Supplementary Material

1 SUPPLEMENTARY DATA

1.1 *btrack* installation

The latest stable release of the package can be installed using: `pip install btrack`

1.2 Configuration and documentation

We refer users to the `CONFIGURATION.md` file in the software repository[1] for detailed descriptions of hyperparameters that can be tuned in the *btrack* package. We encourage users to adapt these hyperparameters to their specific datasets. Detailed documentation and source code can also be found in the repository.

1.3 Hypothesis Generation

The *btrack* Hypothesis Engine creates hypotheses for each of the tracklets in order to perform a global optimization. All tracklets are assigned a default hypothesis, that they are a false positive detection, where $P(\text{FP}) = m^{\ell}$ and m is the segmentation miss rate ($m < 1$) and ℓ the length of the tracklet ($\ell > 1$) (Bise et al., 2011). Intuitively, this means that longer tracks have a lower probability of being a false positive detection. It follows that the true positive probability can be defined as $P(\text{TP}) = 1 - P(\text{FP})$. For all other hypotheses, the hypothesis engine uses tracks intializing or terminating within a certain spatiotemporal bin to decide which hypotheses should be generated. Here, we provide a description of two example hypotheses below.

1.3.1 Example: Broken track linking hypothesis

In a linking hypothesis, the engine finds all tracklets starting within a spatiotemporal bin surrounding the end of a tracklet. For each initializing tracklet (t_j) within that window, a linking hypothesis is proposed, originating from the terminating tracklet (t_i). We calculate:

$$\begin{aligned} P(\text{link } t_i \rightarrow t_j) &= \exp^{(-d/\lambda_{\text{link}})} \\ \rho &= \log P(\text{link } t_i \rightarrow t_j) + \sum_{t \in \{t_i,\, t_j\}} 0.5 \log P(\text{TP } t) \end{aligned} \tag{S1}$$

Where d is the Euclidean distance between the last observation of tracklet t_i and the first observation of t_j. The probability distribution is scaled by the hyperparameter λ_{link}.

1.3.2 Example: Branching hypothesis

In a branching hypothesis, for example mitosis, the engine finds all tracklets starting within a spatiotemporal bin surrounding the end of a tracklet. If there are greater than two initializing tracklets, then for each possible pair of initializing tracklets t_j and t_k, a branching hypothesis is proposed, originating from the terminating tracklet t_i. We calculate:

[1] `https://github.com/quantumjot/BayesianTracker`

$$P(\text{branch } t_i \rightarrow t_j, t_k) = \exp^{(-d/\lambda_{\text{branch}})}$$
$$\rho = \log P(\text{branch } t_i \rightarrow t_j, t_k) + \sum_{t \in \{t_i,\, t_j,\, t_k\}} 0.5 \log P(\text{TP } t) \tag{S2}$$

Where d is a distance representing the angle between the daughter cells (t_j and t_k) and the parent cell (t_i) and the respective states of each of the cells. The probability distribution is scaled by the hyperparameter λ_{branch}. The intuition is that during mitosis, the daughter chromosomes are axis-aligned with the parental metaphase plate (hence the angular metric) and that metaphase preceeds anaphase (hence the states of each cell are used).

For a full description, we refer the reader to the source code for all hypotheses generation, in the file `btrack/src/hypothesis.cc`.

1.4 Workflow Error Analysis

One of the major challenges of cell tracking is the propagation of errors throughout the pipeline. Each step gives rise to the possibility of incorporating errors such as premature track breakage, misidentification of cell splitting events and incorrect assignment of parent-children relationships upon lineage reconstruction. Because accumulation of small unaddressed errors over many time points can lead to a substantial drop in tracking accuracy, we performed a rigorous assessment to test our pipeline performance.

We define 4 different types of observations in the tracking output:

- *true positive* (TP, hit), which represents a true detection where the same object is found in the output and manual annotation
- *false positive* (FP, ghost), which represents a detection in the computer-generated output that does not appear in the manual annotation
- *false negative* (FN, miss), which represents an undetected object in the computer-generated output which is present in the manual annotation
- *identity swap* (IS, mismatch), which represents a tracking-related error where (i) the unique identity of two tracked trajectories are swapped, most likely due to close proximity or with crossing trajectories, (ii) the incorrect linkage event leads to assignment of a new ID label to an existing track, or (iii) the misidentification of mitosis, where one of the children cells becomes falsely concatenated to the parent cell.

1.5 Cell Detection Assessment Metrics

We selected representative time-lapse microscopy movies, and for each, shortlisted 3 representative frames capturing cells grown to low, medium and high confluency and manually annotated the nuclear areas to calculate the fidelity of cell detection and localisation (*Supplementary Table S1*).

To assess the quality of the cell detection step, we measured the multi-object tracking precision (MOTP, Eq S3) in cells appearing in three manually labelled frames capturing fluorescently labelled cell nuclei sampled at low, medium and high confluency. Out of 869 cells in total, 847 cells had their centroid coordinates estimated within threshold of 20 pixels from designated cell in ground truth. Our pipeline localised the cells with highest precision, with mean localisation error of 0.99 pixels.

In contrast to object detection algorithms, our localisation method performs a pixel-wise image classification, which offers the opportunity to segment the whole area belonging to the cell nucleus from which centroid coordinates are calculated. To calculate the accuracy of the nucleus area segmentation, we computed the per-object intersection over union metric (IoU, Eq S4) of the areas for each individual nucleus from the U-Net and how they compare to the manually labelled ground truth areas. We introduced a strict IoU thresholding (0.5 to 0.9 with 0.1 increments, *Supplementary Table S2*), i.e. we only scored a nucleus as hits (true positive, TP) when the computer-generated mask overlapped with the ground truth label by at least 50% (IoU 0.5), and above. We report that out of 616 ground truth nuclei in a fully confluent FoV, 574 nuclei were sharing areas with at least 50% overlap, yielding a Jaccard Index of 0.933 (93% of objects detected in the FoV). As expected, the calculated localisation error was decreasing as only the highest overlapping cells are included in calculation of progressively increasing IoU threshold values (*Supplementary Table S2*). More details can be found at: `https://github.com/quantumjot/unet_segmentation_metrics`.

Below we provide the equations to calculate multiple cell segmentation, cell classification and cell tracking metrics to assess the performance of our workflow.

Multi-Object Tracking Precision (MOTP):

$$MOTP = \frac{\sum_t \sqrt{(x_{GT} - x_{TR})_t^2 + (y_{GT} - y_{TR})_t^2}}{\sum_t (objects_{TP})_t} \tag{S3}$$

Where the Euclidean distance between centroid (x,y) coordinates of manual annotations (GT) and tracker outputs (TR) is computed and divided by sum of all within distance threshold *d* of 20 pixels away from ground truth (GT) observations at time *t*. This metric represents how precisely tracking algorithm can determine the position of an object. It is the ratio of the total error in position to the number of true positive correspondences between the tracker and manual annotation.

Intersection over Union (IoU):

$$IoU = \frac{\sum TP}{\sum (TP + FP + FN)} \tag{S4}$$

Where the ratio between common shared pixels (true positive, TP) and sum of pixels of two segmented objects, including the shared pixels (true positive, TP, false positive, FP, false negative, FN) is computed per single object between the observations and the ground truth.

Jaccard Index (J):

$$J = \frac{\sum TP}{\sum (TP + FP + FN)} \tag{S5}$$

Where the ratio between the correctly identified cells (true positive, TP) and sum of all cells in the FoV, including the correctly localised cells (true positive, TP), ghost cells (false positive, FP) and missed cells (false negative, FN) is computed across multiple whole FoVs between the observations and the ground truth.

Pixel Identity (PI):

$$PI = \frac{\sum TP}{\sum (TP + FP + FN)} \tag{S6}$$

which represents the number of pixels found in both the segmentation and the ground truth which have the same label.

Per-class F1-Score (F1):

$$F1 = \frac{\sum TP}{\sum TP + \frac{1}{2}(\sum FP + \sum FN))} \tag{S7}$$

which represents a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal.

1.6 Cell Tracking Assessment Metrics

We manually reconstructed the ground truth lineage trees of 24 founder cells (black; *Supplementary Figure S1*) and contrasted the observations to automated reconstructions by *btrack* (pink; *Supplementary Figure S2*) and by TrackMate (brown (*Supplementary Figure S3*). These trees represented more than 1/3 of initially seeded cells in a representative movie. The subsequent progeny of these founder cells comprised 1,032 cells (including tree founders) spanned the entire movie duration.

Next, we calculated the multi-object tracking accuracy (MOTA, Eq S8) which scores the tracker's ability to retain cell's identity and trajectory over longer periods of time, defined as:

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IS_t)}{\sum_t (objects_{GT})_t} \tag{S8}$$

Where sum of all trajectory observations subjected to tracking errors (FN, FP and IS) at short time-lapse sequences of length t is divided by the total number of observations in the ground truth (GT) at time series t. Time series t = 20 frames. This metric represents the errors associated with detecting objects and accurately keeping track of them, independent of the ability to precisely localise them.

MOTA score intrinsically penalises the tracking pipeline for static (ghost or missed objects) as well as dynamic (identity switches) errors, which often strongly rely on the detection algorithm performance. We used short movie sequences of up to 20 successive fields of view at three confluency levels as previously. Observing 2,161 individual cell objects over time, our tracking algorithm performed at very high accuracy of 97.66%.

Localization accuracy, as determined by the different algorithms used, accounts for the majority of differences between the MOTA scores for different tracking packages. Importantly, our library is agnostic to the segmentation and localization methods used to identify the cells in the image data. The requirement for the subsequent tracking step is to process the raw sequence of input images into a vector of localisation coordinates of individual objects appearing in each field of view.

Optionally, it is desirable to describe each object with a numerical label which corresponds to the cell state based on its progression towards mitosis (0 - interphase, 1 – pro-/metaphase, 2 - metaphase, 3 – ana-/telophase, 4 - apoptosis), which is an additional input to the tracking algorithm for improved performance.The simple structure of tracking data input enables users to choose, or design, a pipeline appropriate to their data and integrate the tracking algorithm directly into their image analysis pipeline.

1.7 Lineage Tree Reconstruction Fidelity

We used a representative movie comprising 1193 frames with 1600×1200 pixels for cell detection by the tracker. The localisation blob diameter was optimised for the entire movie on 3 fields of view representative of 3 confluency levels (low at 14%, medium at 41% and high at 98% of cell density).

We use the following metrics to validate the lineage tree reconstruction. The Mitotic Branching Correctness (MBC), a measure of the ability of the tracker to identify mitoses, is calculated as:

$$MBC = \frac{\sum TP}{\sum(TP + FP + FN)} \tag{S9}$$

Where true positive (TP) mitotic events are described as track splitting events identified within time distance t of the ground truth (GT) (Supplementary Figure S4). We use a strict threshold of t = ±1 frame. To consider the mitotic event as a TP, both the parent (single dividing track) and progeny (two newly appearing tracks) cells must exist in a correct generational depth relative to the tree root (founder cell). False positive (FP) mitoses were calculated as the difference between the total events detected by the tracking algorithm minus total TP events. False negative (FN) mitoses represent the difference between ground truth mitoses count and total TP events. In addition, we calculate the Leaf Retrieval Score (LRS), a measure of the number of correctly recovered tracks at the end of the movie, is calculated as:

$$LRS = \frac{\sum TP}{\sum(TP + FP + FN)} \tag{S10}$$

Where leaves (a terminal cell of a tree which doesn't further divide) are considered as TP when its entire lifetime matched the GT observation. Additionally, an exclusion of all leaves which were not followed until the last movie frame (Supplementary Movie S1) was introduced to only include TP leaves as those appearing at the movie end. FP leaves were calculated as the difference between the total leaves detected by the tracking algorithm minus total TP leaves. FN leaves represent the difference between the ground truth leaf count and total TP leaves.

The recall (also known as Target Effectiveness), is essentially the ability of the tracker to correctly recall trajectories present within a lineage tree, and is calculated as:

$$Recall = \frac{\sum(observations_{assigned})_{TR}}{\sum(observations_{total})_{GT}} \tag{S11}$$

Where the number of assigned track observations of the target is divided by the overall number of frames in ground truth target. Finally, we calculate the precision (also known as Track Purity), which represents the ability of the tracker to reconstruct the trajectories present within the lineage tree, as is calculated as:

$$Precision = \frac{\sum(observations_{total})_{TR}}{\sum(observations_{assigned})_{GT}} \tag{S12}$$

Where number of total ground truth tracked frames is divided by the overall number of assigned track observations followed by the tracker. The track purity score is designed to account for FP events and expresses the precision of the tracker.

1.8 Benchmarking Standards Calibration

TrackPy: We selected the 23-pixel diameter estimate which yielded the most satisfactory results (66, 180, 574 detected objects vs. 68, 185, 616 objects found in ground truth). All objects below the 'minmass' of 500 were filtered to exclude ephemeral blobs from the analysis. For track linking, the maximum displacement was specified to be 25 pixels and the memory for missed detections was set for 3 frames. The tracking yielded 4,521 trajectories in total, out of which 1,010 trajectories were tracked between the range of 7-42 hours.

TrackMate: The representative movie had the 'z' and 't' dimensions swapped upon loading to TrackMate. The blob detection calibration settings remained unchanged as in default (pixel width, height and depth equal to 1.0 pixel, time interval of 1 frame). Downsample Laplacian of Gaussian (LoG) filter was used for detection, with the sigma suited to the blob estimated size. Estimated diameter was chosen to be 30.0 pixels with threshold set to 0.0 (default), and downsampling factor of 4.0. Initially, 1,302,106 blobs were identified throughout the movie, subjected to initial thresholding for quality of 0.73 to retain 377,291 detected objects. For subsequent tracking step, the Linear Assignment Problem (LAP) mathematical framework was selected. Parameter for maximal distance in frame-to-frame linking was set to 30.0 pixels, track segment gap closes option was allowed and set to maximum distance of 15.0 pixels and 2-frame gap. Track segment splitting was allowed below the 25 pixels threshold. No feature penalties or additional track filtering were introduced in the tracking pipeline. Due to large number of lineage trees in the tracking output, the attempt to visualise all of the lineage data at once using the in-built 'TrackScheme' GUI was unsuccessful. The results tables produced via 'Analysis' option were saved out and computationally processed outside of the TrackMate interface using a custom-written software to reconstruct trees from linked spots information (Supplementary Figure S6). The tracking yielded 2,684 trajectories in total, out of which 1,264 trajectories were tracked for the duration between 7-42 hours with no splitting events.

2 SUPPLEMENTARY FIGURES

Supplementary Figure Legends:

Figure S1 — 2D Tree Representations of 24 Human Annotated Cell Lineages of a Representative Movie. Y-axis represents the time elapsed from start of time-lapse imaging. Vertical lines correspond to cell cycle duration of the particular cell. Horizontal lines represent track splitting (cell division) events.

Figure S2 — 2D Tree Representations of 24 Automatically Reconstructed Cell Lineages from a Representative Movie by our custom-designed bTrack pipeline. Y-axis represents the time elapsed from start of time-lapse imaging. Vertical lines correspond to cell cycle duration of the particular cell. Horizontal lines represent track splitting (cell division) events.

Figure S3 — 2D Tree Representations of 24 Automatically Reconstructed Cell Lineages from a Representative Movie by the benchmarking TrackMate pipeline. Y-axis represents the time elapsed from start of time-lapse imaging. Vertical lines correspond to cell cycle duration of the particular cell. Horizontal lines represent track splitting (cell division) events.

Figure S4 — Visual Overview of Tree Re-Assembly. Highlighted are the regions of ground truth trees (black thick background) which were correctly recapitulated by bTrack (gold branches). Upon branch breakage, two types of assembly actions were applied: subtree attachment (cyan), where the branch underwent further splitting, or branch attachment (pink), where the track did not further branch. In total, out testing tree pool comprised 8 perfectly tracked trees (ID: 13, 14, 15, 19, 22, 30, 38, 56), with 7 trees

Table S1. Residual U-Net performance on segmentation set of 869 cells.

Metric	Score	Equation
Localisation Error (MOTP) [pixels]	0.990	Eq S1
Intersection over Union (IoU)	0.802	Eq S2
Jaccard Index (J)	0.975	Eq S3
Pixel Identity (PI)	0.874	Eq S4

Table S2. Residual U-Net performance on the high-density field of view with strict intersection over union (IoU).

IoU Threshold	0.5	0.6	0.7	0.8	0.9
TP Object Count	574	547	487	335	133
Jaccard Index (J)	0.933	0.889	0.777	0.545	0.216
MOTP Error [pixels]	1.051	0.991	0.896	0.640	0.368

requiring one or more subtree stitch (ID: 2, 11, 27, 33, 52, 58 and 60). In case of tree 27, the right subtree was falsely associated with the tree (brown) and was swapped accordingly.

Figure S5 — Cell cycle heterogeneity does not occur due to the time of cell birth. Cycle lengths of fully resolved cells show no trend over the duration of live-cell imaging with respect to cell birth time relative to start time. For the first 60 hours of time-lapse imaging, the mean intermitotic time was determined to be constant. Due to variable durations of movie time-lapses, ranging between 64 and 120 hours, the mean cell cycling duration appears to decrease at longer imaging durations as only cells born with progressively shorter cycling lengths are sampled. Error bars indicate standard error of the mean calculated for 10-hour bins.

Figure S6 — Large-scale multigenerational analysis of single-cell cycling durations. (Top) Pearson (black circle points, dotted line) and Spearman (white square points, dashed line) rank correlations of cycle lengths between relatives from 5,032 unique lineage trees. Linear regression of each group is shown. Error bars indicate 95% confidence interval. Shaded golden blocks relate to the family relationships as in A, with counts of analysed cell replicates per kinship type listed. Horizontal black bars connect family relatives with equal generational distance to the nearest common ancestor with respect to reference cell, as indicated by numbers above the bars. (Bottom) Linear regression slopes for both coefficient types linearly decrease for at least 4 generational distances to nearest common ancestor.

3 SUPPLEMENTARY MOVIES

Movie S1 — Separation of automatically tracked single-cell lineages. Colours code for 'survivor' cells which can be fully tracked to the movie start through their lineage (cyan), 'incomer' cells which migrated into the field of view throughout the duration of the imaging (yellow) and 'mistracked' lineages where a tracking error, such as trajectory breakage or falsely identified mitosis, has occurred within their lineage (red). Scale bar = 50um. Link: `https://www.youtube.com/watch?v=ZxywQ7LaihI&t=20s`

Movie S2 — Single Cell Proliferation and Colony Expansion Heterogeneity. A sequence of colourised binary masks with segmented individual cells (grey) on background (black), highlighting cell proliferation throughout the duration of a representative 1193 frame-long movie (≈80 hours). Highlighted are the founder cells and progeny corresponding to slow (orange), medium (blue) and fast (green) dividers. Scale bar = 50um. Link: `https://www.youtube.com/watch?v=gScvX89JeYQ`

4 SUPPLEMENTARY TABLES

Table S3. Summary of the Cell Pairs Used for Calculation of Family Tree Cycling Correlations. C.I., confidence interval

Correlation Type		**Pearson Rank**			**Spearman Rank**		
Kinship Type	**Cell Pair Count**	**Correlation Coefficient**	**Lower 95% C.I. Bound**	**Upper 95% C.I. Bound**	**Correlation Coefficient**	**Lower 95% C.I. Bound**	**Upper 95% C.I. Bound**
mother	11696	0.42	0.42	0.45	0.51	0.5	0.52
sister	14380	0.65	0.64	0.66	0.71	0.71	0.72
grandmother	4667	0.24	0.21	0.26	0.25	0.22	0.28
aunt	10286	0.37	0.35	0.38	0.44	0.42	0.45
1st cousins	14144	0.51	0.5	0.52	0.56	0.55	0.57
1x great-grandmother	1002	0.1	0.04	0.16	0.07	0.01	0.13
grandaunt	4242	0.22	0.19	0.24	0.25	0.23	0.28
1st cousins 1-ce rem.	7349	0.32	0.3	0.34	0.37	0.35	0.39
2nd cousins	8524	0.41	0.39	0.43	0.41	0.39	0.42
2x great-grandmother	109	0.17	-0.02	0.34	0.14	-0.05	0.32
1x great-grandaunt	919	0.14	0.07	0.2	0.09	0.02	0.15
1st cousins 2-ce rem.	1647	0.16	0.11	0.21	0.19	0.14	0.24
2nd cousins 1-ce rem.	2695	0.25	0.22	0.29	0.23	0.2	0.27
3rd cousins	2484	0.39	0.35	0.42	0.25	0.22	0.29
3x great-grandmother	10	0.14	-0.54	0.71	0.45	-0.25	0.84
2x great-grandaunt	112	0.11	-0.07	0.29	0.09	-0.1	0.27
1st cousins 3-ce rem.	233	0.04	-0.09	0.17	-0.03	-0.16	0.09
2nd cousins 2-ce rem.	405	0.18	0.09	0.28	0.11	0.02	0.21
3rd cousins 1-ce rem.	596	0.38	0.31	0.45	0.28	0.21	0.36
4th cousins	441	0.62	0.56	0.68	0.32	0.23	0.4

REFERENCES

Bise, R., Yin, Z., and Kanade, T. (2011). Reliable cell tracking by global data association. *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* , 1004–1010doi:10.1109/ISBI.2011.5872571