals were acquired in dual camera mode. The YFP image was then divided with the CFP image resulting in a ratio image, which was used to quantify FRET of EKAR. For each cell a circular region was selected in the nucleus, and its mean intensity of the ratio image (YFP/CFP) was quantified for every time point using ImageJ.

Every single-cell trace was then normalized by dividing every intensity value of a trace by its maximal measured intensity value. Then single-cell traces were aligned so that mitosis is positioned at 0 min.

20. Battich, N., Stoeger, T. & Pelkmans, L. *Nat. Methods* **10**, 1127–1133 (2013).
21. Carpenter, A.E. *et al. Genome Biol.* **7**, R100 (2006).
22. Rämö, P., Sacher, R., Snijder, B., Begemann, B. & Pelkmans, L. *Bioinformatics* **25**, 3028–3030 (2009).