# An objective comparison of cell-tracking algorithms

Vladimír Ulman[1,24,25], Martin Maška[1,25], Klas E G Magnusson[2], Olaf Ronneberger[3,24], Carsten Haubold[4], Nathalie Harder[5,24], Pavel Matula[1], Petr Matula[1], David Svoboda[1], Miroslav Radojevic[6], Ihor Smal[6], Karl Rohr[5], Joakim Jaldén[2], Helen M Blau[7], Oleh Dzyubachyk[8], Boudewijn Lelieveldt[8,9], Pengdong Xiao[10,24], Yuexiang Li[11,24], Siu-Yeung Cho[12], Alexandre C Dufour[13], Jean-Christophe Olivo-Marin[13], Constantino C Reyes-Aldasoro[14], Jose A Solis-Lemus[14], Robert Bensch[3], Thomas Brox[3], Johannes Stegmaier[15], Ralf Mikut[15], Steffen Wolf[4], Fred A Hamprecht[4], Tiago Esteves[16,17], Pedro Quelhas[16], Ömer Demirel[18], Lars Malmström[18], Florian Jug[19], Pavel Tomancak[19], Erik Meijering[6], Arrate Muñoz-Barrutia[20,21], Michal Kozubek[1] & Carlos Ortiz-de-Solorzano[22,23]



**We present a combined report on the results of three editions of the Cell Tracking Challenge, an ongoing initiative aimed at promoting the development and objective evaluation of cell segmentation and tracking algorithms. With 21 participating algorithms and a data repository consisting of 13 data sets from various microscopy modalities, the challenge displays today's state-of-the-art methodology in the field. We analyzed the challenge results using performance measures for segmentation and tracking that rank all participating methods. We also analyzed the performance of all of the algorithms in terms of biological measures and practical usability. Although some methods scored high in all technical aspects, none obtained fully correct solutions. We found that methods that either take prior information into account using learning strategies or analyze cells in a global spatiotemporal video context performed better than other methods under the segmentation and tracking scenarios included in the challenge.**

Cell migration and proliferation are two important processes in normal tissue development and disease[1], and optical microscopy remains the most appropriate imaging modality[2] for visualizing these processes. Imaging techniques, such as phase contrast (PhC) or differential interference contrast (DIC) microscopy, make cells visible without the need of exogenous markers. Fluorescence microscopy, on the other hand, relies on fluorescent reporters to specifically label cell components such as nuclei, cytoplasm or membranes. These labeled structures are then imaged in two or three dimensions by various imaging modalities, including widefield, confocal, multiphoton or light-sheet fluorescence microscopy.

To gain biological insights from time-lapse microscopy recordings of cell behavior, it is often necessary to identify individual cells and follow them over time. The bioimage-processing community has, since its inception, worked on extracting such quantitative information from microscopy images of cultured cells[3,4]. Recently, the advent of new imaging technologies has challenged this community with multi-dimensional, large image data sets following the development of tissues, organs or entire organisms. However, the tasks remain the same: accurately delineating (that is, segmenting) cell boundaries and tracking cell movements over time, providing information about their velocities and trajectories, and detecting cell-lineage changes as a result of cell division or cell death (**Fig. 1**). The level of difficulty of automatically

[1]Centre for Biomedical Image Analysis, Masaryk University, Brno, Czech Republic. [2]ACCESS Linnaeus Centre, KTH Royal Institute of Technology, Stockholm, Sweden. [3]Computer Science Department and BIOSS Centre for Biological Signaling Studies University of Freiburg, Frieburg, Germany. [4]Heidelberg Collaboratory for Image Processing, IWR, University of Heidelberg, Heidelberg, Germany. [5]Biomedical Computer Vision Group, Department of Bioinformatics and Functional Genomics, BIOQUANT, IPMB, University of Heidelberg and DKFZ, Heidelberg, Germany. [6]Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus University Medical Center Rotterdam, Rotterdam, the Netherlands. [7]Baxter Laboratory for Stem Cell Biology, Department of Microbiology and Immunology, and Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, California, USA. [8]Division of Image Processing, Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands. [9]Intelligent Systems Department, Delft University of Technology, Delft, the Netherlands. [10]Institute of Molecular and Cell Biology, A*Star, Singapore. [11]Department of Engineering, University of Nottingham, Nottingham, UK. [12]Faculty of Engineering, University of Nottingham, Ningbo, China. [13]BioImage Analysis Unit, Institut Pasteur, Paris, France. [14]Research Centre in Biomedical Engineering, School of Mathematics, Computer Science and Engineering, City University of London, London, UK. [15]Group for Automated Image and Data Analysis, Institute for Applied Computer Science, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany. [16]i3S - Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal. [17]Facultade de Engenharia, Universidade do Porto, Porto, Portugal. [18]S3IT, University of Zurich, Zurich, Switzerland. [19]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. [20]Bioengineering and Aerospace Engineering Department, Universidad Carlos III de Madrid, Getafe, Spain. [21]Instituto de Investigación Sanitaria Gregorio Marañon, Madrid, Spain. [22]CIBERONC, IDISNA and Program of Solid Tumors and Biomarkers, Center for Applied Medical Research, University of Navarra, Pamplona, Spain. [23]Bioengineering Department, TECNUN School of Engineering, University of Navarra, San Sebastián, Spain. [24]Present address: Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany (V.U.); DeepMind, London, UK (O.R.); Definiens AG, Munich, Germany (N.H.); National Heart Research Institute Singapore (NHRIS), National Heart Centre Singapore (NHCS), Singapore (P.X.); and Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China (Y.L.). [25]These authors contributed equally to this work. Correspondence should be addressed to C.O.-d.-S. (codesolorzano@unav.es).

**RECEIVED 3 JUNE; ACCEPTED 23 SEPTEMBER; PUBLISHED ONLINE 30 OCTOBER 2017; DOI:10.1038/NMETH.4473**

segmenting and tracking cells depends on the quality of the recorded video sequences (**Fig. 2** and Online Methods).

The image-processing community has addressed the above-mentioned tasks using increasingly sophisticated segmentation and tracking algorithms[5–7]. We briefly summarize the most commonly used methods for segmentation and tracking (**Fig. 3**).

For cell segmentation, creating a 'taxonomy of methods' is not a straightforward process, as state-of-the-art methods usually combine different strategies to achieve improved results. We classify existing algorithms by three criteria. First, the principle on which cells are detected, for example, by finding uniform areas, boundaries or at very low resolution by simply finding bright spots and maxima[8]. Second, the image features that are computed to achieve the cell segmentation. These can be simple pixel or voxel intensities, their local averages, or more complex local image descriptors of shapes or textures. Third, we distinguish the segmentation method itself that implements the principle using the features. The methods range from simple methods like thresholding[9,10], hysteresis thresholding[11], edge detection[12] and shape matching[13,14] to more sophisticated approaches like region growing[15–17], machine learning[18,19] and energy minimization[20–26].

Cell-tracking methods can be broadly categorized into two groups. Tracking by contour evolution methods[21,22,24,25] start by segmenting the cells in the first frame of a video and then evolve their contours in consecutive frames, thereby solving the segmentation and tracking tasks simultaneously, one step at a time, under the essential assumption of unambiguous, spatiotemporal overlap between the corresponding cell regions. Tracking by detection methods[14,19,26–29], in contrast, start by segmenting the cells in all frames of a video and later, using mostly probabilistic frameworks, establish temporal associations between the segmented cells. This can be done by either using a two-frame or multiframe sliding window, or even for all frames at once.

The diversity of imaging modalities, cell-tracking tasks and available algorithms makes it difficult for biologists to decide which algorithm to use under certain conditions. Moreover, the developers of image-processing algorithms need to objectively evaluate new cell segmentation and tracking solutions by comparing their performance on standardized data sets. We addressed these problems by organizing three Cell Tracking Challenges (CTC I–III) between 2013 and 2015. For these challenges, we created a diverse repository of annotated microscopy videos and defined quantitative evaluation measures to allow a fair comparison of the competing algorithms[30]. The participating algorithms were examined under the challenge conditions. Here we present an in-depth analysis of the CTC results, provide useful guidelines for users to identify appropriate algorithms for their own data sets and point developers to open challenges that we believe are insufficiently addressed by the algorithms tested. It is important to note that the CTC is an open-source initiative that remains open online, and most of the competing methods are publicly available through the challenge website (http://celltrackingchallenge.net/).

## RESULTS

### Data sets and ground truth

The data set repository (**Fig. 4**, **Supplementary Table 1** and **Supplementary Videos 1–13**) consists of 52 annotated videos from 13 classes, occupying 92 GB of raw image data. Of the 13 data sets, 11 consist of contrast enhancing (PhC, DIC) or fluorescence (wide-field, confocal, light sheet) microscopy recordings of live cells and organisms in two (2D) or three dimensions (3D). The other two data sets are synthetic, generated using a cell simulator that produces

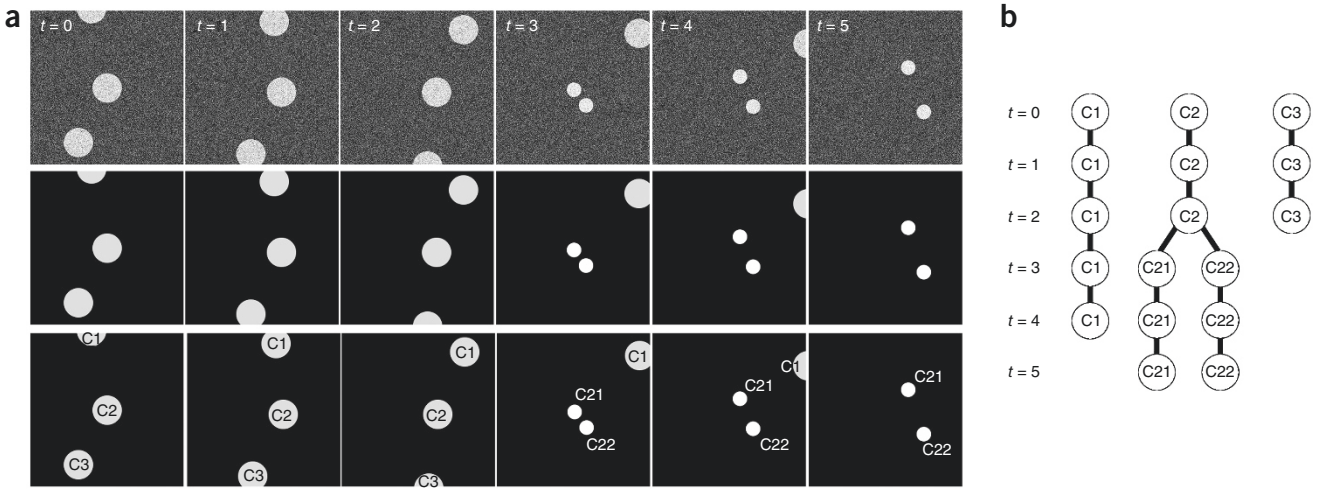


**Figure 1** | Concept of cell segmentation and tracking. (**a**) Top, artificial sequence that simulates six consecutive frames of a time-lapse video. The gray circles represent cells moving on a flat surface. Middle, the goal of a segmentation algorithm is to accurately determine the regions of each individual cell in every frame, constructing a set of binary segmentation masks that correspond to the cells and locate them on a flat background. Bottom, a tracking algorithm finds correspondences between the masks, i.e., the cells, in consecutive frames. If properly designed, a tracking algorithm is able to detect a moving cell (e.g., C1 or C3) while it is in the field of view, determining when the cell enters and leaves the field of view. From the location of the cells in consecutive frames, it is possible to determine the trajectory of each cell and its velocity. A tracking algorithm should also be able to detect lineage changes as a result of, for instance, a cell division event (for example, cell C2 divides into two daughter cells, C21 and C22) or apoptosis. (**b**) Graph-based representation of the cell tracks found by a tracking algorithm in the sequence shown at the top of **a**. Such an acyclic-oriented graph contains, for each cell, the time when the cell enters and leaves the field of view, along with its division or apoptotic events. In a real case scenario, these graphs show the complete genealogy of the cells displayed in the frame of the video, for the entire length of the video. Please note that the orientation of the graph edges follows the temporal sequence starting at $t = 0$ and moving toward $t = 5$.

realistic 2D and 3D renderings of chromatin-stained live cells[31]. **Supplementary Note 1** and supporting **Supplementary Figures 1–11** provide a detailed description of the data sets. **Supplementary Note 2** and supporting **Supplementary Figure 12** describe the simulator used to create the synthetic data sets, applying the parameter configuration provided in **Supplementary Data 1**. **Table 1** provides a quantitative characterization of the quality of each data set, based on the measures described in the Online Methods. In all of the tables, figures and videos, we use a naming convention for data sets that identifies their microscopy modality (fluorescence (Fluo), DIC,

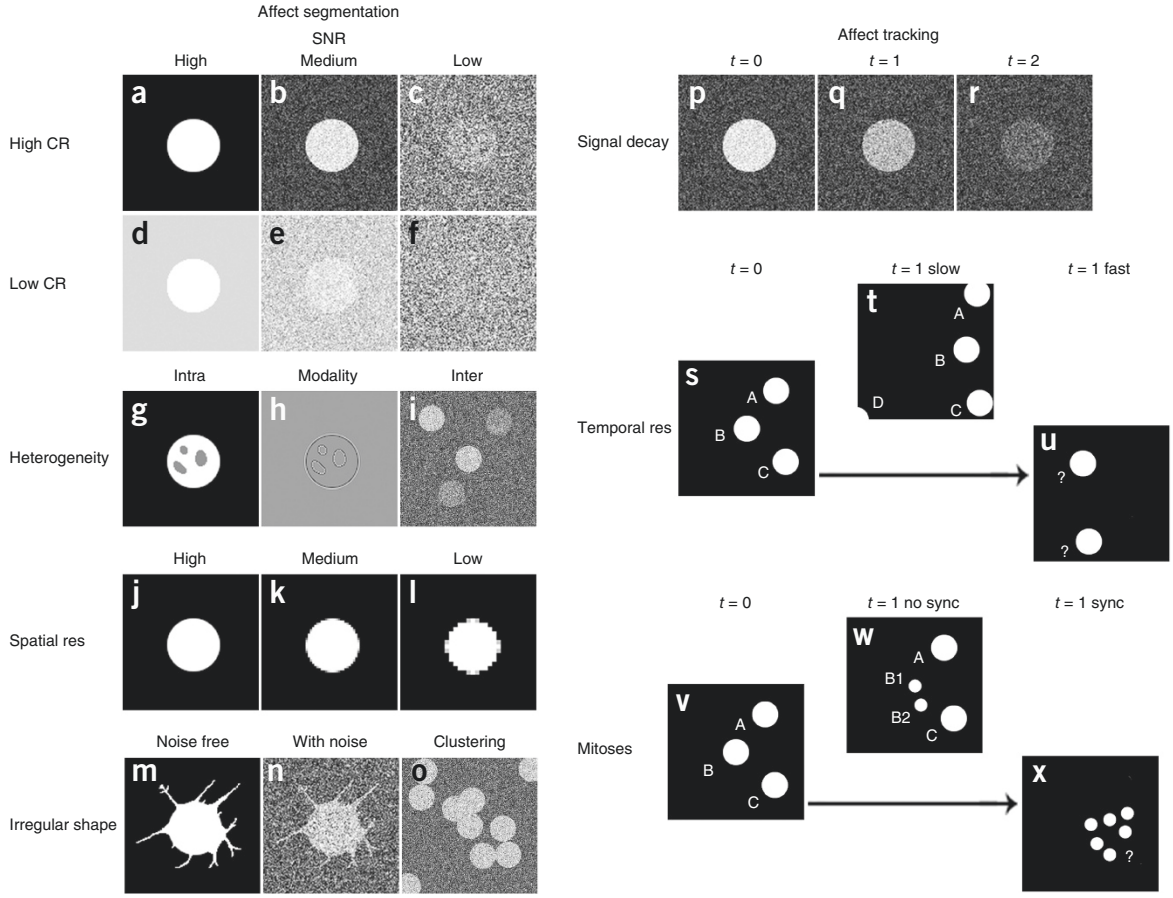


**Figure 2** | Concept of the main factors that determine the quality of cell images and videos. (**a**–**f**). SNR and CR measure the relationship between the signal captured from the cells and the unwanted noise or signal captured at the same time. Decreasing SNR is shown using a cell with 250 intensity units (iu) and no background (0 iu) in three scenarios of increasing s.d. (in iu) of background Gaussian noise: 0 (**a**), 50 (**b**) and 200 (**c**). The effect of decreased CR is displayed using a simulated cell in high background (200 iu) with increasing noise s.d.: 0 (**d**), 50 (**e**) and 200 (**f**). The effect is shown for three increasing noise levels: 0 noise (**a** versus **d**), 50 noise s.d. (**b** versus **e**) and 200 noise s.d. (**c** versus **f**). (**g**,**h**) Intra-cellular signal heterogeneity that can lead to cell over-segmentation when the same cell yields several detections is simulated by a cell with nonuniform distribution of the labeling marker or nonlabel retaining structures (**g**). Signal texture can also be linked to the process of image formation, in this case shown using a simulated cell image imaged by PhC microscopy (**h**). (**i**) Signal heterogeneity between cells, shown by simulated cells with different average intensities can be a result of, for instance, different levels of protein transfection, non-uniform label uptake, or cell cycle stage or chromatin condensation, when using chromatin-labeling techniques. (**j**–**l**) Spatial resolution that can compromise the accurate detection of cell boundaries is displayed using a cell captured with increasing pixel size, i.e., with decreasing spatial resolution: full resolution (**j**), half resolution (**k**) and one fourth of the original full resolution (**l**). (**m**,**n**) Irregular shape that can cause over/under-segmentation, especially when the segmentation methods assume simpler, non-touching objects, is displayed using a simulated cell with highly irregular shape under two background noise s.d. situations: 0 (**m**) and 100 (**n**).This is especially a problem in high-noise situations (**n**). (**o**) High density of cells, which is also a frequent cause of incorrect segmentation, is shown by a cluster of simulated cells. (**p**–**r**) Fluorescence temporal decay that can bring the SNR or CR below detection levels, thereby complicating both segmentation and tracking, is simulated by a cell in a time series showing increasing fluorescence decay as a result of bleaching or quenching of the fluorochrome, and same noise conditions (s.d. of 50 iu): original cell at the beginning of the experiment (**p**), cell with 100 iu decay (**q**) and cell with 200 iu decay (**r**). (**s**–**u**) Cell overlap between consecutive frames is important for correctly tracking the cells, as many algorithms rely on this overlap. Here it is shown using three simulated cells at the beginning of a video ($t = 0$) (**s**) and two possible alternative scenarios for the following time point ($t = 1$): $t = 1$ in a scenario of high temporal resolution and/or low cell speed, allowing relatively simple identification of the correspondence between the cells (**t**); $t = 1$ in a scenario of low temporal resolution and/or high cell speed, complicating the identification of the correspondence between the cells (**u**). (**v**–**x**) Number and synchronization of mitotic events also complicates cell tracking, as tracking a mitotic cell requires correctly assigning the mother to its daughter cells in consecutive frames. This is simulated by cells at the beginning of the video ($t = 0$) (**v**) and two possible alternative scenarios for the following time point ($t = 1$): $t = 1$ in a scenario where only one of the cell divides asynchronously allowing a simple lineage assignment of mother and daughter cells (**w**); $t = 1$ in a scenario of multiple, synchronized division events rendering a complicated lineage assignment of mothers and daughters (**x**).

PhC), the staining (nuclear (N), cellular (C)), the dimensionality (2D, 3D), the resolution (low (L), high (H)), and the cell type or model organism used.

Each data set consists of two training and two competition videos. The training videos, along with their reference annotations, were provided at the time of registration for the CTC, allowing the participants to carry out performance-driven optimization of their algorithms. The competition videos, excluding the reference annotations that were kept secret, were provided at a later time, which allowed the participants to visually fine tune their algorithms on the competition videos before submitting their results.

Three independent human experts created a segmentation solution and a tracking solution (annotation) for each nonsynthetic video[30]. The final segmentation (SEG-GTs) and tracking (TRA-GTs) ground truths were created by combining the three annotations, following a majority-voting scheme[30]. SEG-GTs for the data sets of *Caenorhabditis elegans* (Fluo-N3DH-CE) and the *Drosophila melanogaster* (Fluo-N3DL-DRO) embryos were generated as described above, but in the case of Fluo-N3DL-DRO, only cells of the early nervous system were annotated and used as ground truth. TRA-GTs of both embryonic data sets were not created following the description above. Instead, they were created using published protocols[32,33] by the groups that provided the data sets. For the synthetic videos, SEG-GTs and TRA-GTs were inherently created by the cell simulator used[31].

### Participants, algorithms and handling of submissions

17 teams from 11 countries participated in the three CTC editions, all providing complete tracking results for at least one of the data sets. Two teams submitted more than one algorithm, leading to a total of 21 competing algorithms. **Tables 2** and **3** list the algorithms and classify their segmentation and tracking strategies. **Supplementary Table 2** lists affiliations of the participating teams, and **Supplementary Table 3** contains links to the executable versions of most of the submitted algorithms. Their expanded description is presented in the **Supplementary Note 3**, and the parameter configurations used by each algorithm are listed in the **Supplementary Data 2**. All submissions were received by the CTC organizers as labeled segmentation masks and structured text files containing the cell-lineage graphs. The CTC organizers verified the submitted results by reproducing them on a single computer, using the executable version of each algorithm provided by the participants.

### Quantitative performance criteria

To quantify the performance of all submitted algorithms, we developed three categories of measures that quantified the segmentation and tracking accuracy from the computer science point of view, the biological relevance of the obtained tracking results, and the practical usability of the methods (see Online Methods). It is important to note that only the first set of measures was evaluated in the challenge, and the methods were therefore only fine tuned in this respect. The other two sets were used to analyze aspects that are relevant from the user point of view. **Supplementary Table 3** contains a link to the evaluation software used in the challenge.

The first set of measures examined the segmentation and tracking accuracy of the methods from the developer's point of view. The segmentation accuracy measure (SEG) evaluates the average amount of overlap between the reference segmentation ground truth (SEG-GT) and the segmentation masks computed by an evaluated algorithm. The tracking accuracy measure (TRA) is a normalized weighted distance between the tracking solution submitted by the participant and the reference tracking ground truth (TRA-GT), with weights chosen to reflect the effort it takes a human curator to carry out the edits manually. Both SEG and TRA take values in the interval [0, 1], with higher values corresponding to better performance. For ranking the algorithms, the overall performance (OP) is computed by averaging SEG and TRA values for each pair of competition videos, and then averaging these averages (i.e., $OP = 0.5 \cdot (SEG_{avg} + TRA_{avg})$). In summary, SEG and TRA evaluate results in terms of similarity to the ground truth and are particularly relevant for comparing algorithms with one another. Method developers use such measures to show the superiority of new methods over current state-of-the-art methods.

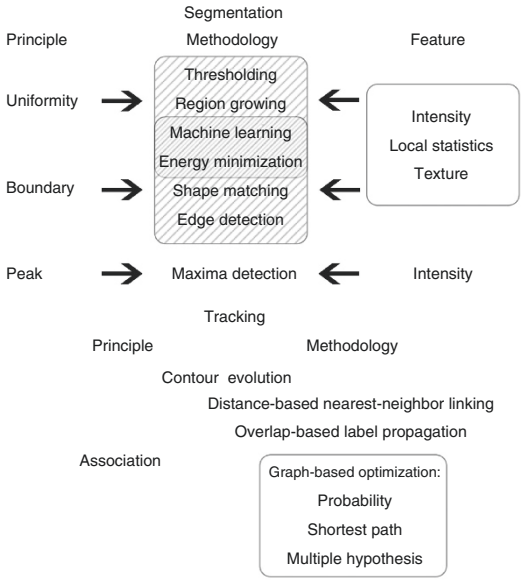


**Figure 3** | Taxonomy of cell segmentation and tracking methods.

Biologists, however, have specific questions when using tracking algorithms and are therefore usually more interested in specific aspects of the final segmentation and tracking analysis. For this reason, we evaluated four additional aspects of biological relevance. Complete tracks (CT) measures the fraction of ground truth cell tracks that a given method is able to reconstruct in their entirety, from the frame they appear in to the frame they disappear from. CT is especially relevant when a perfect reconstruction of the cell lineages is required. Track fractions (TF) averages, for all detected tracks, the fraction of the longest continuously matching algorithm-generated tracklet with respect to the reference track. Intuitively, this can be interpreted as the fraction of an average cell's trajectory that an algorithm reconstructs correctly once the cell has been detected. Branching correctness (BC) measures how efficient a method is at detecting division events. Finally, the cell cycle accuracy (CCA) measures how accurate an algorithm is at correctly reconstructing the length of cell cycles (that is, the time between two consecutive divisions). Both BC and CCA are informative about the ability of the algorithm to detect cell population growth. All of the biologically inspired measures take

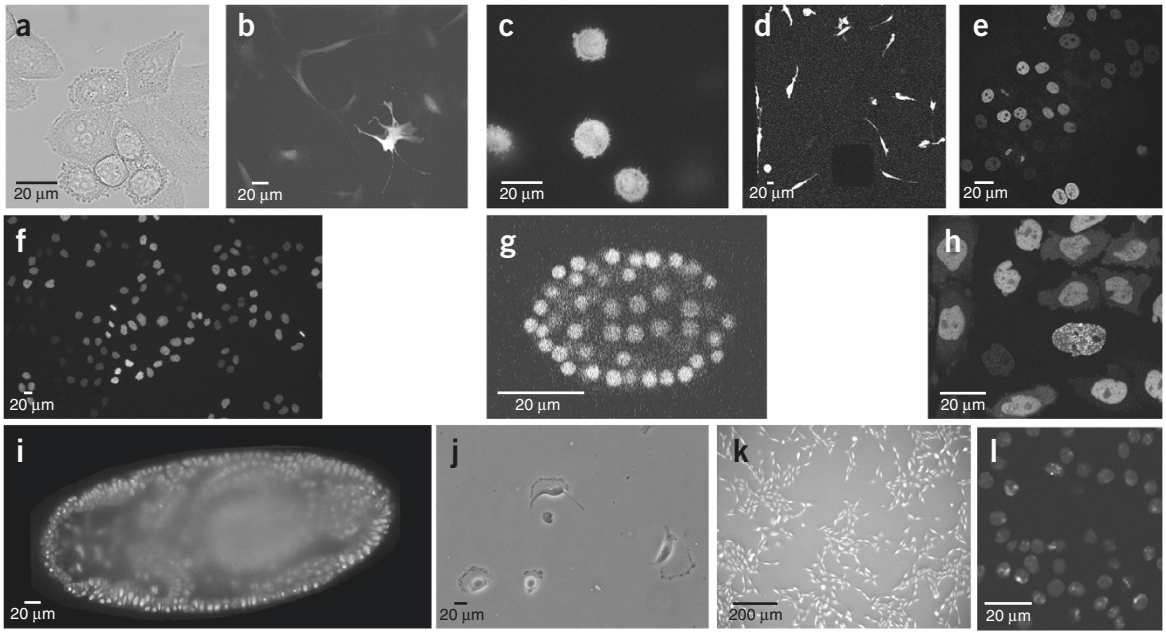


**Figure 4** | Sample images of the challenge data sets. (**a**) DIC-C2DH-HeLa. (**b**) Fluo-C2DL-MSC. (**c**) Fluo-C3DH-H157. (**d**) Fluo-C3DL-MDA231. (**e**) Fluo-N2DH-GOWT1. (**f**) Fluo-N2DL-HeLa. (**g**) Fluo-N3DH-CE. (**h**) Fluo-N3DH-CHO. (**i**) Fluo-N3DL-DRO. (**j**) PhC-C2DH-U373. (**k**) PhC-C2DL-PSC. (**l**) Fluo-N2DH-SIM+ and Fluo-N3DH-SIM+.

values in the interval [0,1], with higher values corresponding to better performance.

The third set of measurable quantities expresses the practical usability of the submitted algorithms. The first indication of an algorithm's usability is the number of tunable parameters (NP) a user is required to manually set, excluding parameters visible only to developers. In general, a lower number of tunable parameters indicates a more usable algorithm. A very different, but important, attribute of an algorithm is its generalizability (GP). This measure quantifies how stable an algorithm is when being applied with the same parameter configuration to new videos acquired under otherwise unchanged imaging conditions. GP values are computed by comparing the results for a particular training and competition video obtained using the same parameter configuration. This measure takes values in the interval [0,1], with higher values corresponding to better generalizability.

The last value we report for each algorithm is its execution time (TIM), in seconds.

### Analysis of the performance of submitted algorithms

All of the measures described have been computed for every data set and competing algorithm. We first evaluated the SEG and TRA measures (**Figs. 5** and **6** and **Supplementary Data 3**). To determine the significance of these values, we calculated SEG and TRA values with respect to the ground truth for the three manual annotations, as they are the best available proxies for evaluating the variability among human annotators. Thus, algorithms with SEG or TRA scores in the range of the average manual scores ($SEG_a$ and $TRA_a$), ±1 s.d., can be considered to perform at the level of human annotators, and algorithms with scores above or below that range can be said to perform better or worse, respectively, than the human annotators.

**Table 1** | Properties of the competition data sets used in the three editions of the Cell Tracking Challenge

| Name | SNR | CR | $Het_i$ | $Het_b$ | Res | Sha | Den | Cha | Ove | Mit | Syn | Ent/Leav | Apo | Deb |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DIC-C2DH-HeLa | 0.74 | 1.00 | 27.28* | 1.35* | 12,032 | 0.68 | 9.8 | 0.43 | 0.91 | 0.02 | N | Y | Y | Y |
| Fluo-C2DL-MSC | 2.81 | 1.50 | 1.19 | 0.74 | 11,787 | 0.32 | 32.8 | 104.78* | 0.72 | 0.01 | N | Y | N | N |
| Fluo-C3DH-H157 | 31.53 | 3.14 | 0.35 | 0.42 | 349,593* | 0.60 | 46.6 | 11.52 | 0.86 | 0.00 | N | Y | N | N |
| Fluo-C3DL-MDA231 | 9.36 | 4.24 | 1.26 | 0.20 | 1,696 | 0.60 | 18.5 | 8.86 | 0.71 | 0.17 | N | Y | N | N |
| Fluo-N2DH-GOWT1 | 6.16 | 11.31 | 0.83 | 0.81 | 3,327 | 0.80 | 40.6 | 0.01 | 0.92 | 0.07 | N | Y | N | Y |
| Fluo-N2DL-HeLa | 57.72 | 1.02 | 0.28 | 0.62 | 561 | 0.80 | 15.8 | 2.58 | 0.88 | 1.45 | N | Y | Y | Y |
| Fluo-N3DH-CE | 6.74 | 3.46 | 0.66 | 0.27 | 6,001 | 0.69 | 4.8 | 0.19 | 0.75 | 1.86 | Y | N | N | N |
| Fluo-N3DH-CHO | 25.96 | 10.43 | 0.59 | 0.27 | 14,494 | 0.58 | 33.7 | 0.01 | 0.87 | 0.06 | N | Y | Y | N |
| Fluo-N3DL-DRO | 2.46 | 3.32 | 0.31 | 0.18 | 1,188 | 0.65 | 12.3 | 0.98 | 0.68 | 1.05 | N | N | N | N |
| PhC-C2DH-U373 | 2.88 | 1.10 | 19.30* | 0.87 | 4,287 | 0.58 | 48.8 | 0.04 | 0.91 | 0.00 | N | Y | N | Y |
| PhC-C2DL-PSC | 4.06 | 1.53 | 0.52 | 0.34 | 114 | 0.60 | 8.5 | 0.04 | 0.90 | 1.99 | N | Y | N | Y |
| Fluo-N2DH-SIM+ | 6.30 | 1.23 | 0.95 | 0.48 | 1,181 | 0.72 | 18.2 | 0.14 | 0.89 | 0.49 | N | Y | N | N |
| Fluo-N3DH-SIM+ | 5.22 | 1.24 | 1.14 | 0.41 | 38,285 | 0.73 | 16.2 | 0.14 | 0.86 | 0.49 | N | Y | N | N |

The displayed values correspond to the image/video quality parameters mathematically described in the Online Methods. SNR, signal-to-noise ratio; CR, contrast ratio; $Het_i$, internal signal heterogeneity of the cells; $Het_b$, heterogeneity of the signal between cells; Res, resolution, measured as the size of the cells in number of pixels (2D) or voxels (3D); Sha, regularity of the cell shape, normalized between 0 (completely irregular) and 1 (perfectly regular); Den, cell density measured as minimum pixel (2D) or voxel (3D) distance between cells; Cha, change of the average intensity of the cells with time; Ove, level of overlap of the cells in consecutive frames, normalized between 0 (no overlap) and 1 (complete overlap); Mit/Syn, number and synchronization of division events; Ent/Leav, cells entering or leaving the field of view; Apo, presence of apoptotic cells; Deb, presence of moving debris. Color code: for each category and data set, the average was computed excluding outlying values (*). The background color of the cell indicates whether the highlighted value is in the categories' average ±1/2 s.d. (yellow) or the value is outside of that range (green or red). A red background indicates a poor value in a given category, and a green background indicates a high value for a given category. In Sha, the 2D and 3D data sets were treated separately because different shape descriptor was used for 2D and for 3D cases.

**Table 2** | Segmentation strategies used by the competing methods

| Algorithm | Preprocessing | Principle | Feature | Methodology | Postprocessing |
|---|---|---|---|---|---|
| COM-US | Noise suppression Intensity normalization | Homogeneity | Intensity | Thresholding | Size filtering |
| CUL-UK | Noise suppression Illumination correction | Homogeneity | Intensity | Thresholding | Size filtering |
| CUNI-CZ | Noise suppression | Homogeneity | Intensity | Thresholding | Size filtering Cluster separation |
| FR-Be-GE | Intensity normalization Illumination correction | Homogeneity Boundary | Intensity | Energy minimization | Size filtering Hole filling |
| FR-Ro-GE | Intensity normalization Illumination correction | Homogeneity | Texture descriptor | Machine learning | None |
| HD-Har-GE | Noise suppression Intensity clipping | Homogeneity | Intensity | Thresholding | Hole filling Cluster separation |
| HD-Hau-GE | None | Homogeneity | Texture descriptor | Machine learning | Size filtering |
| IMCB-SG (1) | Noise suppression Illumination correction | Homogeneity | Intensity | Thresholding | Size filtering Cluster separation |
| IMCB-SG (2) | Image resampling Noise suppression Illumination correction | Homogeneity | Intensity | Thresholding | Size filtering Cluster separation |
| KIT-GE | Noise suppression | Homogeneity | Local descriptor | Thresholding | None |
| KTH-SE (1) | Intensity normalization Noise suppression Illumination correction | Homogeneity | Intensity | Thresholding | Size filtering Hole filling Cluster separation |
| KTH-SE (2) | Intensity normalization Noise suppression Illumination correction | Homogeneity | Intensity | Thresholding | Size filtering Hole filling Cluster separation |
| KTH-SE (3) | Intensity normalization Illumination correction | Homogeneity | Local descriptor | Thresholding | Boundary refinement |
| KTH-SE (4) | Intensity normalization Noise suppression | Boundary | Intensity | Thresholding | Size filtering Region merging |
| LEID-NL | Noise suppression | Homogeneity | Intensity | Energy minimization | Cluster separation |
| MU-CZ | Noise suppression | Homogeneity | Intensity | Energy minimization | Cluster separation |
| NOTT-UK | Intensity normalization | Homogeneity | Intensity | Thresholding | None |
| PAST-FR | Intensity normalization Noise suppression | Homogeneity Boundary | Intensity | Energy minimization | None |
| UP-PT | Image subsampling Noise suppression | Homogeneity Peak | Intensity | Thresholding | Boundary refinement |
| UPM-ES | Noise suppression | Homogeneity | Intensity | Thresholding | Size filtering Hole filling Boundary refinement |
| UZH-CH | Intensity normalization Noise suppression Illumination correction | Homogeneity | Intensity | Region growing | Size filtering Hole filling |

Principle, feature and methodology used in the segmentation phase of the competing algorithms (following the taxonomy shown in **Fig. 3**) along with the preprocessing and postprocessing strategies employed.



We first examine the results trying to pinpoint the features that underlie the good and not so good performance of the competing methods (**Fig. 5**). We observed that some algorithms reached very good values (OP > 0.9) for data sets Fluo-N2DH-GOWT1, PhC-C2DH-U373, Fluo-N2DL-HeLa, Fluo-C3DH-H157 and Fluo-N3DH-CHO. In all but one of these data sets (Fluo-C3DH-H157), one or more algorithms reached human-quality results. Notably, all but one of these results were obtained on fluorescence data with high signal-to-noise ratio (SNR) or contrast ratio (CR) values. Some also showed high spatial (Fluo-C3DH-H157, Fluo-N3DH-CHO) and/or temporal (Fluo-N2DH-GOWT1, Fluo-N2DL-HeLa, Fluo-N3DH-CHO) resolution and displayed rather low cell densities (Fluo-C3DH-H157, Fluo-N2DH-GOWT1, PhC-C2DH-U373, Fluo-N3DH-CHO).

A second group of data sets was solvable with OP values between 0.75 and 0.9 (DIC-C2DH-HeLa, PhC-C2DL-PSC, Fluo-C3DL-MDA231, Fluo-N2DH-SIM+ and Fluo-N3DH-SIM+). For these data sets, the SEG and TRA values are near, but below, the performance of the human annotators, meaning that after automatic tracking some additional curation work is required to reach the level of the human-level solutions. The difficulty for DIC-C2DH-HeLa and PhC-C2DL-PSC appeared to be the low SNR and CR values and high cell density, and for DIC-C2DH-HeLa also the rather complex image texture of the cells (**Supplementary Figs. 1** and **11**). For Fluo-C3DL-MDA231, the low SNR and CR values were paired with low spatial and temporal resolution and substantial photobleaching (**Supplementary Fig. 4**). The two synthetic data sets (Fluo-N2DH-SIM+, Fluo-N3DH-SIM+) showed average SNR, low CR, average cell density and average-to-high heterogeneity in and between cells.

Three data sets (Fluo-C2DL-MSC, Fluo-N3DH-CE and Fluo-N3DL-DRO) turned out to be the hardest to segment and track fully automatically (OP < 0.75). For these data sets, a substantial amount of manual work would be needed to curate the computed results to reach human-level annotations. Fluo-C2DL-MSC suffered mostly from low SNR and CR values, low temporal resolution

**Table 3** | Tracking strategies used by the competing methods

| Method | Principle | Methodology | Temporal support | Postprocessing | Division detection |
|---|---|---|---|---|---|
| COM-US | Association | Graph-based multiple hypothesis tracking | All | Distance-based track refinement | None |
| CUL-UK | Association | Motion prediction-based label propagation | 3 | Cell-collision-based track refinement | None |
| CUNI-CZ | Association | Distance-based nearest neighbor linking | 2 | None | Specific |
| FR-Be-GE | Association | Maximum-overlap-based label propagation | 2 | None | None |
| FR-Ro-GE | Association | Maximum-overlap-based label propagation | 2 | None | None |
| HD-Har-GE | Association | Constrained distance-based nearest neighbor linking | 3 | Location- and length-based track refinement | Specific |
| HD-Hau-GE | Association | Probability-graph-based global optimization | All | None | Inherent |
| IMCB-SG (1) | Association | Overlap-based label propagation | 2 | None | Inherent |
| IMCB-SG (2) | Association | Distance-based nearest neighbor linking | 2 | None | Specific |
| KIT-GE | Association | Distance-based nearest neighbor linking | 2 | None | Specific |
| KTH-SEM (1) | Association | Graph-based shortest path global optimization | All | Adjacency- and overlap-based track refinement | Inherent |
| KTH-SEM (2) | Association | Graph-based shortest-path global optimization with detection preprocessing | All | Adjacency based track refinement | Inherent |
| KTH-SEM (3) | Association | Graph-based shortest-path global optimization | All | Adjacency based track refinement | Inherent |
| KTH-SEM (4) | Association | Graph-based shortest-path global optimization | All | Adjacency based track refinement | Inherent |
| LEID-NL | Contour evolution with motion compensation | | 2 | None | Specific |
| MU-CZ | Contour evolution with bleaching compensation | | 2 | Location-based track refinement | Inherent |
| NOTT-UK | Association | Distance-based nearest neighbor linking | 2 | None | Inherent |
| PAST-FR | Contour evolution | | 2 | None | Inherent |
| UP-PT | Association | Distance-based nearest neighbor linking | 2 | Location- and length-based track refinement | Specific |
| UPM-ES | Association | Overlap-based label propagation | 2 | None | None |
| UZH-CH | Association | Distance-based nearest neighbor linking | 2 | None | Specific |

Principle and methodology used in the tracking phase of all the competing algorithms (following the taxonomy shown in **Fig. 3**) along with postprocessing strategies employed, the temporal support given, and the scheme followed for the division detection.



and substantial photobleaching. This data set was also difficult to segment correctly as a result of its prominent cell protrusions (**Supplementary Fig. 2**). For Fluo-N3DH-CE and Fluo-N3DL-DRO, the two whole-embryo data sets, the algorithms mostly struggled to segment and track the very noisy cell nuclei in 3D. In addition, these data sets showed very low spatial resolution, relatively low temporal resolution and increasingly dense frames toward the end of the videos, which strongly complicated tracking of the segmented cells (**Supplementary Figs. 7** and **9**).

Next, we examined the results from the viewpoint of the algorithms, asking which ones showed the best overall performance (**Fig. 6**). The algorithms KTH-SE, FR-Ro-GE and HD-Hau-GE ranked first for one or more data sets. Looking more globally at the number of top-three occurrences, KTH-SE, FR-Ro-GE and HD-Har-GE outperformed the others. Their common denominator was reliance on the tracking by detection paradigm. In particular, KTH-SE algorithms performed extraordinarily well, and they were ranked among the top-three algorithms for all data sets. These methods rely on a simple thresholding for segmentation, the results of which are highly enriched by the use of global information in the tracking process. In some data sets, however, the tracking by contour evolution methods (LEID-NL, MU-CZ and PAST-FR) reached the level of the tracking by detection methods. This can be attributed to their high segmentation performance on data sets with high temporal and spatial resolution (Fluo-N3DH-CHO, Fluo-N2DH-GOWT1, Fluo-N2DH-SIM+ and Fluo-N3DH-SIM+). These results highlight how these methods rely on substantial cell-to-cell overlaps between successive frames to work properly. Finally, it is interesting to note the exceptional performance of the machine-learning methods (FR-Ro-GE, HD-Hau-GE) on contrast enhancement microscopy (PhC and DIC) data sets. Indeed, these methods obtained performance values on DIC-C2DH-HeLa, PhC-C2DH-U373 and PhC-C2DL-PSC that did not match their predicted level of complexity. This can be explained by the fact that the internal texture of the cells in these data sets is not detrimental for the segmentation. On the contrary, it seems to improve the learning capacity of the algorithms.

Notably, the evolution of the average of the top-three OP values during the three CTC editions showed progress toward the objective of reaching the level of the human expert annotators (**Supplementary Fig. 13**). Across all data sets, the average top-three OP values rose by $0.03 \pm 0.03$ (CTC II versus CTC I) and $0.05 \pm 0.07$ (CTC III versus CTC I).

We studied the robustness of the OP-based rankings (see Online Methods and **Supplementary Fig. 14**) and found that the rankings were indeed robust for up to 45% of possible weight changes. Furthermore, we analyzed the correlation (i.e., interdependence) of SEG and TRA scores using the Kendall's $\tau$ correlation coefficient (**Supplementary Table 4**) and found moderate global correlation (0.55) with only a few cases of very high (DIC-C2DH-HeLa and Fluo-N3DH-CE) or high (PhC-C2DL-PSC and Fluo-C2DL-MSC) correlation.

Given that segmentation and tracking are meant to answer biological questions in the hands of practicing biologists, we next

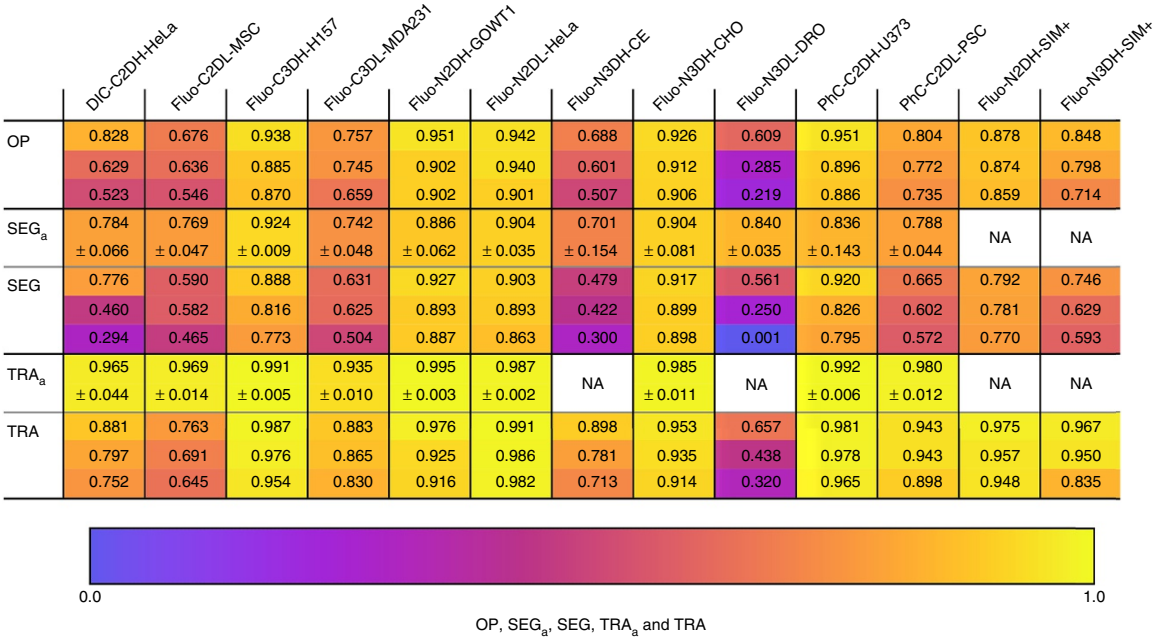


**Figure 5** | Top-three technical performance values (SEG, TRA and OP) obtained by the competing algorithms. Both the SEG and TRA sections start with $SEG_a$ and $TRA_a$, respectively, which are the average plus s.d. values of the measures obtained by three manual annotations used to create the ground truths (SEG-GTs and TRA-GTs), which were considered as if they were also regular submissions. The color code below correlates with the values in the [0, 1] interval for the SEG, TRA and OP scores. NA, not applicable because only one tracking annotation exists (Fluo-N3DH-CE and Fluo-N3DL-DRO) or because no manual annotation was necessary as a result of the existence of an absolute ground truth (simulated data sets Fluo-N2DH-SIM+ and Fluo-N3DH-SIM+).

analyze the biologically inspired and usability measures. **Figure 7** shows the top-three biological scores: CT, TF, BC and CCA, and the average values obtained by the annotators ($CT_a$, $TF_a$, $BC_a$ and $CCA_a$). When looking at CT across data sets, we observed very low values overall, but especially so for DIC-C2DH-HeLa, Fluo-C2DL-MSC, PhC-C2DL-PSC and the two embryonic developmental data sets (Fluo-N3DH-CE and Fluo-N3DL-DRO). The low CT values are especially relevant for the embryonic data sets, as tracking completeness is critical for a correct genealogical reconstruction of embryo development. The TF values were at a higher level, meaning that the methods are reasonably competent at measuring cell speeds and trajectories, but some work is still required to bring them to the level of the human annotators. Finally, Fluo-N2DL-HeLa, Fluo-N2DH-SIM+ and Fluo-N3DH-SIM+ showed high BC and CCA values, which indicates that the methods are able to correctly detect cell divisions and cell-population growth;

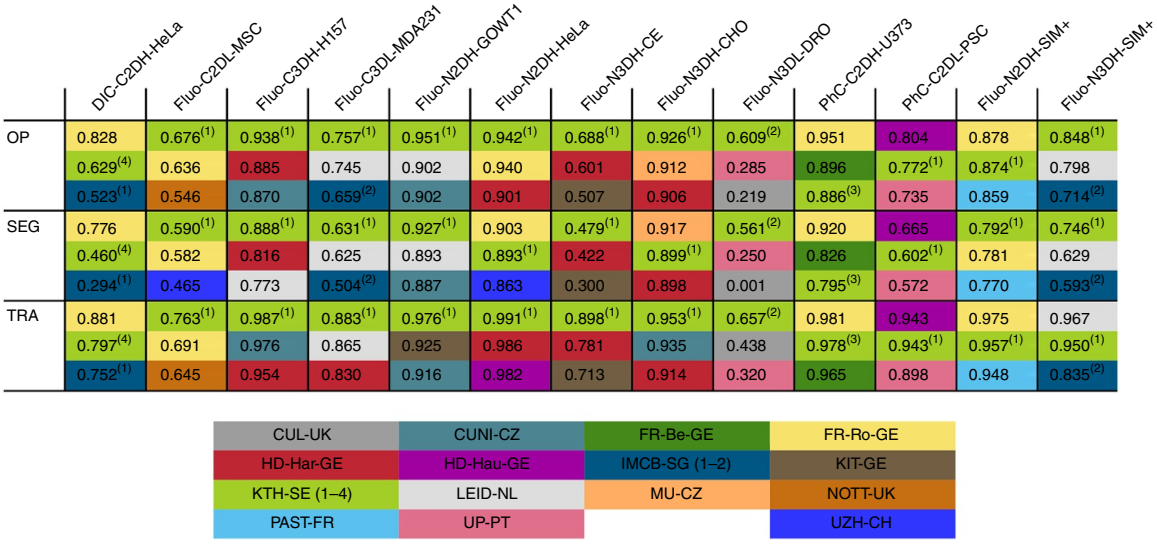


**Figure 6** | Top-three performing methods of the three challenge editions. For each data set, the table shows the OP and its corresponding average SEG and TRA scores computed over the two competition videos. Note that the methods submitted by the same participant are displayed in the same color, with super-indices denoting the particular method of the respective participant.

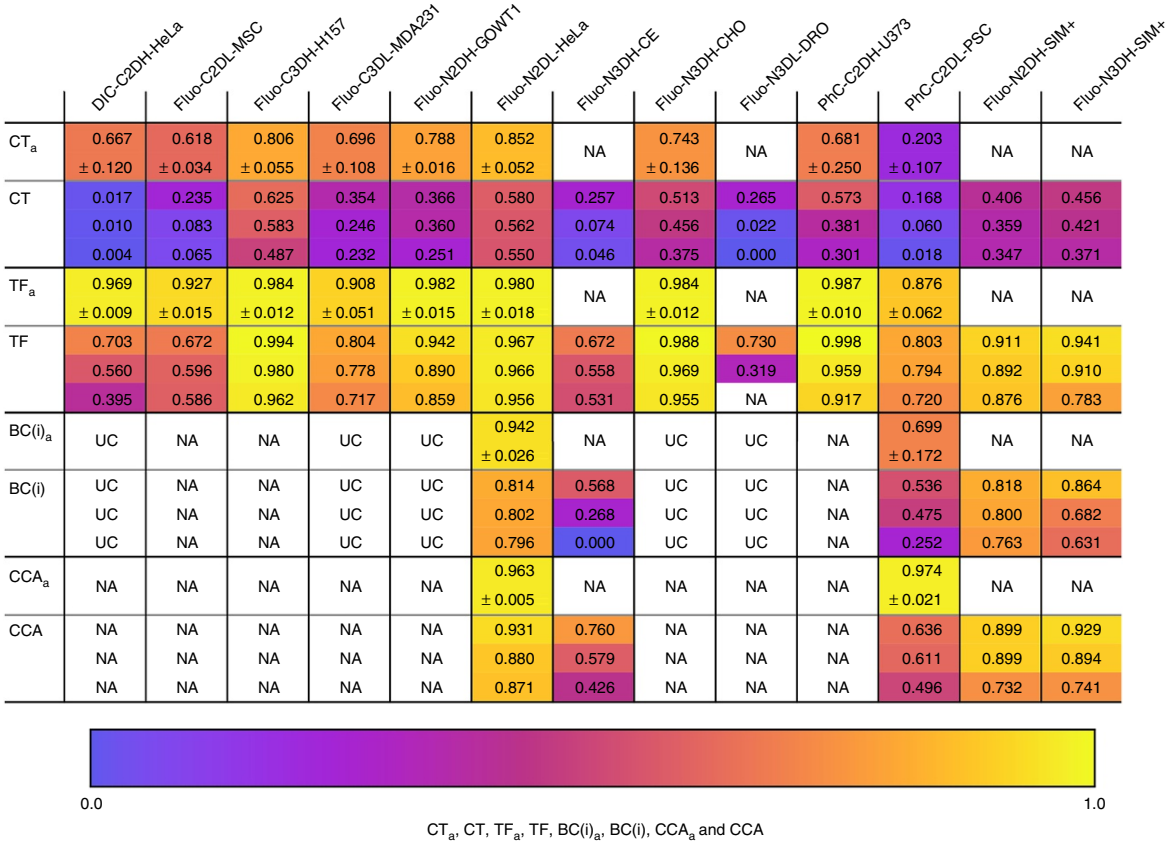


**Figure 7** | Top-three biological performance values (CT, TF, BC(i) and CCA) measures obtained by the competing algorithms. All four CT, TF, BC(i) and CCA sections start with $CT_a$, $TF_a$, $BC(i)_a$ and $CCA_a$, respectively, which are the average plus s.d. values of the measures obtained by three manual annotations used to create the ground truths (SEG-GTs and TRA-GTs), which were considered as if they were also regular submissions. If not available, the values are labeled NA. The color code below correlates with the values in the [0, 1] interval. The BC(i) measure was not calculated for the data sets that do not feature any division event (NA) or a minimum number of 50 division events in each video (UC). The tolerance parameters i used for each data set were: Fluo-N2DL-HeLa ($i$ = 1, corresponding to a 30-min tolerance window), Fluo-N3DH-CE ($i$ = 1, 1 min), PhC-C2DL-PSC ($i$ = 2, 20 min), Fluo-N2DH-SIM+ ($i$ = 3, 87 min) and Fluo-N3DH-SIM+ ($i$ = 3, 87 min). The CCA measure was not calculated for the data sets where no evidence of entire cell cycles was found (NA).

whereas PhC-C2DL-PSC, Fluo-N3DH-CE and, presumably, Fluo-N3DL-DRO would benefit from improved management of division events as revealed by their low BC and CCA values.

When analyzing the performance of the individual algorithms in terms of CT and TF (**Fig. 8** and **Supplementary Data 4**), we saw similar, but not completely matching, pictures compared with the ranking compiled using SEG and TRA (**Fig. 6**). This is because TF and CT consider only tracking correctness, regardless of the accuracy of the segmentation, and have much stricter requirements on correctly reconstructed tracks. This means that solutions with a high TRA score and low TF and CT scores still contain errors that need to be fixed to enable sound biological conclusions. The KTH-SE algorithms remained the top-ranked ones in most data sets, which highlights the importance of the inclusion of global information in the linking process, as it yields longer, correctly reconstructed tracklets. However, similar to the above-discussed SEG and TRA scores, the tracking by contour evolution method LEID-NL managed to break the dominance of tracking by detection approaches (it is top ranked twice for TF and four times for CT). This highlights the fact that tracking by contour evolution methods can be superior at following cells once a track has been initiated if the temporal resolution of the image data permits. As a final comment, methods that inherently (KTH-SE, HD-Hau-GE, IMCB-SG) or specifically (HD-Har-GE, LEID-NL) detect cell division events showed higher BC and CCA values than those that do not use specific cell division detection routines. Especially relevant is the excellent behavior of HD-Har-GE, which was ranked first three out of five possible times in the CCA category, and can therefore safely be distinguished as the best method when it comes to detecting complete cell cycles and therefore measuring cell population growth.

Finally, given that competing solutions need to be deployed by biologists who normally have little computer science experience, we analyzed the usability, speed and general applicability of all top-ranked algorithms. We found that the superior performance of the KTH-SE algorithms came, unfortunately, with the disadvantage of an elevated number of parameters compared with most other methods (in particular with the close contender FR-Ro-GE; **Table 4** and **Supplementary Data 5**). Conversely, the KTH-SE algorithms were faster than most other methods, including FR-Ro-GE (for which, however, a much faster implementation using graphics cards exists). Finally, we found that the KTH-SE methods generalized very well to similar data (high GP values). This indicates that, given a well-chosen parameter configuration, this method is likely to obtain good results also when applied on previously unseen image data of the same kind.

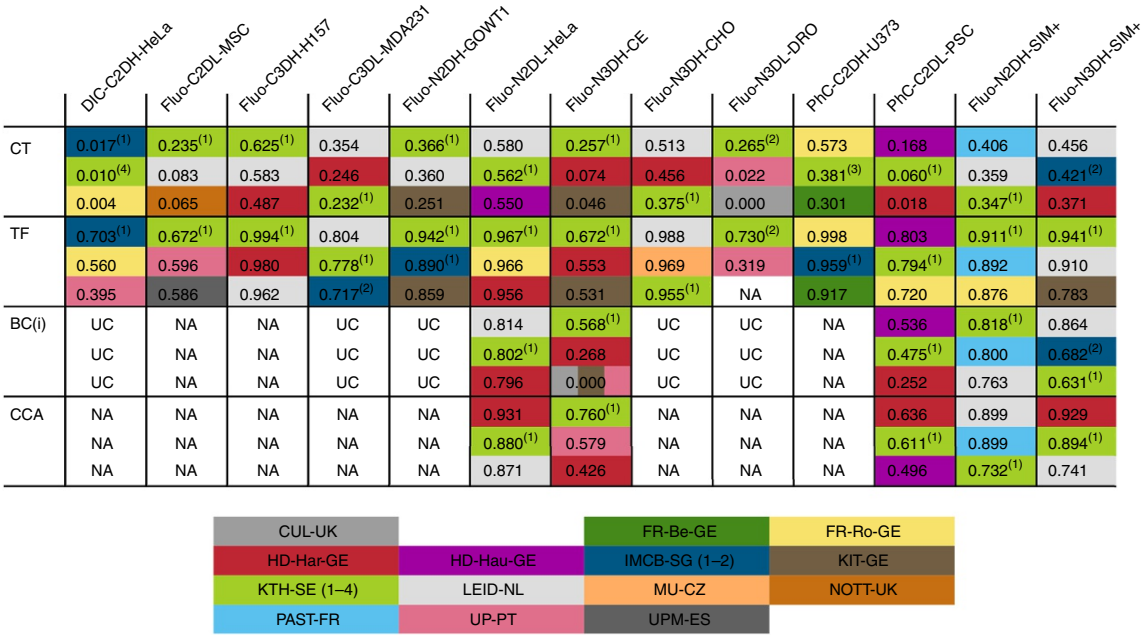


**Figure 8** | Top-three performing methods of the three challenge editions in terms of the CT, TF, BC(i) and CCA scores. Note that the methods submitted by the same participant are displayed in the same color, with super-indices denoting the particular method of the respective participant. The BC(i) measure was not calculated for the data sets that do not feature any division event (NA) or at least a minimum number of 50 division events in each video (UC). The data set Fluo-N2DL-HeLa, Fluo-N3DH-CE, PhC-C2DL-PSC, Fluo-N2DH-SIM+ and Fluo-N3DH-DIM+ was evaluated with $i = 1$ (corresponding to a 30-min tolerance window), $i = 1$ (1 min), $i = 2$ (20 min), $i = 3$ (87 min) and $i = 3$ (87 min), respectively. The CCA measure was not calculated for the data sets where no evidence of entire cell cycles was found (NA).

## DISCUSSION

Here we present the results of three editions of the CTC, a benchmarking effort aimed at improving cell tracking in multi-dimensional microscopy. The prerequisite for our study was the compilation of a large corpus of exemplar video sequences of biological samples imaged with a variety of microscopy modalities and displaying a broad range of image qualities known to be challenging for automated segmentation and tracking of cells. Our work makes a number of important contributions. First, the compilation of expert-driven annotations of cell regions and trajectories in these videos. We also include artificially generated image data at an intermediate level of complexity, for which an absolute ground truth inherently exists. Together, this represents a unique and rich resource of annotated, real and simulated image data that distinguishes our challenge from similar events that relied exclusively on simulated data[34]. Second, we developed a set of measures that quantitatively evaluate the performance of submitted solutions against the ground truth data in terms of accuracy, biological relevance of the results and usability for biologists. Third, over the course of three challenges, we assembled a diverse collection of competing solutions that represent all of the main algorithmic approaches to cell segmentation and tracking problems in biology. Fourth, we analyzed the accumulated results and provide useful guidelines for both users and developers of tracking software.

From the comparison of the competing algorithms, we found that in most practical scenarios tracking by detection methods outperformed tracking by contour evolution methods. A notable exception to this can be observed in data sets with high temporal resolutions that have substantial interframe cell overlaps. Indeed, in these situations tracking by contour evolution methods seem to be able to track cells for longer stretches of the videos than the tracking by detection methods. Paradoxically, this means that even if the results of tracking by contour evolution methods are less similar to the ground truth solution, their biologically relevant performance might be sometimes higher. Another important result of this study is that the algorithms that make use of modern machine-learning approaches performed best in most segmentation scenarios. For example, the methods that use machine-learning strategies to classify pixels as being either part of a cell or the background tended to produce better segmentation results than other methods. Furthermore, tracking by detection methods that consider larger, possibly global, spatiotemporal contexts to reason about track linking tended to outperform algorithms that only look at the nearest neighbors in space and time. The conclusion that algorithms that use prior and contextual information perform better than those that do not use it was also reached in the aforementioned Particle Tracking Challenge[34]. We found this conclusion to also be true in real data sets of moving cells with nonlinear lineages (i.e., with division events).

From the user perspective, complete and perfect unsupervised tracking remains a distant dream. When a certain level of remaining errors or manual postprocessing is acceptable, the top-scoring algorithms offer good performance. However, as a result of a large number of tunable parameters, practical deployment of the software on new data may prove to be cumbersome. Potentially, long runtimes of complex algorithmic solutions can be offset by running them on graphics hardware whenever such implementation is feasible and/or available. The good news is that once parameters have been optimized manually or using automatic supervised or unsupervised algorithms and the software runs on decent hardware, the best methods will perform well on all similar microscopy recordings. Finally, we acknowledge that, as a result of the complexity of

**Table 4** | Usability evaluation of the top-three ranked algorithms based on the overall performance measure

| | 1st ranked | | | 2nd ranked | | | 3rd ranked | | |
|---|---|---|---|---|---|---|---|---|---|
| | NP | GP | TIM | NP | GP | TIM | NP | GP | TIM |
| DIC-C2DH-HeLa | FR-Ro-GE 0.828 | | | KTH-SEM (4) 0.629 | | | IMCB-SG (1) 0.523 | | |
| | 4 | 0.912 | 4818 | 14 | 0.928 | 622 | 5 | 0.924 | 236 |
| Fluo-C2DL-MSC | KTH-SEM (1) 0.676 | | | FR-Ro-GE 0.636 | | | NOTT-UK 0.546 | | |
| | 17 | 0.893 | 79 | 4 | 0.893 | 2630 | 5 | 0.920 | 342 |
| Fluo-C3DH-H157 | KTH-SEM (1) 0.938 | | | HD-Har-GE 0.885 | | | CUNI-CZ 0.870 | | |
| | 17 | 0.966 | 16156 | 10 | 0.882 | 14110 | 8 | 0.836 | 952 |
| Fluo-C3DL-MDA231 | KTH-SEM (1) 0.757 | | | LEID-NL 0.745 | | | IMCB-SG (2) 0.659 | | |
| | 16 | 0.947 | 217 | 9 | 0.958 | 992 | 9 | 0.936 | 3506 |
| Fluo-N2DH-GOWT1 | KTH-SEM (1) 0.951 | | | LEID-NL 0.902 | | | CUNI-CZ 0.902 | | |
| | 17 | 0.955 | 632 | 9 | 0.932 | 1333 | 8 | 0.950 | 479 |
| Fluo-N2DL-HeLa | KTH-SEM (1) 0.942 | | | FR-Ro-GE 0.940 | | | HD-Har-GE 0.901 | | |
| | 17 | 0.967 | 304 | 3 | 0.963 | 22878 | 10 | 0.966 | 609 |
| Fluo-N3DH-CE | KTH-SEM (1) 0.688 | | | HD-Har-GE 0.601 | | | KIT-GE 0.507 | | |
| | 17 | 0.895 | 13475 | 9 | 0.889 | 14518 | 10 | 0.872 | 4258 |
| Fluo-N3DH-CHO | KTH-SEM (1) 0.926 | | | MU-CZ 0.912 | | | HD-Har-GE 0.906 | | |
| | 17 | 0.954 | 202 | 8 | 0.936 | 223 | 10 | 0.923 | 1495 |
| Fluo-N3DL-DRO | KTH-SEM (2) 0.609 | | | UP-PT 0.285 | | | CUL-UK 0.220 | | |
| | 20 | 0.885 | 85272 | 8 | 0.916 | 13772 | 3 | 0.973 | 6902 |
| PhC-C2DH-U373 | FR-Ro-GE 0.951 | | | FR-Be-GE 0.896 | | | KTH-SEM (3) 0.886 | | |
| | 5 | 0.965 | 11450 | 8 | 0.953 | 621 | 11 | 0.964 | 81 |
| PhC-C2DL-PSC | HD-Hau-GE 0.804 | | | KTH-SEM (1) 0.772 | | | UP-PT 0.735 | | |
| | 15 | 0.952 | 924 | 17 | 0.971 | 3481 | 11 | 0.959 | 8246 |
| Fluo-N2DH-SIM+ | FR-Ro-GE 0.878 | | | KTH-SEM (1) 0.874 | | | PAST-FR 0.859 | | |
| | 3 | 0.979 | 20124 | 17 | 0.983 | 301 | 9 | 0.978 | 370 |
| Fluo-N3DH-SIM+ | KTH-SEM (1) 0.848 | | | LEID-NL 0.798 | | | IMCB-SG (2) 0.714 | | |
| | 17 | 0.985 | 13115 | 9 | 0.973 | 66773 | 9 | 0.988 | 69549 |

NP, number of parameters; GP, generalizability measure, normalized between 0 (no generalizability) and 1 (complete generalizability); TIM, execution time in seconds. Color code: for each data set and parameter, red background indicates the worst value of the three methods, yellow indicates the intermediate value and green indicates the best value out of the three listed.

relevant factors (biological, imaging and algorithmic) that affect the results of segmentation and tracking, there is no simple way to point out the right algorithm for a given data set. This is supported by the fact that none of the presented problems were solved completely when judged from a biologist's viewpoint.

For algorithm developers, the results of the challenge indicate that their job is far from being complete. Despite the very good results the submitted algorithms achieved on many data sets, additional development is crucially required for scenarios with low SNR or CR or for tracking cells with more complex shapes or textures. Large 3D data sets, such as those of developing embryos, present additional challenges. Not only do such videos show very high cell densities in later frames, the size of the image data itself causes very long runtimes. Tracking by detection approaches fail on these data sets because they crucially depend on high-quality segmentation results, something difficult to achieve in these challenging data sets. Tracking by contour evolution approaches often fails because of their low temporal resolution.

In most circumstances, tracking is contingent on segmentation, and the submitted algorithms mix and match different segmentation and tracking strategies. By equally weighting both segmentation and tracking accuracy when calculating the overall performance of the methods, we assign equal importance to both tasks; although, as we found, the resulting ranking is robust against changes in those weights. Furthermore, the overall correlation of both measures is moderate, with only a few exceptions in data sets in which the performance of a tracking solution seems to be heavily influenced by the performance of the segmentation approach.

Although the challenge was broadly taken on by the community, and many algorithms competed, it is important to stress that the voluntary nature of participation necessarily resulted in substantial omissions. In particular, this affected the submissions attempting to meaningfully solve the 3D tracking problems in embryos that are the most challenging data sets and for which efficient methods are published and available[32,33].

The CTC, which remains open for online submissions, is a powerful resource for algorithm developers and users alike. Along with the data sets, we offer an open-source Fiji plugin[35] with the evaluation suite, which is capable of computing the technical and biologically oriented measures, as well as the data set quality parameters; and we provide executable versions of most of the participants' algorithms. Furthermore, we encourage participants to make their submitted algorithms available to biologists via easy to install and intuitive graphical user interfaces. In the future, new data sets of existing and new microscopy modalities will be incorporated to the data set repository. It will be particularly important to collect and annotate complex tissue, organ and whole-embryo image data. Finally, we intend to add new synthetic data sets that closely mimic the variety of cell types and microscopy scenarios. These synthetic image data will model different cell labeling, cell shapes and cell behaviors and migration patterns in 2D and 3D. Given that artificially generated data sets implicitly bear absolute ground truth, they can be tuned to challenge algorithms to improve specific aspects of the problem (for example, how to deal with increasing noise or signal heterogeneity levels) or provide training data for segmentation and tracking approaches based on promising machine-learning methods.

## METHODS

Methods, including statements of data availability and any associated accession codes and 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.*


### ACKNOWLEDGMENTS

We acknowledge the work of A. Urbiola, C. Ederra, T. España, S. Venkatesan, D.M.W. Balak, P. Karas, T. Bolcková, M. Štreitová, M. Charousová and L. Zátopková, who manually annotated the data sets to create the ground truths used to evaluate the performance of the algorithms. We also would like to thank F. Prósper (CIMA-University of Navarra), E. Bártová (Institute of Biophysics, Academy of Sciences of the Czech Republic), J. Essers (Erasmus University Medical Center), the Mitocheck consortium, A. Rouzaut (CIMA-University of Navarra), R. Kamm (Massachussets Institute of Technology), the Waterston Lab (The George Washington University), P. Keller (Howard Hughes Medical Institute), S. Kumar (University of California at Berkeley), G. van Cappellen (Erasmus University Medical Center) and T. Becker (Fraunhofer Institution for Marine Biology), who provided the data sets used in the three challenge editions. Finally, we thank R. Stoklasa for technical support. The participants would like to acknowledge the contributions of M. Schiegg, D. Stöckel, J. Crowe, M. Temerinac-Ott and P. Fischer. This work was funded by Spanish Ministry of Economy MINECO grants DPI2012-38090-C03-02 (C.O.-d.-S.) and DPI2015-64221-C2-2 (C.O.-d.-S.), TEC2013-48552-C2-1-R (A.M.B.), TEC2015-73064-EXP (A.M.B), and TEC2016-78052-R (A.M.B.); Netherlands Organization for Scientific Research (NWO) grants 612.001.018 (M.R. and E.M.) and 639.021.128 (I.S.); Dutch Technology Foundation (STW) grant 10443 (I.S. and E.M.); Czech Science Foundation (GACR) grant P302/12/G157 (M.K. and Pavel Matula); the Czech Ministry of Education, Youth and Sports grant LTC17016 in the frame of EU COST NEUBIAS project (M.M., Pavel Matula, Petr Matula, D.S. and M.K.); Helmholtz Association (J.S. and R.M.) and DFG grant MI 1315/4-1 (J.S. and R.M.); the Excellence Initiative of the German Federal and State Governments EXC 294 (O.R., T.B. and R.B.); the Swiss Commission for Technology and Innovation, CTI project 16997 (Ö.D. and L.M.); the BMBF, projects ENGINE (NGFN+), RNA-Code (e:Bio) and de.NBI, as well as the DFG, SFB 1129 and RTG 1653 (N.H. and K.R.);

the HGS MathComp Graduate School, the SFB 1129 for integrative analysis of pathogen replication and spread, the RTG 1653 for probabilistic graphical models, and the CellNetworks Excellence Cluster/EcTop (C.H., S.W. and F.H.); the Baxter Foundation and US National Institutes of Health grant AG020961 (H.M.B.) and the Swedish Research Council VR Grant 2015-04026 (K.M. and J.J.); and the BMBF, project de.NBI, grant 031L0102 (V.U. and F.J.).

**AUTHOR CONTRIBUTIONS**

V.U. actively participated in the organization and management of the CTC challenges by handling submissions, producing synthetic data sets, evaluating the submitted results and globally analyzing the participant's contributions, and creating annotations for data set evaluation. V.U. contributed to the writing of the manuscript and produced the tables and plot results, as well as the Fiji plugin with the evaluation suite. M.M. actively participated in the organization and management of the CTC challenges by handling and evaluating submissions, providing evaluation and annotation software, supervising annotations, and creating consensual ground truths for the evaluation of the submitted results. M.M. contributed to the writing of the manuscript and was a challenge participant. K.E.G.M., O.R. and C.H. were top ranked challenge participants and contributed to the writing of the manuscript. N.H. was a top ranked challenge participant. Pavel Matula actively participated in the organization of the CTC challenges by leading the development of a suitable tracking measure and assessing the behavior of various measures on challenge data sets. Petr Matula, M.R. and I.S. actively participated in the organization of the CTC challenges by preparing data and supervising data annotation. D.S. actively participated in the organization of the CTC challenges by leading the development of synthetic data generator and creation of suitable collection of synthetic time-lapse sequences with absolute ground truth. K.R., J.J., H.M.B., O.D., B.L., P.X., Y.L., S.-Y.C., A.C.D., J.-.C.O.-M., C.C.R.-A., J.A.S.-L., R.B., T.B., J.S., R.M., S.W., F.A.H., T.E., P.Q., Ö.D. and L.M. were challenge participants. F.J. contributed to the revision of the manuscript and supported V.U. with the related data processing. P.T., E.M., A.M.-B. and M.K. were challenge organizers and contributed to the revision of the manuscript. C.O.-d.-S. was a challenge organizer, coordinated the work of the committee that organized the challenges and wrote the manuscript with input from all of the authors.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.



1. Franz, C.M., Jones, G.E. & Ridley, A.J. Cell migration in development and disease. *Dev. Cell* **2**, 153–158 (2002).
2. Bullen, A. Microscopic imaging techniques for drug discovery. *Nat. Rev. Drug Discov.* **7**, 54–67 (2008).
3. Walter, R.J. & Berns, M.W. Digital image processing and analysis. in *Video Microscopy* (ed. Inoué, S.) 327–392 (Springer Sciences, 1986).
4. Schneider, C.A., Rasband, W.S. & Eliceiri, K.W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
5. Meijering, E. Cell segmentation: 50 years down the road. *IEEE Signal Process. Mag.* **29**, 140–145 (2012).
6. Dufour, A.C. *et al.* Signal processing challenges in quantitative 3-D cell morphology: more than meets the eye. *IEEE Signal Process. Mag.* **32**, 30–40 (2015).
7. Zimmer, C. *et al.* On the digital trail of mobile cells. *IEEE Signal Process. Mag.* **23**, 54–62 (2006).
8. Wuttisarnwattana, P., Gargesha, M., van't Hof, W., Cooke, K.R. & Wilson, D.L. Automatic stem cell detection in microscopic whole mouse cryo-imaging. *IEEE Trans. Med. Imaging* **35**, 819–829 (2016).
9. Lerner, B., Clocksin, W.F., Dhanjal, S., Hultén, M.A. & Bishop, C.M. Automatic signal classification in fluorescence in situ hybridization images. *Cytometry* **43**, 87–93 (2001).
10. Chen, X., Zhou, X. & Wong, S.T.C. Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. *IEEE Trans. Biomed. Eng.* **53**, 762–766 (2006).
11. Henry, K.M. *et al.* PhagoSight: an open-source MATLAB package for the analysis of fluorescent neutrophil and macrophage migration in a zebrafish model. *PLoS One* **8**, e72636 (2013).
12. Wählby, C., Sintorn, I.M., Erlandsson, F., Borgefors, G. & Bengtsson, E. Combining intensity, edge and shape information for 2D and 3D segmentation of cell nuclei in tissue sections. *J. Microsc.* **215**, 67–76 (2004).
13. Cicconet, M., Geiger, D. & Gunsalus, K. Wavelet-based circular hough-transform and its application in embryo development analysis. in *Proc. of the International Conference on Computer Vision Theory and Applications* 669–674 (Science and Technology Publications, 2013).
14. Türetken, E., Wang, X., Becker, C.J., Haubold, C. & Fua, P. Network flow integer programming to track elliptical cells in time-lapse sequences. *IEEE Trans. Med. Imaging* **36**, 942–951 (2017).
15. Malpica, N. *et al.* Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry* **28**, 289–297 (1997).
16. Ortiz de Solórzano, C. *et al.* Segmentation of confocal microscope images of cell nuclei in thick tissue sections. *J. Microsc* **193**, 212–226 (1999).
17. Cliffe, A. *et al.* Quantitative 3D analysis of complex single border cell behaviors in coordinated collective cell migration. *Nat. Commun.* **8**, 14905 (2017).
18. Ronneberger, O., Fisher, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. in *Proc. MICCAI 2015 LNCS* **9351**, 234–241 (Spring, Cham, 2015).
19. Schiegg, M. *et al.* Graphical model for joint segmentation and tracking of multiple dividing cells. *Bioinformatics* **31**, 948–956 (2015).
20. Zimmer, C., Labruyère, E., Meas-Yedid, V., Guillén, N. & Olivo-Marin, J.-C. Segmentation and tracking of migrating cells in videomicroscopy with parametric active contours: a tool for cell-based drug testing. *IEEE Trans. Med. Imaging* **21**, 1212–1221 (2002).
21. Dufour, A., Thibeaux, R., Labruyère, E., Guillén, N. & Olivo-Marin, J.C. 3-D active meshes: fast discrete deformable models for cell tracking in 3-D time-lapse microscopy. *IEEE Trans. Image Process.* **20**, 1925–1937 (2011).
22. Maška, M. *et al.* Segmentation and shape tracking of whole fluorescent cells based on the Chan-Vese model. *IEEE Trans. Med. Imaging* **32**, 995–1006 (2013).
23. De Solorzano, C.O., Malladi, R., Lelièvre, S.A. & Lockett, S.J. Segmentation of nuclei and cells using membrane related protein markers. *J. Microsc.* **201**, 404–415 (2001).
24. Dzyubachyk, O., van Cappellen, W.A., Essers, J., Niessen, W.J. & Meijering, E. Advanced level-set-based cell tracking in time-lapse fluorescence microscopy. *IEEE Trans. Med. Imaging* **29**, 852–867 (2010).
25. Dufour, A. *et al.* Segmenting and tracking fluorescent cells in dynamic 3-D microscopy with coupled active surfaces. *IEEE Trans. Image Process.* **14**, 1396–1410 (2005).
26. Bensch, R. & Ronneberger, O. Cell segmentation and tracking in phase contrast images using graph cut with asymmetric boundary costs. In *Proc. 2015 IEEE Int. Symp. Biomed. Imaging (ISBI)* 1120–1123 (2015).
27. Harder, N. *et al.* Automatic analysis of dividing cells in live cell movies to detect mitotic delays and correlate phenotypes in time. *Genome Res.* **19**, 2113–2124 (2009).
28. Bise, R., Yin, Z. & Kanade, T. Reliable cell tracking by global data association. in *Proc. 2011 IEEE Int. Symp. Biomed. Imaging (ISBI)* 1004–1010 (2011).
29. Magnusson, K.E.G., Jaldén, J., Gilbert, P.M. & Blau, H.M. Global linking of cell tracks using the Viterbi algorithm. *IEEE Trans. Med. Imaging* **34**, 911–929 (2015).
30. Maška, M. *et al.* A benchmark for comparison of cell tracking algorithms. *Bioinformatics* **30**, 1609–1617 (2014).
31. Svoboda, D. & Ulman, V. MitoGen: A framework for generating 3D synthetic time-lapse sequences of cell populations in fluorescence microscopy. *IEEE Trans. Med. Imaging* **36**, 310–321 (2017).
32. Murray, J.I. *et al.* Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. *Nat. Methods* **5**, 703–709 (2008).
33. Amat, F. *et al.* Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nat. Methods* **11**, 951–958 (2014).
34. Chenouard, N. *et al.* Objective comparison of particle tracking methods. *Nat. Methods* **11**, 281–289 (2014).
35. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

## ONLINE METHODS

**Data set quality parameters.** To assess the quantitative video parameters (see **Table 1**), we had to calculate those parameters, ideally, on a complete ground truth of the competition data sets, meaning having appropriate cell masks and tracking information for all the cells in the videos. The ground truth used to evaluate the performance of the algorithms (SEG-GT and TRA-GT) was obtained manually from three annotators. TRA-GT indeed contains the manually annotated tracks of all the cells in the videos. However, due to the monumental task that it would have required, SEG-GT includes a subset of complete segmentation masks per video, which constitutes a representative amount for the evaluation of segmentation performance. To extend the manual ground truth to cover as many as possible of the cells in the videos, we first combined the manual tracking ground truth (TRA-GT) with the segmentation masks provided by the participants. For any marker in TRA-GT, we automatically merged the top-performing participants' segmentation masks that overlap the majority of this tracking marker. The number of masks used was determined manually for each video. On average, a majority of the total number of available masks were used. The process led occasionally to colliding situations, i.e., when obtained segmentation masks for two different tracking markers were overlapping. If the overlap was less than 10% of the mask area/volume, the intersecting pixels/voxels were removed from both colliding masks in an expectation that 10% loss will not significantly influence the measured quantities. Otherwise, both entire masks were discarded. In this way, a rich consensus-based segmentation with reliable linking was obtained for all real challenge videos. The synthetic data sets did not require this process, since they are accompanied with the absolute segmentation and tracking ground truth, inherently generated during the simulation process.

Next, a mask for the background region of each video was established as the complement to the union of all objects' consensus segmentation masks taken over all frames of the given video. This results in a constant -stationary over the video- background mask that fits to all images of that video. A background mask for synthetic data sets was established also like this. For Fluo-N3DH-CE and Fluo-N3DL-DRO data sets, however, the background masks had to be established on per-frame basis, encompassing interior region of the embryos as well as the surrounding medium.

From the consensus segmentation and tracking ground truth, we calculated quantitative parameters as follows. Let $\mathrm{FG}_{i,t}$ and $\mathrm{BG}_t$ represent the sets of image elements that form $i$-th cell and (single) background mask, respectively, in $t$-th image of the video. Furthermore, let avg($S$) and s.d.($S$) denote average and s.d. of intensities found at image elements in the set $S$, and let dist($a$, $b$) be a chamfer distance[36] between image elements a and b in their coordinate units (pixels/voxels in 2D/3D). The reported values of the signal-to-noise ratio (SNR), contrast ratio (CR), internal signal heterogeneity of the cells ($\mathrm{Het}_i$), resolution ($Res$), regularity of the cell shape (Sha), cell density (Den), and level of cell overlap in consecutive frames (Ove) were established as averages of $SNR_{i,t}$, $CR_{i,t}$, $\mathrm{HETi}_{i,t}$, $\mathrm{Res}_{i,t}$, $\mathrm{Sha}_{i,t}$, $\mathrm{Den}_{i,t}$,and $\mathrm{Ove}_{i,t}$ values, respectively, calculated for every object in every image in both competition videos

$$\mathrm{SNR}_{i,t} = \frac{|\mathrm{avg}(\mathrm{FG}_{i,t}) - \mathrm{avg}(\mathrm{BG}_t)|}{\mathrm{std}(\mathrm{BG}_t)}$$

$$\mathrm{CR}_{i,t} = \frac{\mathrm{avg}(\mathrm{FG}_{i,t})}{\mathrm{avg}(\mathrm{BG}_t)}$$

$$\mathrm{HETi}_{i,t} = \frac{\mathrm{std}(\mathrm{FG}_{i,t})}{|\mathrm{avg}(\mathrm{FG}_{i,t}) - \mathrm{avg}(\mathrm{BG}_t)|}$$

$$\mathrm{HETb}_{i,t} = \frac{\mathrm{avg}(\mathrm{FG}_{i,t}) - \mathrm{avg}(\mathrm{BG}_t)}{\sum_{j \in I(t)} |\mathrm{avg}(\mathrm{FG}_{j,t}) - \mathrm{avg}(\mathrm{BG}_t)| / |I(t)|}$$

$$\mathrm{Res}_{i,t} = |\mathrm{FG}_{i,t}|$$

$$\mathrm{Den}_{i,t} = \min\{50, \mathrm{dist}(a,b) \,||\, a \in \mathrm{FG}_{i,t}, b \in \mathrm{FG}_{j,t}, j \in I(t), j \neq i\}$$

$$\mathrm{Ove}_{i,t} = \frac{|\{a \in \mathrm{FG}_{i,t} \,||\, \exists b \in FG_{i,t-1} : \mathrm{dist}(a,b)=0\}|}{|FG_{i,t}|}$$

where $|S|$ is the size of the set $S$ and $I(t)$ is the set of indices of all cells or nuclei segmented in the $t$-th image. The heterogeneity of the signal between cells ($\mathrm{Het}_b$) is calculated as the s.d. of $\mathrm{HETb}_{i,t}$ values for every object in every image in both competition videos. $\mathrm{Sha}_{i,t}$ is the circularity[37] for 2D objects, which is given as the normalized ratio of perimeter of a circle having the same area as the object to the actual area of the object, and sphericity[37] for 3D objects, which is given as the normalized ratio of the surface area of a sphere having the same volume as the object to the actual surface area of the object. Note that in the latter case the actual (anisotropic) voxel size was taken into account. The $\mathrm{Den}_{i,t}$ was evaluated only up to the distance of 50 image elements away from $i$-th object. The distance tells how many (background) pixels/voxels there are between two nearby objects. Clearly, higher number expects separating nearby objects easier. To calculate Cha, the absolute difference between the average object intensity at the end and the beginning of a video was divided by the number of its frames minus one and averaged over both videos in a data set. The number of division events (Mit) is computed as average of $\mathrm{Mit}_t$ taken over images from both videos, where $\mathrm{Mit}_t$ is the number of objects whose tracks end in the $t$-th image because of subsequent division events (which are marked in the tracking ground truth TRA-GT). The remaining qualitative parameters, synchronization of division events (Syn), cells entering or leaving the field of view (Ent/Leav), apoptotic cells (Apo), and the presence of moving debris (Deb), were set after manual inspection of the data sets.

**Performance criteria (technical measures).** *Segmentation Accuracy.* We quantify the amount of overlap between the reference annotations and the computed segmentation results using the Jaccard similarity index, defined as

$$J(R,S) = \frac{|R \cap S|}{|R \cup S|}$$

where $R$ is the reference segmentation of a cell in SEG-GT and $S$ is its corresponding cell segmentation. The Jaccard index always falls in the [0, 1] interval, where 1 means total overlap and 0 means no overlap. The final SEG value for a particular video is calculated as the mean Jaccard index over all reference cells in the video.

*Tracking accuracy.* To evaluate the ability of an algorithm to track cells in time, the tracking results are first represented as acyclic oriented graphs, as trees that capture the genealogy of the cells during the duration of the video. We then assess how difficult it is to transform a computed tracking graph into the corresponding reference graph, TRA-GT, using a normalized version of the Acyclic Oriented Graph Matching (AOGM) measure[38]

$$\mathrm{TRA} = 1 - \frac{\min(\mathrm{AOGM}, \mathrm{AOGM}_0)}{\mathrm{AOGM}_0}$$

where $\mathrm{AOGM}_0$ is the AOGM value required for creating the reference graph from scratch (i.e., it is the AOGM value for empty tracking results). The minimum operator in the numerator prevents from having a final negative value when it is cheaper to create the reference graph from scratch than to transform the computed graph into the reference graph. TRA always falls in the [0, 1] interval, with higher values corresponding to better tracking performance.

*Overall performance.* For each algorithm and data set, SEG and TRA are first averaged over the two competition videos. Then, the averaged values, $\mathrm{SEG}_{\mathrm{avg}}$ and $\mathrm{TRA}_{\mathrm{avg}}$, are averaged again (i.e., $\mathrm{OP} = 0.5 \cdot (\mathrm{SEG}_{\mathrm{avg}} + \mathrm{TRA}_{\mathrm{avg}})$), and the result is used to compile the final ranking.

**Performance criteria (biologically inspired measures).** *Complete tracks.* CT[39] examines how good a method is at reconstructing complete reference tracks (i.e., the tracks in TRA-GT). A reference track is considered completely reconstructed if and only if each of its track points has an assigned track point in the corresponding computed track, and both tracks have the same temporal support. The final CT value for a particular video is computed as the $F_1$ score of completely reconstructed reference tracks, defined as:

$$\mathrm{CT} = \frac{2T_{rc}}{T_c + T_{gt}}$$

where $T_{rc}$ is number of completely reconstructed reference tracks, $T_{gt}$ is number of all reference tracks, and $T_c$ is the number of all computed tracks.

*Track fractions.* TF targets the longest, correctly reconstructed, continuous fraction of a detected reference track. The final TF value for a particular video is computed by averaging these fractions over all detected reference tracks.

*Branching correctness.* BC(i)[28,29] examines how good a method is at reconstructing mother-daughter relationships. Division events often happen during several frames, thus complicating matching of the provided result and the ground truth. Therefore, for two division events to be considered matching[29,30] (i.e., one provided by the method and one in the ground truth), they are allowed to be separated by no more than $i$ frames. More specifically, we allowed the reconstruction of division events using a tolerance window of $(2.i + 1)$ frames. The tolerance value i used for each data set was fixed by analyzing how the performance of the participating methods depends on $i$. Namely, the value $i$ was selected as the minimum value that was large enough to ensure that the BC(i) values of all competitive methods remain constant. The actual i values used for individual data sets were: Fluo-N2DL-HeLa ($i$ = 1, corresponding to a 30-min tolerance window), Fluo-N3DH-CE ($i$ = 1, 1 min), PhC-C2DL-PSC ($i$ = 2, 20 min), Fluo-N2DH-SIM+ ($i$ = 3, 87 min), and Fluo-N3DH-SIM+ ($i$ = 3, 87 min). The final BC(i) value for a particular video is computed as the $F_1$ score of correctly reconstructed division events in the corresponding reference graph.

*Cell cycle accuracy.* CCA reflects the ability of an algorithm to discover true distribution of cell cycle lengths in a video, considering only those tracks that are both initiated and terminated by a branching event. Each such track witnesses the development of a cell from its birth until its next division, and its length, therefore, corresponds to the cell cycle length of that cell. The CCA measure is defined as:

$$\mathrm{CCA} = 1 - \max_l(|\,\mathrm{CDF}_r(l) - \mathrm{CDF}_{gt}(l)\,|)$$

where $\mathrm{CDF}_r$ and $\mathrm{CDF}_{gt}$ are cumulative distribution functions of cell cycle length occurrence probabilities in the reference annotation and the computed result, respectively, adopting a common non-parametric approach to discovering dissimilarities between two sample distributions[40].

It is important to note that CT, TF, BC(i) and CCA always fall into the [0, 1] interval, with higher values corresponding to better performance.

**Performance criteria (usability measures).** *Number of required tunable paramters.* NP corresponds to the number of parameters that need to be provided, and possibly tuned, to obtain the evaluated results. Although there are methodologies that allow for automatic tuning of the parameters, having to do so adds a level of complexity to the task that might prevent a very efficient algorithm from being used by a user non-proficient in those methods.

*Generalizability.* GP examines how stable the algorithm is when being applied to similar image data using the set of parameters provided. Being evaluated for all 21 algorithms, we ran the algorithms on the training videos using the same parameters provided for the competition videos and evaluated how much the results for the training videos differ from those for the competition videos in terms of the technical measures:

$$\mathrm{GP} = \frac{(1 - \mathrm{SEG}_{\mathrm{avg}}^{\mathrm{GP}}) + (1 - \mathrm{TRA}_{\mathrm{avg}}^{\mathrm{GP}})}{2}$$

where $\mathrm{SEG}_{\mathrm{avg}}^{\mathrm{GP}}$ and $\mathrm{TRA}_{\mathrm{avg}}^{\mathrm{GP}}$ are average absolute differences in the SEG and TRA scores, respectively, between the results obtained for the competition and training videos. Note that GP always falls into the [0, 1] interval, with higher values corresponding to higher generalizability.

*Execution time.* For each data set, we accumulated the time (in seconds) that was required to analyze each competition video.

**Ranking robustness.** For each dataset, we ranked all methods based on their SEG and TRA scores using the formula $0.5 \cdot (a \cdot \mathrm{SEG} + \mathrm{b} \cdot \mathrm{TRA})$, $a, b \in \{0, 0.001, 0.002, \ldots, 1\}$, and calculated the number of changes between each such ranking and the one compiled using OP (i.e., when $a$ equals to $b$). **Supplementary Figure 14** plots the number of changes for every combination of weights. As can be seen, 45% of the area (that is of possible weight





doi:10.1038/nmeth.4473

configurations) causes no more than two changes in the rankings across all data sets.

**Code availability.** All the code used to produce the results reported in this article, namely a Fiji plugin that implements the entire evaluation suite (used to produce the numbers listed in **Tables 1** and **4**, **Figs. 5–8**, and **Supplementary Figs. 13** and **14**), is freely available through the link to the CTC website given in **Supplementary Table 3**, along with the links to the executable versions of individual algorithms of those participants who agreed to share their tools. The parameters used by the participants to produce their submitted results are listed in **Supplementary Data 2**.

**Data availability statement.** All the data sets used in the challenge (referred to in **Fig. 4**, **Supplementary Figs. 1–11**, **Supplementary Videos 1–13**, and described in **Table 1** and **Supplementary Table 1** and **Supplementary Note 1**), along with the annotations of the training data sets, are available through the challenge website: http://celltrackingchallenge.net/datasets.html. Access to the data sets is granted after free registration for the challenge.

The set of parameters used for the generation of the synthetic data sets (referred to in **Fig. 4**, **Supplementary Fig. 12**, **Supplementary Videos 12** and **13**, and described in **Table 1** and **Supplementary Table 1**) is given in **Supplementary Data 1**.

The entire set of evaluation measures obtained and used to compare the algorithms (used to produce **Figs. 5–8**, **Table 4**, **Supplementary Figs. 13** and **14**, and **Supplementary Table 4**) is provided with this article as **Supplementary Data 3** (SEG, TRA and OP), **4** (CT, TF, BC and CCA), and **5** (NP, GP and TIM).

A **Life Sciences Reporting Summary** is provided.


36. Klette, R. & Zamperoni, P. *Handbook of Image Processing Operators* (New York, Wiley, 1996).
37. Lin, C.L. & Miller, J.D. 3D characterization and analysis of particle shape using X-ray microtomography (XMT). *Powder Technol.* **154**, 61–69 (2005).
38. Matula, P. *et al.* Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. *PLoS One* **10**, e0144959 (2015).
39. Li, K. *et al.* Cell population tracking and lineage construction with spatiotemporal context. *Med. Image Anal.* **12**, 546–566 (2008).
40. Brown, M.R. *et al.* Flow-based cytometric analysis of cell cycle via simulated cell populations. *PLOS Comput. Biol.* **6**, e1000741 (2010).

nature research

Corresponding author(s): Carlos Ortiz de Solórzano

☐ Initial submission ☐ Revised version ☒ Final submission

# Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

## Experimental design

1. Sample size

   Describe how sample size was determined.

   Our manuscript does not report on experimental work. We evaluate and rank the performance of software -cell tracking algorithms on videos- based on a set of performace measures. Therefore, no descriptive statistics have been used and accordingly all the following questions have been answered (N/A).
   Regarding the sample size, as explained in the Results section, "Datasets and ground truth" subsection (page 6), we used 52 annotated videos, 4 videos of 13 types, covering a wide range of microscopy and experimental conditions. From each type, two videos were used for training the algorithms and two videos were used to evaluate the performance of the algorithms. The number of videos per dataset (4) was considered appropriate taking into account the labor intense annotation required, the amount of work given to the participants, and the availability of good quality videos of each type.

2. Data exclusions

   Describe any data exclusions.

   N/A

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   N/A

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   N/A

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   N/A

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.





Nature Methods: doi:10.1038/nmeth.4473

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

| n/a | Confirmed |
|---|---|
| ☒ | ☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) |
| ☒ | ☐ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☐ A statement indicating how many times each experiment was replicated |
| ☒ | ☐ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) |
| ☒ | ☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons |
| ☒ | ☐ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted |
| ☒ | ☐ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range) |
| ☒ | ☐ Clearly defined error bars |

*See the web collection on statistics for biologists for further resources and guidance.*

## Software

Policy information about availability of computer code

7. Software

| | |
|---|---|
| Describe the software used to analyze the data in this study. | We have developed code to analyze the performance of the algorithms, and quantify the properties of the videos, to help with the interpretation of the results. A beta version of a Fiji plugin that contains the software in provided as a link in Supplementary Table 3, along with links to the executable versions of the participant's algorithms. |

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

## Materials and reagents

Policy information about availability of materials

8. Materials availability

| | |
|---|---|
| Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company. | N/A |

9. Antibodies

| | |
|---|---|
| Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species). | M/A |

10. Eukaryotic cell lines

| | |
|---|---|
| a. State the source of each eukaryotic cell line used. | N/A |
| b. Describe the method of cell line authentication used. | N/A |
| c. Report whether the cell lines were tested for mycoplasma contamination. | N/A |
| d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use. | N/A |





Nature Methods: doi:10.1038/nmeth.4473

## Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

| | |
|---|---|
| Provide details on animals and/or animal-derived materials used in the study. | N/A |

Policy information about studies involving human research participants

12. Description of human research participants

| | |
|---|---|
| Describe the covariate-relevant population characteristics of the human research participants. | N/A |





Nature Methods: doi:10.1038/nmeth.4473