An Objective Comparison of Cell Tracking Algorithms

Vladimír Ulman[1¶**], Martin Maška[1**], Klas E. G. Magnusson[2], Olaf Ronneberger[3], Carsten Haubold[4], Nathalie Harder[5¶¶], 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¶¶¶], Yuexiang Li[11¶¶¶¶], Siu-Yeung Cho[12], Alexandre 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 Tomančák[19], Erik Meijering[6], Arrate Muñoz-Barrutia[20,21], Michal Kozubek[1], Carlos Ortiz-de-Solorzano[22*]

[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 Signalling Studies University of Freiburg, Germany
[4] Heidelberg Collaboratory for Image Processing, IWR, University of Heidelberg, Germany
[5] Biomedical Computer Vision Group, Dept. 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, CA, 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, United Kingdom
[12] Faculty of Engineering, University of Nottingham, Ningbo, China
[13] BioImage Analysis Unit, Institute Pasteur, Paris, France
[14] Research Centre in Biomedical Engineering, School of Mathematics, Computer Science and Engineering, City University of London, United Kingdom
[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, Switzerland
[19] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
[20] Bioengineering and Aerospace Engineering Department, Universidad Carlos III de Madrid, Spain
[21] Instituto de Investigación Sanitaria Gregorio Marañon, Madrid, Spain
[22] CIBERONC, and Program of Solid Tumors and Biomarkers, Center for Applied Medical Research, University of Navarra, Pamplona, Spain

[¶] Current affiliation: Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
[¶¶] Current affiliation: Definiens AG, Munich, Germany
[¶¶¶] Current affiliation: National Heart Research Institute Singapore (NHRIS), National Heart Centre Singapore (NHCS), Singapore
[¶¶¶¶] Current affiliation: Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[**] These authors contributed equally to this work
[*] Corresponding author

Abstract

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 tracking algorithms in multidimensional time-lapse microscopy. With seventeen participants providing results of twenty-one algorithms on a data repository consisting of thirteen datasets of various microscopy modalities and experimental conditions, this challenge displays a good sample of today's state of the art in the field. We first determine the quality of all datasets and assess the complexity of the associated segmentation and tracking tasks. Next, we analyze the results received during the three editions of the challenge. To this end, we use performance measures for segmentation and tracking that rank all participating methods. For each dataset we highlight the particular issues that remain to be addressed by future algorithmic advances. Finally, we analyze the performance of all algorithms in terms of biological measures and their practical usability. This allows us to directly address the biologist/user point of view. Even though some methods score high in all technical aspects, not a single one obtains fully correct solutions. This is especially true when considering measures that are of biological relevance, such as the correct detection of entire tracks or cell lineages. We also show that methods that either take prior information into account using learning strategies or analyze cells in a larger spatio-temporal image or video context perform better than other methods under the segmentation and tracking scenarios included in this challenge.

Introduction

Cell migration and proliferation are two important processes in normal tissue development and disease[1]. To visualize these processes, optical microscopy remains the most appropriate imaging modality[2]. Some 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 requires internalized, transgenic, or transfected fluorescent reporters to specifically label cell components such as nuclei, cytoplasm, or membranes. These are then made visible in 2D by wide-field fluorescence microscopy or in 3D by using the optical sectioning capabilities of confocal, multiphoton, or light sheet microscopes.

In order to gain biological insights from such time-lapse 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 quantitative information from microscopy images of cultured cells[3,4]. Recently, the advent of new imaging technologies has challenged the field with multi-dimensional, large image datasets following the development of tissues, organs, or entire organisms. Yet the tasks remain the same, accurately delineating (i.e. segmenting) cell boundaries and tracking cells movements over time, providing information about their velocities and trajectories, and detecting cell lineage changes due to cell division or cell death (see **Fig. 1** for a graphical description of these concepts). The level of difficulty of automatically segmenting and tracking cells depends on the quality of the recorded video sequences. In most practical scenarios, this quality is the result of a number of occasionally mutually contradicting biological[a] and technical[b] factors. The main properties that determine the quality of time-lapse videos with respect to the subsequent segmentation and tracking analysis are discussed in the following paragraphs, graphically illustrated in **Fig. 2**, and expressed as a set of quantitative measures in the **Online Methods** (section **Dataset quality parameters**).

The main factors that influence the outcome of image segmentation are the signal to noise and contrast ratios (***SNR*** and ***CR***), measuring the relationship between the signal captured from the cells and the unwanted noise or signal captured at the same time (**Fig. 2a-f**). Additionally, **intra-cellular heterogeneity** can lead to cell over-segmentation when the same cell yields several detections (**Fig. 2g-h**) while **heterogeneity between cells** has the opposite effect resulting in undetected cells (**Fig. 2i**). Insufficient **spatial resolution** compromises the accurate detection of cell boundaries (**Fig. 2j-l**). Besides the factors that relate to the imaging process, biological features such as ***irregularity of cell shapes*** (**Fig. 2m,n**) or ***high density of cells*** (**Fig. 2o**) can cause over/under-segmentation, especially when the segmentation methods

[a] e.g. type and efficiency of the labeling process, cell type, cell movement speed, cell viability, phototoxicity, etc.

[b] e.g. acquisition time, sensitivity of the detector, image resolution, objective lens used, etc.

assume simpler, non-touching objects. Moreover, the imaging and biological effects often conspire. High noise levels, for instance, can lead to discontinuities in segmentation of thin, elongated filopodium-like structures (**Fig. 2n)**. In the time domain, **changes in the average intensity** of cells complicate their segmentation by bringing *SNR* or *CR* to levels below detection, causing signal heterogeneity between frames, and affecting both segmentation and tracking (**Fig. 2p-r**)**.**

While the above-mentioned properties of microscopic recordings influence primarily the segmentation of cells, other properties impact specifically the tracking problem. For instance, the **temporal resolution** combined with the speed of cell movement determines the amount of overlap between the cells in consecutive frames (**Fig. 2s-u**). Since many algorithms rely on this overlap, tracking is compromised when overlap decreases or even vanishes completely. Similarly, **cell divisions** pose challenges on tracking **(Fig. 2v-x**), since tracking a mitotic cell requires correctly assigning the mother to its daughter cells in consecutive frames. This is particularly difficult when the cell density is high and/or when the mitotic events are synchronized **(Fig. 2x**). Finally, cells may die during the recording or enter/leave the field of view, which further complicates their tracking.

The image processing community has addressed the above-mentioned issues by increasingly sophisticated segmentation and tracking algorithms[5-7]. Below we briefly summarize the most commonly used methods for segmentation and tracking, respectively (**Fig. 3**).

For segmentation, creating a 'taxonomy of methods' is not straightforward since the state-of-the-art methods usually combine different strategies to achieve improved results. We classify existing cell segmentation algorithms based on three criteria: *(i)* The *principle* upon which cells are detected, e.g. by finding uniform areas, boundaries, or at very low resolution by simply finding bright spots/maxima[8]. *(ii)* The image *features* that are computed to achieve the cell segmentation. These can be simple pixel/voxel or average region intensities, or more complex local image descriptors of shapes or textures. *(iii)* Finally, we distinguish the segmentation *method* itself that implements the *principle* using the *features*. The methods range from simple thresholding[9,10], hysteresis thresholding[11], edge detection[12], or shape matching[13,14], to more sophisticated region growing[15-17], machine learning[18,19], or energy minimization[20-26] approaches.

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

The diversity of imaging modalities, cell tracking tasks, and available algorithms make 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 datasets. We addressed these problems by organizing three Cell Tracking Challenges (CTC I-III), under the auspices of the IEEE International Symposium on Biomedical Imaging (ISBI) between 2013 and 2015. For these challenges, we created a diverse repository of widefield and confocal fluorescence microscopy videos, created a set of reference annotations, and defined quantitative evaluation measures to allow a fair comparison of the competing algorithms[30].

Here we present a combined report on all three CTC editions. We introduce all datasets and quantify the level of difficulty each dataset poses on segmentation and tracking. Next, we show the results obtained by the participating algorithms on the CTC datasets. The analysis of results provides useful guidelines for users to identify appropriate algorithms for their own datasets, and point developers to open challenges that we believe are insufficiently addressed by the competing algorithms. It is important to note that this is an open-source initiative that remains open online, and most of the competing methods are publicly available through the challenge website and server (http://www.codesolorzano.com/Challenges/CTC/).

Results

Datasets and ground truth

The dataset repository consists of 52 annotated videos from 13 classes, in total representing 92 GB of raw image data. Eleven datasets are contrast enhancing (PhC, DIC) or fluorescence (widefield, confocal, light sheet) microscopy recordings of live cells and organisms in 2D and 3D. The other two datasets are synthetic, generated using a cell simulator that produces realistic 2D and 3D renderings of chromatin-stained live cells[31]. Each dataset consists of two training and two competition videos. The training videos were provided at the time of registration for the CTC. The competition videos were provided at a later time, but at least two months before the submission deadline, to allow the participants to optimize their algorithms on them before submitting their results. The training videos are provided with the corresponding reference annotations, while the annotations for the competition videos are kept secret.

Fig. 4 shows a representative 2D frame from all datasets used in the three CTC editions, and **Supplementary Videos 1** to **13** contain renderings of fragments of one video per dataset. **Supplementary Material: Table 1** lists experimental and technical details of the datasets, and the **Online Methods** (section **Description of datasets**) contains a more detailed description of the datasets, including their possible biological uses. The **Online Methods** (section **Simulation system used and its parameters**) briefly describes the simulator used to create the synthetic datasets and provides the parameters used in the simulations. Finally, and more importantly, **Table 1** provides a quantitative characterization of the quality of each dataset, based on the set of measures described in the **Online Methods** (section **Dataset quality parameters**). In all tables, figures, and videos we use a naming convention for datasets that identifies their microscopy modality (**Fluo**rescence, **DIC**, **PhC**), the staining (**N**uclear, **C**ellular), the dimensionality (**2D**, **3D**), the resolution (**L**ow, **H**igh), and the cell type or model organism used.

Three independent human experts annotated all microscopy videos of cells in culture. Each expert created a segmentation and a tracking solution (annotation) for each video[30]. At the end, to account for inter-annotator variability, the final segmentation (**SEG-GT**) and tracking (**TRA-GT**) ground truths were created by combining the three user-generated annotations, following a majority-voting scheme[30]. The **SEG-GT** for the embryonic datasets (the *C.elegans* embryo Fluo-N3DH-CE and the *Drosophila melanogaster* embryo datasets Fluo-N3DH-CE and Fluo-N3DL-DRO) were generated as described above, but in the case of Fluo-N3DL-DRO, only a small subset of cells of the early nervous system was annotated and used as ground truth. The **TRA-GT** of both embryonic datasets was not created following the description above. Instead, it was created by the groups that provided the datasets, using published protocols[32,33].

Participants, algorithms, and handling of submissions.

Seventeen teams from 11 countries participated in the three CTC editions, all providing complete tracking results for at least one of the datasets. 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 (**Table 2**) and tracking (**Table 3**) strategies. **Supplementary Material: Table 2** lists affiliations of the participating teams[c] and **Supplementary Material: Table 3** contains links to executable versions of the submitted algorithms. An expanded description of the algorithms is also presented in the **Supplementary Material: Cell Tracking Algorithms** and the parameter configurations used by each algorithm are listed in the **Supplementary Data 1**. In all tables, the algorithms are named starting with an acronym that identifies the institution that hosts the participating group (e.g., KTH, COM, etc.), followed by an acronym that identifies the country where the institution is located (e.g., SE, USA, etc). When two or more participants belong to the same institution, an intermediate acronym identifies the person responsible for the submission. Finally, if the same participant submits more than one algorithm, the name is followed by a bracketed numeral that identifies the algorithm.

All submissions were received by the CTC organizers as labeled segmentation masks and structured text files containing the cell lineage graphs[d]. The organizers evaluated the data using the set of technical measures described below to generate a provisional ranking. This ranking was later confirmed by reproducing the results on a single computer, using the executable version of each algorithm provided by the participants.

Quantitative performance criteria.

In order to quantify the performance of all submitted algorithms, we developed three categories of measures that are meant to quantify the *(i)* segmentation and tracking accuracy from the computer science point of view, *(ii)* biological relevance of the obtained tracking results, and *(iii)* practical usability of the methods. A rigorous description of all measures can be found in the **Online Methods** (section **Performance criteria**). Please note that only the first set of measures was evaluated in the challenge and, therefore, the methods were only fine-tuned in this respect. The other two sets are used to analyze aspects that are of relevance from the user point of view.

The first set measures 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. **SEG** always takes values in the interval [0,1], with 1 meaning total overlap (congruency) and 0 meaning that not even one foreground pixel or voxel was common to both. The **tracking accuracy measure** (**TRA**) evaluates the accuracy of the tracking results of each computed solution. It is a weighted

[c] See also http://www.codesolorzano.com/Challenges/CTC/Challenge_Participants.html for an updated list of Challenge participants

[d] File formats and conventions can be found on the CTC website (http://ctc2015.gryf.fi.muni.cz/Public/Documents/Naming%20and%20file%20content%20conventions.pdf)

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. **TRA** is normalized in order to take values between 0 and 1, where higher values stand again for fewer errors with respect to the reference solution. For ranking the algorithms, the **overall performance** (**OP**) is computed by averaging **SEG** and **TRA** values for each pair of competition movies, and then averaging these averages, i.e. **OP** = (**SEG_{avg}** + **TRA_{avg}**)/2. 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 the state-of-the-art. **Supplementary Material: Table 3** contains a link to the evaluation software used in the challenge.

Biologists however, when using tracking algorithms, have specific biological questions 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. The **Complete Tracks** (**CT**) focuses on the fraction of ground truth cell tracks that a given method is capable to reconstruct in their entirety. The higher **CT** is, the larger is the fraction of cells that is correctly tracked throughout the entire movie, 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. The **Track Fractions** (**TF**) selects for each reference track its longest matching algorithm-generated tracklet (continuous cell tracking subsequence), computes the percentage of overlap of these subsequences with respect to the full tracks, and takes the average of these values. Intuitively, this can be interpreted as the fraction of an average cell's trajectory that an algorithm reconstructs correctly, and therefore gives an indication of the algorithm's ability to measure cell speeds or trajectories. In cases where the reliable detection of dividing cells is critical, **Branching Correctness** (**BC**) measures how efficient a method is at correctly detecting division events. Finally, the **Cell Cycle Accuracy** (**CCA**) measures how accurate an algorithm is at correctly reconstructing the length of the life of a cell, i.e., the time between two consecutive divisions. Both **BC** and **CCA** are informative about the ability of the algorithm to detect cell population growth. All biologically inspired measures take values in the interval [0,1], with higher values corresponding to better performance.

The third set of measurable quantities that we report 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 set. This does not include parameters visible only to developers. Instead, it is concerned specifically with the ones that are entered by the user. In general, lower number of tunable parameters signifies 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 data, i.e. other movies acquired under otherwise unchanged imaging conditions. **GP** values are computed by comparing the results for a particular training and competition movie, obtained using the same parameter configuration. This measure takes values in the interval [0,1], with higher values corresponding to better generalizability, leading to better applicability of the

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

Analysis of the performance of submitted algorithms.

All measures we described above have been computed for every dataset and competing algorithm. We first evaluated the segmentation (**SEG**) and tracking (**TRA**) accuracy measures. Top-three values for each dataset are listed in **Table 4**, and the algorithms that obtained those scores are listed in **Table 5** (see **Supplementary Data 2** for the complete list of values). In order to help determining the significance of these values, we calculated **SEG** and **TRA** values with respect to the ground truth data not only for each algorithm's result, but also for the three manual annotations, as if they were submitted results, since they are the best available proxies for evaluating the variability among human annotators. Therefore, algorithms with **SEG** or **TRA** scores within the range of the average manual scores ($\mathbf{SEG_a}$ and $\mathbf{TRA_a}$) plus/minus one standard deviation 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.

We first examine the results from the viewpoint of the datasets (**Table 4**), trying to pinpoint the features that underlie the good and not so good performance of the competing methods. We observe that some algorithms reached very good values (**OP** > 0.9) for specific datasets, such as Fluo-N2DH-GOWT1, PhC-C2DH-U373, Fluo-N2DL-HeLa, Fluo-C3DH-H157, and Fluo-N3DH-CHO. In all but one of these datasets (Fluo-C3DH-H157), one or more algorithms reached human quality results, i.e. values close to or above the average annotator values $\mathbf{SEG_a}$ and $\mathbf{TRA_a}$. Interestingly, all but one of these results are obtained on fluorescence data with high *SNR* or *CR* values. Some also show high spatial (Fluo-C3DH-H157, Fluo-N3DH-CHO) and/or temporal (Fluo-N2DH-GOWT1, Fluo-N2DL-HeLa, Fluo-N3DH-CHO) resolution and display rather low cell densities (Fluo-C3DH-H157, Fluo-N2DH-GOWT1, PhC-C2DH-U373, Fluo-N3DH-CHO).

A second group of datasets 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 datasets, the **SEG** and **TRA** values are near but below the performance of the manual annotators, meaning that after automatic tracking some additional curation work is required to reach the level of the human annotators. The difficulty for DIC-C2DH-HeLa and PhC-C2DL-PSC appears to be the low *SNR* and *CR* values and high cell density, and for DIC-C2DH-HeLa also the rather complex image texture within cells (see **Supplementary Material: Figs. 1** and **11**). For Fluo-C3DL-MDA231, the low *SNR* and *CR* values are paired with low spatial and temporal resolution and significant photobleaching (see **Supplementary Material: Fig. 4**). The two synthetic datasets (Fluo-N2DH-SIM+, Fluo-N3DH-SIM+) show average *SNR*, low *CR*, average cell density, and average to high heterogeneity within and between cells.

Three datasets (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 datasets, a substantial amount of manual work would be needed to curate the computed results in order to reach

human-level annotations. Fluo-C2DL-MSC suffers mostly from low *SNR* and *CR* values, low temporal resolution, and significant photobleaching. This dataset is difficult to segment correctly also due to its prominent cell protrusions (see **Supplementary Material: Fig. 2**). For Fluo-N3DH-CE and Fluo-N3DL-DRO, the two whole embryo datasets, the algorithms mostly struggle to segment and track the very noisy cell nuclei in 3D. Additionally, these datasets show very low spatial resolution, relatively low temporal resolution, and increasingly dense cells toward the end of the movies which strongly complicate tracking of the segmented cells (see **Supplementary Material: Figs. 7** and **9**).

Next, we examine the results from the viewpoint of the algorithms, asking which ones show best overall performance. Among the algorithms that obtained the best values (**Table 5**), KTH-SE, FR-Ro-GE, and HD-Hau-GE ranked first for one or more datasets. Looking more globally at the number of top-3 occurrences in **Table 5**, the methods KTH-SE, FR-Ro-GE and HD-Har-GE outperform the others. Their common denominator is the reliance on the *tracking by detection* paradigm. In particular, KTH-SE algorithms perform extraordinarily well, being ranked among the top-three for all datasets. These methods rely on a simple thresholding segmentation highly enriched by the use of global information in the tracking process. In some datasets, however, the overall lower performing *tracking by contour evolution* methods (LEID-NL, MU-CZ, and PAST-FR) reach the level of the leading *tracking by detection* methods. This can be attributed to their high segmentation performance on datasets with high temporal and spatial resolution (Fluo-N3DH-CHO, Fluo-N2DH-GOWT1, Fluo-N2DH-SIM+, and Fluo-N3DH-SIM+). These results highlight how this class of methods relies on significant 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) datasets. Indeed, these methods obtain performance values on DIC-C2DH-HeLa, PhC-C2DH-U373, and PhC-C2DL-PSC that do not match their predicted level of complexity. This can be explained by the fact that the internal texture of the cells in these datasets is not detrimental for the segmentation. On the contrary, it seems to be helpful by improving the learning capacity of the algorithms. Finally for this part of the analysis, as shown in **Supplementary Material: Fig. 12,** the evolution of the average of the top-3 **OP** values during the three CTC editions shows clear progress towards the objective of reaching the level of the human expert annotators.

We have also studied the robustness of the **OP**–based rankings, obtained as described in the **Online Methods** (section **Ranking robustness**) and summarized in **Supplementary Material: Fig. 13**, which shows that the rankings are indeed robust for up to 45% of possible weight changes. Furthermore we have analyzed the correlation, i.e. interdependence of **SEG** and **TRA** scores using the Kendall's τ correlation coefficient (**Supplementary Material: Table 4**) to show moderate global correlation (0.55) with only a few cases of very high (**DIC-C2DH-HeLa**, **Fluo-N3DH-CE**) or high **(PhC-C2DL-PSC**, **Fluo-C2DL-MSC**) correlation.

Since segmentation and tracking are meant to answer biological questions in the hands of practicing biologists, we next analyze the biologically inspired and usability measures. **Table 6** shows the top-three biological scores: **CT** (Complete tracks), **TF** (Track fractions), **BC**

(Branching correctness) and **CCA** (Cell cycle accuracy) and the average values obtained by the annotators ($\mathbf{CT_a}$, $\mathbf{TF_a}$, $\mathbf{BC_a}$, $\mathbf{CCA_a}$). When looking at **CT** along the dataset dimension (columns), we observe very low values overall, but especially so for DIC-C2DH-HeLa, Fluo-C2DL-MSC, PhC-C2DL-PSC, and the two embryonic developmental datasets (Fluo-N3DH-CE and Fluo-N3DL-DRO). The low **CT** values are especially relevant for the embryonic datasets since tracking completeness is indeed critical for a correct genealogical reconstruction of embryo development. The **TF** values are at a higher level, meaning that the methods are reasonably competent at measuring cells 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+ show high **BC** and **CCA** values, meaning that the methods are able to correctly detect cell divisions and cell population growth, while PhC-C2DL-PSC, the *C. elegans* dataset Fluo-N3DH-CE (and presumably also *Drosophila* 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** (**Table 7**, and **Supplementary Data 3**), we see similar but not completely matching pictures compared to the ranking compiled using the **SEG** and **TRA** values (**Table 5**). This is because **TF** and **CT** are only considering tracking correctness, regardless of the accuracy of the segmentation, and have much more strict requirements on correctly reconstructed tracks (for instance three fragmented tracks overlapping a single ground truth track in multiple places would contribute towards a high **TRA** score but will cause low **TF** and zero **CT** scores). This means that solutions with a high **TRA** score but low **TF** and **CT** scores, do still contain errors that need to be fixed in order to enable sound biological conclusions. The KTH-SE algorithms remain the top-ranked ones in most datasets, highlighting the importance of the inclusion of global information in the linking process, which yields longer, correctly reconstructed tracklets. However, similarly to the above-discussed **SEG** and **TRA** scores, the *tracking by contour evolution* method LEID-NL manages to break the dominance of *tracking by detection* approaches (it is top-ranked three times for **TF** and two times for **CT**). This highlights that *tracking by contour evolution* methods can be superior at following cells (once a track is initiated) if the temporal resolution of the data permits. As a final comment, methods that inherently (KTH-SE, HD-Har-GE, IMCB-SG) or specifically (HD-Har-GE, LEID-NL) detect cell division events show 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 that, is ranked first three out of five possible times in the **CCA** category, and can therefore safely be distinguished as the best overall method when it comes to detecting complete cell cycles, and therefore, measuring cell population growth.

Finally, since competing solutions need to be deployed by biologists normally having little computer science experience, we analyzed the usability, the speed, and the general applicability of all top-ranked algorithms. From the results shown in **Table 8** (see **Supplementary Data 4** for a complete list), when focusing on the two best methods, we can see that the superior performance of the KTH-SE algorithms comes, unfortunately, with the disadvantage of an elevated number of parameters compared to most other methods (in particular the close contender FR-Ro-GE). Conversely, the KTH-SE algorithms are faster than most other methods including FR-Ro-GE (for which, however, a much faster implementation

using graphics cards exists). Finally, we see that the KTH-SE methods generalize 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 for previously unseen data of the same kind.

Discussion

We have presented the results of three editions of the Cell Tracking Challenge, a benchmarking effort aimed at improving cell tracking in multidimensional 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. The most important contribution of our work is the compilation of expert-driven annotations of cell regions and trajectories in these videos. We also include artificially generated 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 selection of competing solutions that represent all main algorithmic approaches to cell segmentation and tracking problems in biology. Fourth, in this report we analyze the accumulated data and provide useful guidelines for both users and developers of tracking software.

From the comparison of the competing algorithms, we can conclude that in most practical scenarios *tracking by detection* methods outperform *tracking by contour evolution* methods. A notable exception to this can be observed in datasets with high temporal resolutions that have significant inter-frame 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 perform 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 tend to produce better segmentation results than other methods. Furthermore, *tracking by detection* methods that consider larger spatiotemporal contexts to reason about track linking tend 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]. In this study, we prove

that to be true also in real datasets of moving cells with non-linear 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 post-processing is acceptable, the top-scoring algorithms offer good performance. However, due to 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/available. The good news is that once parameters are 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 due to the combinatorial explosion of colliding factors (biological, imaging, algorithmic) that affect the results of segmentation and tracking, there is no simple way to point out the right algorithm for a given dataset. 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 datasets, additional method development is crucially required for scenarios with low *SNR* or *CR* or for tracking cells with more complex shapes or textures. Large 3D datasets, such as those of developing embryos, bear additional challenges. Not only do such movies show very high cell densities at later frames, the size of the image data itself causes very long runtimes. *Tracking by detection approaches* fail on these datasets because they crucially depend on high quality segmentation results, something difficult in these challenging datasets. *Tracking by contour evolution approaches* often fail on them due to 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 **SEG** and **TRA** when calculating the overall performance of the methods, we assign equal importance to both tasks although, as we show, 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 datasets where the performance of a tracking solution seems to be heavily influenced by a the performance of segmentation approach.

It is important to stress that, although the challenge was broadly taken on by the community and many algorithms competed, the voluntary nature of participation necessarily resulted in significant omissions. This affected, in particular, the submissions attempting to meaningfully solve the 3D tracking problems in embryos that are the most challenging datasets and for which potent methods are published and available[32,33]. This situation was made worse by the lack of complete ground truth for these massive datasets.

The Cell Tracking Challenge, which remains open for online submissions, is a powerful resource for algorithm developers and users alike. While additional submissions are currently

handled manually, we plan to automate this in the near future. New datasets of existing and new microscopy modalities will over time be incorporated to the dataset repository. It will be particularly important to collect and annotate complex tissue, organ, and whole embryo data. We offer the evaluation framework, capable of computing all measures we have introduced, as an open-source Fiji plugin[35], and provide executable versions of the participants' algorithms. Furthermore, will encourage past and future participants to make their submitted algorithms available to biologists via easy to install and intuitive graphical user interfaces. Finally, we are planning to add new synthetic datasets that closely mimic the variety of cell types and microscopy scenarios. These synthetic data will model different cell labeling, cell shapes, and cell behaviors and migration patterns in 2D and 3D. Since artificially generated datasets implicitly bear absolute ground truth, they can be tuned to challenge algorithms to improve specific aspects of the problem (i.e. 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.

Acknowledgments

We would like to acknowledge the following funding sources: The Spanish Ministry of Economy MINECO grants DPI201238090-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., E.M.) and 639.021.128 (I.S.); Dutch Technology Foundation (STW) grant 10443 (I.S., E.M.); Czech Science Foundation (GACR) grant P302/12/G157 (M.K., Pa.M.); Helmholtz Association (J.S., R.M.), DFG grant MI 1315/4-1 (J.S., 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., L.M.); the BMBF, projects ENGINE (NGFN+) and RNA-Code (e:Bio), and the DFG within the SFB 1129 (N.H., 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, the CellNetworks Excellence Cluster / EcTop (C.H., S.W., F.H.); the Baxter Foundation and NIH grant AG020961 (H.M.B.), the Swedish Research Council VR Grant 2015-04026 (K.M., J.J.); the BMBF, project de.NBI, grant 031L0102 (V.U., F.J.).

We acknowledge the work of those who manually annotated the datasets to create the ground truths used to evaluate the performance of the algorithms: A. Urbiola, C. Ederra, T. España, S. Venkatesan, D.M.W. Balak, P. Karas, T. Bolcková, M. Štreitová, M. Charousová, L. Zátopková.

We also would like to thank those who provided the datasets used in the three challenge editions: Dr. F. Prósper, Dr. E. Bártová, Dr. J. Essers, the Mitocheck consortium, Dr. A. Rouzaut, Dr. R. Kamm, the Waterston Lab, Dr. P. Keller, Dr. S. Kumar, Dr. G. van Cappellen, and Dr. T. Becker.

Finally, we thank R. Stoklasa for technical support. The participants would like to acknowledge the contributions of participants not listed among the authors: M. Schiegg, D. Stöckel, J. Crowe, M. Temerinac-Ott.

Author's contributions

Vladimír Ulman: Actively participated in the organization and management of the CTC challenges by handling submissions, producing synthetic datasets, evaluating the submitted results and globally analyzing the participant's contributions, created annotations for dataset evaluation. Contributed to the writing of the manuscript and produced the tables and plot results.

Martin Maška: Actively participated in the organization and management of the CTC challenges: handled and evaluated submissions, provided evaluation and annotation software, supervised annotations, created consensual ground truths for the evaluation of the submitted results. Contributed to the writing of the manuscript. Challenge participant.

Klas E. G. Magnusson: Top ranked challenge participant. Contributed to the writing of the manuscript.

Olaf Ronneberger: Top ranked challenge participant. Contributed to the writing of the manuscript.

Carsten Haubold: Top ranked challenge participant. Contributed to the writing of the manuscript.

Nathalie Harder: Top ranked challenge participant.

Pavel Matula: Actively participated in the organization of the CTC challenges: Led the development of a suitable tracking measure and assessed the behavior of various measures on challenge datasets.

Petr Matula: Actively participated in the organization of the CTC challenges: prepared data and supervised data annotation.

David Svoboda: Actively participated in the organization of the CTC challenges: Led the development of synthetic data generator and creation of suitable collection of synthetic time-lapse sequences with absolute ground truth.

Miroslav Radojevic: Actively participated in the organization of the CTC challenges: prepared data and supervised data annotation.

Ihor Smal: Actively participated in the organization of the CTC challenges: prepared data and supervised data annotation.

Karl Rohr: Challenge participant.

Joakim Jaldén: Challenge participant.

Helen M. Blau: Challenge participant.

Oleh Dzyubachyk: Challenge participant.

Boudewijn Lelieveldt: Challenge participant.

Pengdong Xiao: : Challenge participant.

Yuexiang Li: Challenge participant.

Siu-Yeung Cho: Challenge participant.

Alexandre Dufour: Challenge participant.

Jean Christophe Olivo-Marin: Challenge participant.

Constantino C. Reyes-Aldasoro: Challenge participant.

Jose A. Solis-Lemus: Challenge participant.

Robert Bensch: Challenge participant.

Thomas Brox: Challenge participant.

Johannes Stegmaier: Challenge participant.
Ralf Mikut: Challenge participant.
Steffen Wolf: Challenge participant.
Fred. A. Hamprecht: Challenge participant.
Tiago José Arieria Esteves: Challenge participant.
Pedro Quelhas: Challenge participant.
Ömer Demirel: Challenge participant.
Lars Malmström: Challenge participant.
Florian Jug: Contributed to the revision of the manuscript and supported V.U. with the related data processing.
Pavel Tomančák: Challenge organizer. Contributed to the revision of the manuscript.
Erik Meijering: Challenge organizer. Contributed to the writing of the manuscript.
Arrate Muñoz-Barrutia: Challenge organizer. Contributed to the writing of the manuscript.
Michal Kozubek: Challenge organizer. Contributed to the writing of the manuscript.
Carlos Ortiz-de-Solorzano: Challenge organizer. Coordinated the work of the committee that organized the challenges. Wrote the manuscript with the input from all authors.

Competing financial interests

All the authors declare not to have competing financial interests.

References

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. Microscopy imaging techniques for drug discovery. *Nat. Rev. Drug Discov.* **7**, 54-67 (2007).
3. Walter, R.J., Berns, M.W. Digital image processing and analysis, in Video Microscopy. (S. Inoué Ed.) Springer Sciences, p. 327-392 (1986)
4. Schneider, C.A., Rasband, W.S. & Elicieri, 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 Proc. Mag.* **29**, 140-145 (2012)
6. Dufour, A.C, Liu, T-Y., Ducroz, C., Tournemenne, R., Cummings, B., Thibeaux, R., Guillen, N., Hero, A.O., Olivo-Marin, J.C. Signal processing challenges in quantitative 3-D cell morphology: More than meets the eye. *IEEE Signal Proc. Mag* **32**, 30-40 (2015).
7. Zimmer, C., Zhang, B., Dufour, A., Thebaud, A., Berlemont S., Meas-Yedid, J., Olivo-Marin, J.C. On the digital trail of mobile cells. *IEEE Signal Proc. Mag.* **23**, 54-62 (2006)
8. Wuttisarnwattana, P., Gargesha, M., vant't Hof. W., Cooke, K.R., Wilson, D.L. Automatic stem cell detection in microscopic whole mouse cryo-imaging. *IEEE Trans. Med. Imag.* **35**, 819-829, (2016)
9. Lerner, B., Clocksin, W.F., Dhanjal, S., Hultén, S., 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., Pase, L., Ramos-Lopez, C.F., Lieschke, G.J., Renshaw, S.A., Reyes-Aldasoro, C.C. 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., Elandsson, F., Borgefors, G., Bengtsson, E. Combining intensity, edge and shape information for 2D and 3D segmentation of cell nuclei in tissue sections. J. Microsc-Oxford, **215**, 67-76 (2004)
13. Cicconet, M., Geiger, D., Gunsalus, K., Wavelet-based circular hough-transform and its application in embryo development analysis. VISAPP 2013, Proceedings of the International Conference on Computer Vision Theory and Applications, 669-674, (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. Imag.* **36**, 942-951, (2016)
15. Malpica, N., Ortiz-de-Solorzano, C., Vaquero, J.J., Santos, A., Vallcorba, I., Garcia-Sagredo, J.M., Pozo, F. Applying watershed algorithms to the segmentation of clustered nuclei. Cytometry Part A. **28**, 289-297 (1997)

16. Ortiz-de-Solorzano, C., García-Rodríguez, E., Jones, A., Pinkel, D., Gray, J.W., Sudar, D., Lockett, S.J. Segmentation of confocal microscopy images of cell nuclei in thick tissue sections. J. Microsc-Oxford, **193**, 212-226 (1999)

17. Cliffe, A., Doupé, D.P., Sung, H., Lim, I.K.H., Ong, K.H., Cheng, L., Yu, W. 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 LCNS* **9351**, 234-241, (2015).

19. Schiegg, M., Hanslovsky, P., Haubold, C., Koethe, U., Hufnagel, L., Hampretch, F.A., Graphical model for joint segmentation and tracking of multiple dividing cells. *Bioinformatics* **31**, 948-56 (2015)

20. Zimmer, C., Labruyere, E., Meas-Yedid, V., Guillen, 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. Imag.* **21**, 1212-1221, (2002)

21. Dufour, A., Thibeaux, R., Labruyere, E., Guillen, N., Olivo-Marin, J.C. 3D active meshes: fast discrete deformable models for cell tracking in 3D time-lapse microscopy. *IEEE Trans. Image Process.* **20**, 1925–37 (2011).

22. Maška, M., Daněk. O., Garasa, S., Rouzaut, A., Muñoz-Barrutia, A., Ortiz-de-Solorzano, C., Segmentation and shape tracking of whole fluorescent cells based on the Chan-Vese model. *IEEE Trans. Med. Imag.* **32**, 995-1005, (2013)

23. Ortiz-de-Solorzano, C., Malladi, R., Lelievre, S.A., Lockett, S.J. Segmentation of nuclei and cells using membrane related protein markers. J. Microsc-Oxford, **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. Imag.* **29**, 852-867, (2010)

25. Dufour, A., Shinin, V., Tajbakhsh, S., Guillen-Aghion, N., Olivo-Marin, J.C., Zimmer, C. Segmenting and tracking fluorescent cells in dynamic 3D microscopy with coupled active surfaces. *IEEE Trans. Image Process.* **14**, 1396–1410 (2005)

26. Bensch, R., and 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., Mora-Bermúdez, F., Godinez, W.J., Wünsche, A., Elis, R., Ellenberg, J., Rohr, K. 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. Imag.* **34**, 1–19 (2015).

30. Maška, M., Ulman, V., Svoboda, D., Matula, Pt., Matula, Pv., Ederra, C. Urbiola, A., España, T., Venkatesan, S., Balak, D.M.W., Karas, P., Bolcková, T., Štreitová, M., Carthel, C., Coraluppi, S., Harder, N., Rohr, K., Magnusson, K.E.G., Jaldén, J., Blau, H.M., Dzyubachyk, O., Křížek, P., Hagen, G.M., Pastor-Escuredo, D., Jimenez-Carretero, D., Ledesma-Carbayo, M.J., Muñoz-Barrutia, A., Meijering, E., Kozubek, M.

& Ortiz-de-Solorzano, C. 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., Bao, Z., Boule, T.J., Boeck, M.E., Mericle, B.L., Nicholas, T.J., Zhao, Z., Sandel, M.J., Waterston, R.H. Automated analysis of embryonic gene expression with cellular resolution in C. elegans. *Nat. Methods.* **5**, 703-9 (2008)

33. Amat, F., Lemon, W., Mossing, D.P., McDole, K., Wan, Y., Branson, K., Myers, E.W., Keller, P.J. Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nat. Methods.* **11**, 951-8 (2014)

34. Chenouard, N, Smal, I., de Chaumont, F., Maška, M., Sbalzarini, I.F, Gong, Y. *et al*. Objective comparison of particle tracking methods. *Nat. Methods* **11**, 281-289 (2014)

35. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Rueden, C., Saafeld, S., Schmid, B., Tinevez, J.Y., White, D.J., Hartenstein, V., Elliceri, K., Tomancak, P., Cardona, A. Fiji: an open source platform for biological-image analysis. *Nat. Methods* **9**, 676-82 (2012)

Online Methods

Dataset quality parameters

In order to assess the quantitative video parameters (see **Table 1**), we had to calculate those parameters –ideally- on a complete ground truth of the competition datasets, 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 consists of 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 tracking point in **TRA-GT**, we automatically merged the top-performing participants' segmentation masks that included this tracking point. The number of masks used was determined manually for each video. On average, majority of the total number of available masks were used. The process led occasionally to colliding situations, that is, when obtained segmentation masks for two different tracking points 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 datasets 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 datasets was established also like this. For Fluo-N3DH-CE and Fluo-N3DL-DRO datasets, 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 $\mathbf{FG_{i,t}}$ and $\mathbf{BG_t}$ 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 **std(S)** denote average and standard deviation 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 *SNR*, *CR*, Het_i, *Res*, *Sha*, *Den*, and *Ove*

parameters were established as averages of $\boldsymbol{SNR}_{i,t}$, $\boldsymbol{CR}_{i,t}$, $\mathbf{HETi}_{i,t}$, $\mathbf{Res}_{i,t}$, $\mathbf{Sha}_{i,t}$, $\mathbf{Den}_{i,t}$,and $\mathbf{Ove}_{i,t}$ values, respectively, calculated for every object in every image in both competition videos:

$$SNR_{i,t} = \frac{\left|avg(FG_{i,t}) - avg(BG_t)\right|}{std(BG_t)}$$

$$CR_{i,t} = \frac{avg(FG_{i,t})}{avg(BG_t)}$$

$$HETi_{i,t} = \frac{std(FG_{i,t})}{\left|avg(FG_{i,t}) - avg(BG_{t)}\right|}$$

$$HETb_{i,t} = \frac{\left|avg(FG_{i,t}) - avg(BG_t)\right|}{\sum_{j \in I(t)} \left|avg(FG_{j,t}) - avg(BG_t)\right| / |I(t)|}$$

$$Res_{i,t} = \left|FG_{i,t}\right|$$

$$Den_{i,t} = min\{50, dist(a,b) \;\|\; a \in FG_{i,t}\,, b \in FG_{j,t}, j \in I(t), j \neq i\}$$

$$Ove_{i,t} = \frac{\left|\{a \in FG_{i,t} \;\|\; \exists b \in FG_{i,t-1} : dist(a,b) = 0\}\right|}{\left|FG_{i,t}\right|}$$

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 Het_b is calculated as the standard deviation of $\mathbf{HETb}_{i,t}$ values for every object in every image in both competition videos. $\mathbf{Sha}_{i,t}$ 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 $\mathbf{Den}_{i,t}$ 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 difference between average object intensity at the end and the beginning of a video was divided by the number of images comprising this video, and *Cha* reported for a dataset is the average over both videos. *Mit* is the average of $\mathbf{Mit}_t$ taken over images from both videos, where $\mathbf{Mit}_t$ is the number of objects whose tracks end in the **t**-th image because of subsequent division event (which is marked in the tracking ground truth **TRA-GT**). The remaining qualitative parameters, *Syn*, *Ent*/*Leav*, *Apo*, and *Deb* were set after manual inspection of the datasets.

Description of datasets

DIC-C2DH-HeLa (**Supplementary Video 1, Fig. 4a, Supplementary Material: Fig. 1**): HeLa cells on a flat glass substrate. The uses of this cell line and setup are similar to Fluo-N2DL-Hela, with the physiological advantages of transmission microscopy over fluorescence described for PhC-C2DH-U373.
This dataset presents low *SNR* and *CR* values characteristic of phase-enhancement microscopy techniques. Het_i and Het_b are high due to the presence of DIC-highlighted internal structures and organelles (Het_i), and the fact that in most frames co-exist well spread interphase cells with brighter, rounded shaped cells undergoing mitosis (Het_b). Finally, another relevant problem of this dataset is the high density of the cells, which are highly clustered, occupy the majority of the image area and barely show intensity changes between neighboring cells.
These videos are courtesy of Dr. Gert van Cappellen, Optical Imaging Center, Department of Pathology, Erasmus University Medical Center, Rotterdam (the Netherlands).

Fluo-C2DL-MSC (**Supplementary Video 2**, **Fig. 4b**, **Supplementary Material: Fig. 2**): Rat mesenchymal stem cells (MSCs) on a flat polyacrylamide substrate stained by stable transfection with Actin-GFP. Mesenchymal stem cells are non-hematopoietic cells located in the bone marrow that can differentiate into several cell types such as osteoblasts, adipocytes and hematopoietic-supporting stromal cells. The *ex vivo* expansion and *in vivo* differentiation of these cells is of high therapeutic value, and has been used in both cell and tissue engineering therapies to treat acute graft-versus-host disease after allogeneic hematopoietic stem cell transplant, promote heart recovery after ischemic heart disease and congestive heart failure, treat cirrhosis, hepatitis and other liver diseases, etc. In this context, the study of the migratory properties of MSCs is relevant since it is indeed related to their ability to access sites of inflammation and their homing and engrafting properties. Especially relevant is the dependence of MSCs migration on the biomechanical properties of their substrate, which has a quantifiable effect both on the morphology of the cells (segmentation) and the dynamics of their migration patterns (tracking).
The *SNR* and *CR* values are low, due to the low level of emission of the fluorescent cytoplasmic reporter, especially in the long, thin filopodial extensions of the cell. The intensity is also quite variable in different parts of the cell, causing high Het_i. The cells present different levels of intensity, possibly due to different levels of expression of the transfected fluorescent reporter, thus producing high Het_b values. The shape of the cells is highly irregular (*Sha*) due to the long filopodial extensions, show a significant degree of bleaching (*Cha*) and move fast, causing low overlap (*Ove*) of the cells between consecutive frames.
These videos are courtesy of Dr. F. Prósper, Center for Applied Medical Research, Pamplona (Spain).

Fluo-C3DH-H157 (**Supplementary video 3**, **Fig. 4c**, **Supplementary Material: Fig. 3**): GFP-transfected H157 lung cancer cells embedded in a 3D Matrigel matrix. H157 is a non-small cell lung cancer cell line. These cells are highly metastatic, thus providing a good benchmark for the study of cell migration. Tracking, in this context, provides a great deal of information regarding the morphological changes that the cells suffer during migration and the dynamics of migration

itself as a response to different molecular stimuli. It also gives information about the relationship between the morpho-mechanic properties of the extracellular environment and cell migration. These cells are key for pharmacological studies aimed at blocking migration by interfering with the mechanosensing and mechanotransducing properties of the cells which, in turn, produce changes in cell morphology and migration dynamics.
This dataset displays reasonably good values for most properties, with the exception of some signal decay due to photobleaching (*Cha*). This is a negative side-effect of the high-resolution of the images, which requires a dense optical sectioning, and therefore high acquisition times and thus lenghty exposures of the fluorochrome to the light. The presence of prominent blebs, and some heterogeneity between cell intensities can also complicate accurately segmenting and delineating the cell boundaries.
The videos are courtesy of Dr. A. Rouzaut, Cell Adhesion and Metastasis Laboratory, Center for Applied Medical Research, Pamplona (Spain).

Fluo-C3DL-MDA231 (**Supplementary Video 4**, **Fig. 4d**, **Supplementary Material: Fig. 4**): MDA231 human breast carcinoma cells infected with a plasmic murine stem cell virus (pMSCV) vector including the GFP sequence, embedded in a 3D collagen matrix. These cells are also metastatic, and their uses in the context of cell tracking are similar to the ones described for Fluo-C3DH-H157.
The *SNR* and *CR* of this dataset are relatively low due both to low signal intensity and increased background which affects the quality of the signal especially in the long migration-related filopodial extensions. This noisy signal efficiency causes high internal heterogeneity (Het_i.). To complicate the segmentation and tracking even further, the images are acquired at low resolution, especially in the axial direction (*Res*), and also in the temporal dimension (*Ove*) and suffer from significant photobleaching (*Cha*).
These videos are courtesy of Dr. Roger Kamm, Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA (USA).

Fluo-N2DH-GOWT1 (**Supplementary Video 5**, **Fig. 4e**, **Supplementary Material: Fig. 5**): Multipotent mouse embryonic stem cell nuclei, chromatin-stained by stable transfection with histone H2B-GFP. Embryonic stem cells are used in many areas of research, most notably in the study and application of cell differentiation, with strong therapeutic potential particularly for neural regeneration, cardiology, and hemato-oncology. These cells do not have a clear motile phenotype, but segmentation and tracking is still of interest to give a spatial frame to intracellular molecular trafficking events, to detect and quantify cell division as part of the differentiation process and to capture the dynamics of tissue or organogenesis.
This dataset presents average to good values in all properties, except the internal heterogeneity of the nuclear signal due to the existence of prominent, unlabeled nucleoli (Het_i) and the heterogeneity of the average cell intensities (Het_b) due possibly to different levels of efficiency of the transfected reporter.
These videos are courtesy of Dr. E. Bártová, Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno (Czech Republic).

Fluo-N2DL-Hela (**Supplementary Video 6**, **Fig. 4f**, **Supplementary Material: Fig. 6**): H2B-GFP stably transfected HeLa cells, the oldest immortalized human cell line, obtained from a cervical cancer. They have been used innumerable times in biological research. Cells, as in the previous two cases, are not properly motile, and tracking is mainly a tool for nuclear delineation and detection of mitotic expansion.
This dataset displays average or good values for most properties, except for *CR*, which is alleviated by a high *SNR*, provided by the high dynamic range of the detector used. Also poor are the values corresponding to the signal heterogeneity between cells (Het_b), the spatial resolution (*Res*), cell density (*Den*) and the presence of division events (*Mit*).
These videos were kindly provided by the Mitocheck Consortium (http://www.mitocheck.org).

Fluo-N3DH-CE (**Supplementary Video 7**, **Fig. 4g**, **Supplementary Material: Fig. 7**): Early stage *C. elegans* developing embryo with nuclei stained by GFP transfection. This nematode is the simplest and most commonly used model for the study of the genetic expression and regulatory networks that control embryonic development. It is also widely applied to study other cellular processes such as cell-to-cell communication and wound healing. In this context, automatic cell tracking can simplify the process of quantifying migration capacity and building cell lineages. With this dataset we wanted to see the ability of algorithms to keep track of significantly increasing population of cells, where divisions events are equally important to nuclei tracking in order to construct proper lineages.
The most significant problems of this dataset are high cell density (*Den*) low cell overlap between frames (*Ove*) caused by large temporal acquisition step and the abundance of mitotic cells typical of a developing embryo (*Mit*). This is aggravated by average to low values in most other categories, which turn this dataset into one of the most challenging ones provided by the challenge.
These videos are courtesy of the Waterston Lab, The George Washington University, Washington DC (USA).

Fluo-N3DH-CHO (**Supplementary Video 8**, **Fig. 4h**, **Supplementary Material: Fig. 8**): Chinese Hamster Ovarian (CHO) cell nuclei, chromatin-stained by transfection with PCNA-GFP. A well-established cell line derived from the ovaries of Chinese hamsters. It constitutes a commonly used mammalian cell model in biomedical research. In addition, they are frequently used to manufacture therapeutic recombinant proteins. As in the case of Fluo-N2DH-GOWT1, these cells do no have a motile phenotype. The emphasis of tracking is also in accurate nuclear segmentation and detection of mitotic events.
All the property values of this dataset are high, rendering this dataset one of the least challenging ones. Only two are just average: the internal heterogeneity of the staining (Het_i), clearly visible in the images and caused by the fact that the nuclear staining does not label the nucleoli of the cells, and the relatively high cell density (*Den*).
These videos are courtesy of Dr. J. Essers, Departments of Genetics, Vascular Surgery, and Radiation Oncology, Erasmus University Medical Center, Rotterdam (the Netherlands).

Fluo-N3DL-DRO (**Supplementary Video 9**, **Fig. 4i**, **Supplementary Material: Fig. 9**): Developing *Drosophila melanogaster* embryo. The uses of this model are similar to the ones

described for Fluo-N3DH-CE, but with one additional level of information and complexity due to the higher developmental level of the imaged animal.
This is the most challenging dataset provided, due to low *SNR*, low spatial (*Res*) and temporal (*Ove*) resolution characteristic of SPIM, and the presence of frequent mitosis typical of a developing embryo.
These videos are courtesy of Dr. Philipp Keller, Howard Hughes Medical Institute, Janelia Research Campus, Ashburn VA (USA).

PhC-C2DH-U373 (**Supplementary Video 10**, **Fig. 4j**, **Supplementary Material: Fig. 10**): Glioblastoma-astrocytoma U373 cells on a polyacrylamide 2D substrate. This cell line and setup are instrumental to study the morphology and migration of cancer cells, related to the biomechanical properties of their substrate or the presence of chemotactic factors or interfering drugs, through a quantitative look at cell morphology and migration patterns. In contrast to its fluorescence counterpart, imaging with phase contrast simplifies the preparation of the experiments, eliminates the uncertainty regarding the effect of vector-GFP transfection in the migration phenotype and reduces toxicity caused by the fluorescent excitation.
As explained for **DIC-C2DH-Hela**, at this level of resolution the *SNR*, *CR*, Het_i and Het_b are deficient, as expected for a contrast enhancement microscopy modality. All other values are either average or good, which seems to compensate the deficient values for the segmentation and tracking task. Especially beneficial seems to be a high spatial (*Res*) and temporal (*Ove*) resolution, and a relatively low cells density (*Den*).
These videos are courtesy of Dr. Sanjay Kumar, Department of Bioengineering, University of California at Berkeley, Berkeley CA (USA).

PhC-C2DL-PSC (**Supplementary Video 11**, **Fig. 4k**, **Supplementary Material: Fig. 11**): Pancreatic stem cells on a flat polystyrene substrate. These cells allow similar uses as Fluo-N2DH-GOWT1. As in the previous two cases, phase enhancement instead of fluorescence imaging adds to the simplicity of the microscopy setup and physiological value of the results, at the expense of a more complex analysis.
Most of the parameters are in the average to low range, especially those already mention for brightfield modalities, and the very low spatial resolution (*Res*) -to some extent compensated by a high temporal resolution (*Ove*) – and significant number of mitotic events (*Mit*)
These datasets are courtesy of Dr. Tim Becker, Fraunhofer Institution for Marine Biology, Lübeck (Germany).

Fluo-N2DH-SIM+ and Fluo-N3DH-SIM+ (**Supplementary Videos 12-13**, **Fig. 4l**): While the real datasets aimed at discovering the performance of algorithms in real biological situations, the simulated datasets allow us to evaluate their full performance due to the existence of an absolute ground truth which is not available in real samples. The datasets display nuclear dynamics throughout the complete cell cycle, including cell division, of loosely synchronized motile cell populations. The cells were largely inspired by HL60 and HeLa nuclei stained with Hoechst 33342 dye. The simulations included the effects of photobleaching and uneven illumination. They also take into account the impulse response measured in a real optical system Zeiss Axiovert 100S, equipped with a Yokogawa CSU-10 confocal unit, and added

artifacts that occur during image acquisition such as dark current noise, photon-shot noise and readout noise based on the Micromax 1300-YHS camera documentation. The simulated datasets owe to the limits of underlying simulation model. As a consequence, they do not feature much of what can be called natural variability of cells. This is exemplified with the absence of apoptosis, relatively smooth texture of the nuclei during interphase (low Het_i), and regular nuclei shape (high *Sha*). Also, imaging artifacts, such as debris, are absent. The datasets introduce segmentation difficulties mainly by means of rendering nuclei at low intensities (very low *CR*) in the presence of noise (average to low *SNR*), by medium clustering of cells (average *Den*) and strong effect of the photobleaching (average *Cha*). Note that the datasets display increased number of cell divisions, which typically spans across 5 frames and renders nuclei at increased intensities, this altogether increases the variability between nuclei (average to high Het_b). The tracking abilities are examined by maintaining average, compared to real datasets, values of minimal inter-nuclei distance (*Den*), nuclei speed versus temporal sampling (which is represented with the degree of overlap, *Ove*), and amount of simultaneous divisions in an image (*Mit*).

Simulation system used and its parameters.

The simulator used to produce all simulated videos has been thoroughly described elsewhere[31]. However we now describe its main features to facilitate the explanation of its use in the CTCs. The simulator was designed to handle artificial populations of chromatin-stained cells that develop according to known relevant principles of cell biology, throughout a complete cell cycle including mitosis. The cells autonomously exhibit movement and avoid collisions. Technically, it is a suite of computer programs that implement the simulation in two main stages. First, a sequence of digital *phantom* images with associated ground truth labeled masks and lineage file is created. During this stage, the simulation of the cell population is carried on rendering the stained nuclei into the phantom images. Second, the sequence of phantom images is artificially acquired using a virtual microscope and virtual digital camera simulators that, producing a sequence of images containing the final renderings of the nuclei. These sequences of images are the ones used in the Cell Tracking Challenge. The simulator works inherently in three spatial dimensions and produces always 3D time-lapse images. The 2D datasets were obtained by choosing 2—3 planes and taking maximum projection from them. The images were consequently cropped to introduce events of leaving and entering cells, and ground-truth data were automatically curated as a part of the simulator functionalities. Note that an updated version simulator used is freely available.

Computational models of several biological structures are used during the simulation. Namely, the virtual cell comprises of models of cell membrane, nuclear membrane, nucleoli, chromatin strands and centrosomes as well as models for movement and shape changes of the cell body. These models develop simultaneously over the course of simulation while interacting with each other. For example, nuclear and cellular membranes are modeled as surface voxels of underlying nucleus and cell body masks, and it is established (and computationally assured) that nucleus mask is always fully included in the cell body mask. Another example is the representation of a single chromatin strand as a chain of molecules that, for instance, double during S-phase or that move towards a common center during Prophase, thus visually

mimicking condensation of chromosomes. Besides, the modeled chromatin molecules are assured to always reside inside a nucleus and outside nucleoli as yet another example of the mutual interaction between the models. These models are focused predominantly only on phenomena of the cell cycle that are visible using fluorescence microscopy. In our case, the system simulates chromatin-labeled cells through 3 interphase and 5 mitotic phases.

Cells in the G1-phase, are simulated roughly half the size of a mature cell and contains one centromere. The cell body gradually grows during this phase radially outwards its center, pulling chromatin strands molecules, nucleus and remaining structures in synchrony. The subsequent S-phase replicates the chromatin chains. This is modeled as gradually duplicating the chromatin molecules until the chains double in length. Closing the interphase, the second centromere is created at the end of the G2-phase. This phase is otherwise assumed resting prior to the upcoming mitosis, and therefore the only visible changes here are due to simulation of the Brownian motion of the chromatin molecules. During mitotic phases, dramatic changes occur that affect both the cell body shape and the simulated chromatin. In the Prophase, the first phase of the mitosis, all chromatin strands collapse into small regions. These regions are dragged to align at the plate in the next phase called Metaphase. Subsequently, all chromosome chains are split into two parts. During Anaphase, the split parts are displaced towards the two centrosomes accumulating the chromatin molecules into two separated bulks. Finally, the cell body is elongated pulling the chromatin bulks slightly away from each other, and two new (future daughters) nuclei membranes are formed around the two bulks. This finishes the Telophase, last but one phase of the mitosis. In the Cytokinesis, cell body mask is split into two cells of roughly half the size of a mature cell and the virtual cell heads towards the G1-phase again. At any point of the cell cycle simulation, chromatin molecules are subject to random Brownian motion. During the interphase and some mitotic phases, the cell body shape is randomly changing, yet coherently both in space and time, with occasional abrupt deformation. As for the cell motion, a certain degree of motion persistency is included in the model. Furthermore, cells are not allowed to overlap but they can touch. In the population of virtual cells, each cell is allowed to perform one simulation step only. This guarantees that each cell does develop and the whole population will seem to be developing in parallel. Detailed description of the simulator is, however, outside the scope of this paper, and can be found elsewhere[31].

The simulator features two sets of parameters. External, and thus run-time configurable, parameters of the simulator affect mostly quantifiable parameters of the underlying models such as number of chromosome strands, length of G1-phase, or mean displacement size of the simulated cells due to their movement between consecutive frames. These parameters are typically expressed in physical units, and are collected in one configuration file along with their documentation. Another external parameter is an initial input image that is in fact a labeled mask with initial geometry (defines cell shapes and positioning inside an image) of the cells that the simulation shall start with. The configuration file used to generate the competition videos is included verbatim in **Supplementary Data 5**. A sample of the initial input image is shown in **Supplementary Material: Fig 14**. Note that the initial images are identical to the first ground truth images of the obtained simulated video except that the ground truth images are smaller

due to the cropping, as explained in the first paragraph of this section. Internal parameters, and thus compile-time configurable, is the second set of parameters of the simulator. These allow choosing the type of processes that shall be used during the simulation. Here, the two main parameters to adjust are synchronization of cell divisions in the population and the imaging process itself. The former parameter triggers mechanisms in the simulator that assure that virtual cells are initiated in the same cell cycle phase and that the first round of divisions happens within a narrow temporal window. The latter parameter influences the simulated staining as well as the processing of the phantom image in the virtual microscope and camera (explained above). The simulator also internally recognizes a command with which programmer can influence, at chosen temporal point, the behavior of the cell population. This affects chosen number of cells at the periphery of the population to switch their motion model and start moving preferentially towards image boundary and bounce back. In this way, and due to the final cropping, an increased number of cells leaving and entering the simulation can be achieved. Another command exists for influencing the length of the simulation to control how many images the output sequence shall consists of.

The videos used in the Cell Tracking Challenge in the datasets were all obtained using the same single configuration file (see **Supplementary Data 5**) The videos, however, differ mainly by regulating initial cell proximity and population size (via different initial images), number of cells leaving and entering the images, and adjusting the extent to which cell divisions are synchronized. By careful combination of the latter parameters, two types of populations were devised. One type should resemble embryonic development imaging which is typical by low-to-no number of cells leaving and entering the images, relatively synchronized divisions and thus exponential increase of the displayed population size during the course of the simulation. The second type is designed otherwise and should correspond to imaging of cells dispersed on a microscopy slide, with linear growth of the number of displayed cells. The parameters used to generate the simulated videos using in the competition phase of the CTC are listed next:

Video	Population size (cells)	Video length (frames)	Number of generations	Type	Imaging parameters
2D_01	30-64	110	2-3	slide	High intensity
2D_02	13-68	138	3-4	embryonic	Low intensity
3D_01	12-67	150	3-4	embryonic	High intensity
3D_02	29-66	110	2-3	slide	Low intensity

Performance criteria.

Technical measures:

- Segmentation accuracy (**SEG**): 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 (**TRA**): 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]:

$$\mathbf{TRA} = 1 - \frac{\min(\mathrm{AOGM}, \mathrm{AOGM}_{\mathbf{0}})}{\mathrm{AOGM}_{\mathbf{0}}}$$

where $\mathrm{AOGM}_{\mathbf{0}}$ 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 (OP): For each algorithm and dataset, **SEG** and **TRA** are first averaged over the two competition videos. Then, the averaged values, $\mathbf{SEG}_{\mathbf{avg}}$ and $\mathbf{TRA}_{\mathbf{avg}}$, are also average, i.e ($\mathbf{SEG}_{\mathbf{avg}}$ + $\mathbf{TRA}_{\mathbf{avg}}$)/2 and the result is used to compile the final ranking.

Biologically inspired measures:

- **Complete Tracks** (**CT**)[39]: **CT** 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:

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

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. Note that this equation follows from expressing harmonic mean of the standard recall and precision measures.

- **Track Fractions** (**TF**): **TF** targets the longest, correctly reconstructed, continuous fraction of a reference track. The final **TF** value for a particular video is computed by averaging these fractions over all tracks.
- **Branching Correctness** (**BC(i)**)[28,29]: **BC(i)** 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., the one provided by the method and 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 dataset 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 datasets were: Fluo-N2DL-HeLa (**i**=1, corresponding to a 30-minute 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-DIM+ (**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**): **CCA** reflects the ability of an algorithm to discover true distribution of cell cycle lengths in a video. Given an annotation of the video, be it participant's result or ground truth annotation, the branching events are first discovered. A branching event is given by a mother track and at least two daughter tracks. Then, only all tracks that are a mother track in one and daughter track in another branching event are considered. Such tracks witness a development of a cell from its birth till its next division, and the length of such track, therefore, correspond to the cell cycle length of that cell. For a given video, a cumulative frequency distribution (CDF) of detected cell cycle lengths can be calculated, and can be normalized by a number of all lengths detected yielding a function CDF of cycle lengths on a range [0:1]. Given a participant result on a video and ground truth for that video, the CCA measure is defined as:

$$CCA = 1 - max_l(|CDF_r(l) - CDF_{gt}(l)|)$$

where CDF_r and CDF_{gt} are the normalized cumulative frequency distribution functions extracted from the result and ground truth, respectively. Note that this approach does not assume any specific statistical distribution of the cell cycle lengths, and that it is adopting a common approach to discovering dissimilarities between two 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.

Usability Measures:
As an additional way to compare the algorithms, we also measure practical usability of the algorithm, based on three elements:

- **Number of required tunable parameters** (**NP**): **NP** corresponds to the number of parameters that need to be provided, and possibly tuned, to obtain a reasonable result. 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**): **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:

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

where SEG_{avg}^{GP} and TRA_{avg}^{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 fall into the [0, 1] interval, with higher values corresponding to higher generalizability.
- **Execution time** (**TIM**): For each dataset, we measured 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 $a/2 \cdot \mathbf{SEG} + b/2 \cdot \mathbf{TRA}$, $a, b \in \{0, 0.001, 0.002, \ldots, 1\}$ and calculated the number of changes between each such ranking and the one compiled using **OP** (i.e., when *a* equals to *b*). **Supplementary Material: Fig 13** plots the number of changes for every combination of weights. As can be seen, 45% of the area (i.e. of possible changes) causes no more than two changes in the rankings.

Data availability

All the datasets used in the challenge (referred to in **Fig. 4**, **Supplementary Material: Figs. 1-11**, **Supplementary Material: Videos 1-13**, and described in **Table 1** and **Supplementary Material: Table 1**), along with the annotations of the training datasets are available through the challenge website: http://www.codesolorzano.com/celltrackingchallenge/Cell_Tracking_Challenge/Datasets.html. Access to the datasets is granted after free registration to the challenge.

The set of parameters used in the generation of the synthetic datasets (referred to in **Fig. 4**, **Supplementary Material: Fig. 14**, **Supplementary Material: Videos 12-13**, and described in **Table 1** and **Supplementary Material: Table 1**) are given in **Supplementary Data** file **5**.

The entire set of evaluation measures obtained and used to compare the algorithms (used to produce **Tables 4-8**, **Supplementary Material: Figs. 12-13** and **Supplementary Material: Table 4**) is provided with this article as **Supplementary Data** files **2** (SEG, TRA, OP) **3** (CT, TF, BC, CCA), and **4** (GP)

Code availability

All the code used to produced the results reported in this article is freely available through links to the CTC server in **Supplementary Materials: Table 3**, namely a version of a FIJI plugin that contains the entire evaluation suite (used to produce **Tables 4-8**, **Supplementary Material: Fig. 12-13**), and the software used evaluate the main properties of the videos (used to produce **Table 1**), along and links to binary executable versions of all participants that agreed to share their code. The parameters used by the participants to produce their submitted results are listed in **Supplementary Data 1**.

References

36. Klette, R., & Zamperoni, P. (1996). Handbook of image processing operators. Handbook of image processing operators, by Klette, Reinhard.; Zamperoni, Piero. Chichester; New York: Wiley, 1996.

37. Lin, C. L., & Miller, J. D. 3D characterization and analysis of particle shape using X-ray microtomography (XMT). Powder Technology, **154**, 61-69 (2005)

38. Matula, Pa., Maška, M., Sorokin, D.V., Matula, Pe., Ortiz-de-Solorzano, C. & Kozubek, M. Cell Tracking Accuracy Measurement Based on Comparison of Acyclic Oriented Graphs. *PLoS One* **10**, e0144959 (2015)

39. Li, K., Miller, E.D., Chen, M., Kanade, T., Weiss, L.E., Campbell, P.G. Cell population tracking and lineage construction with spatiotemporal context. *Med. Image Anal.* **12**, 546-566 (2008)

40. Brown, M.R., Summers, H.D., Rees, P., Smith, P.J., Chappell, S.C., Errington, R.J. Flow-based cytometric analysis of cell cycle via simulated cell populations. *PLoS Comput Biol* **6**, e10000741 (2010)

Figures

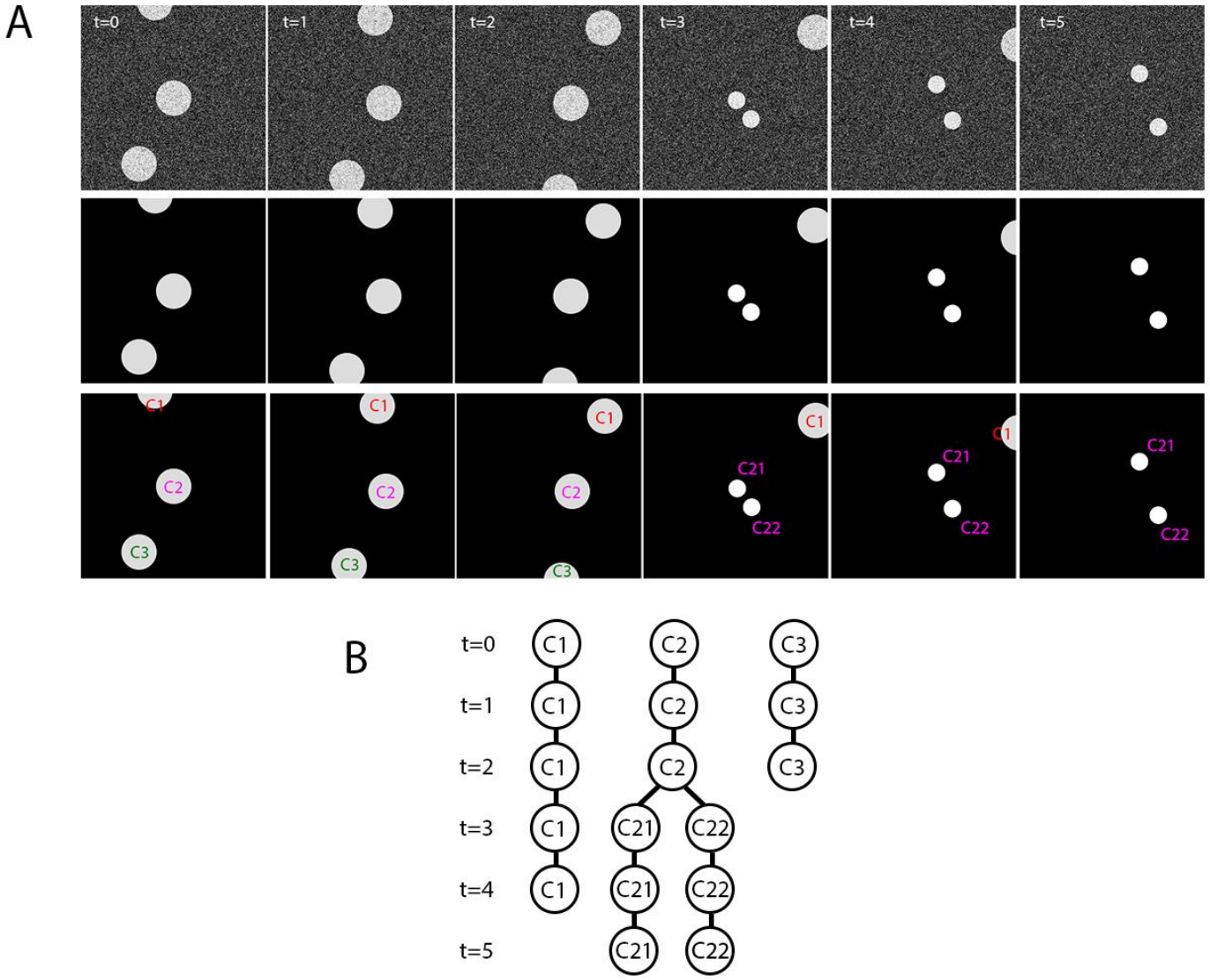

Figure 1. Concept of cell segmentation and tracking. A. *Top row*: Artificial sequence that simulates six consecutive frames of a time-lapse video. The gray circles represent cells moving on a flat surface. *Middle row*: The goal of a segmentation algorithm is to accurately determine the regions of each 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 row*: 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 being within 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 due, for instance to a cell division event (e.g., cell C2 divides into two daughter cells, C2-1 and C2-2) or apoptosis. **B**. Graph-based representation of the cell tracks found by a tracking algorithm in the sequence shown at the top of panel **A**. Such an acyclic oriented graph contains, for each cell, the time when the cells 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, all through the length of the video. Please note that the direction of the graph follows the temporal sequence starting at t=0 and moving towards t=5.

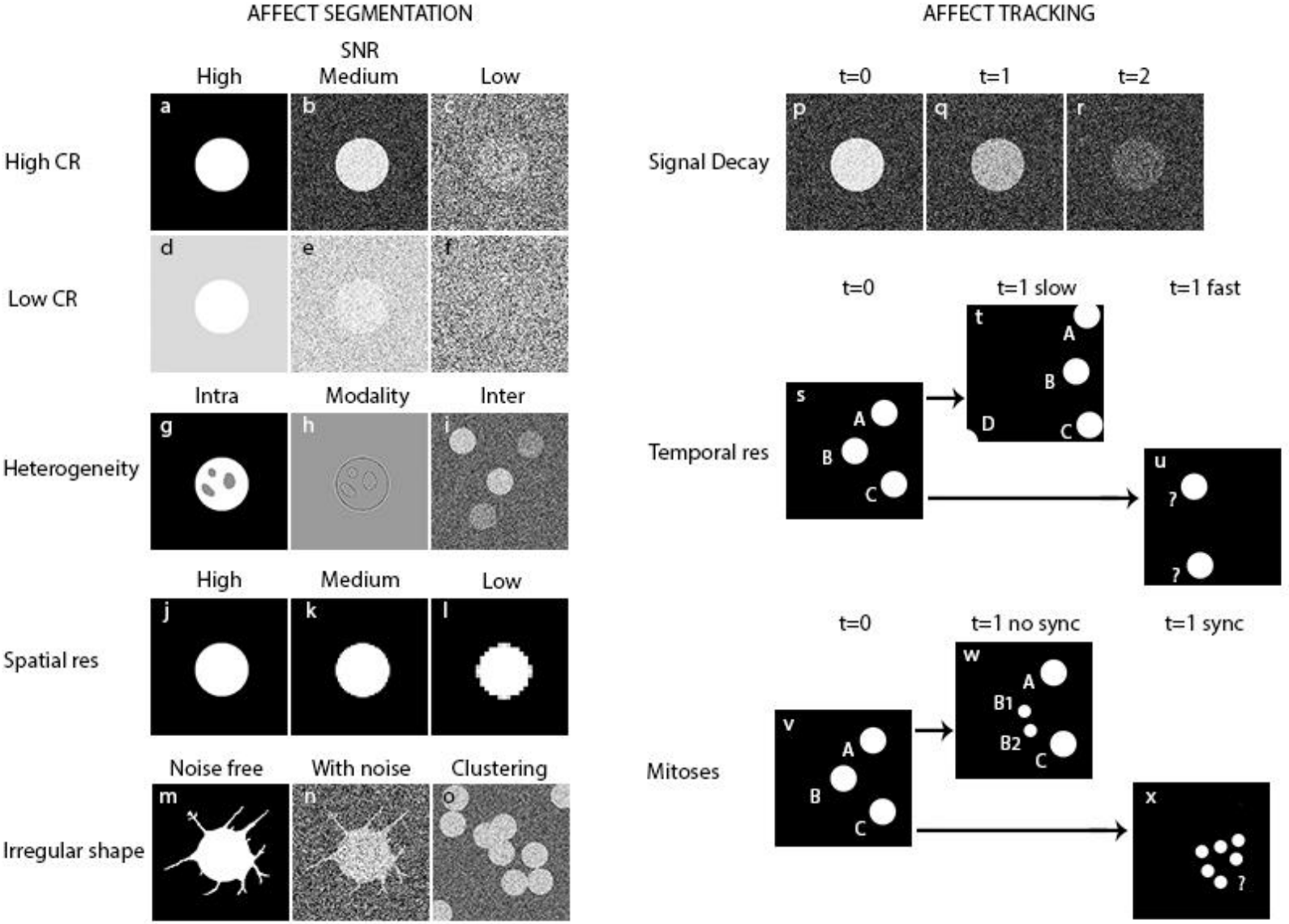
AFFECT SEGMENTATION
SNR
High
Medium
Low
High CR
Low CR
Intra
Modality
Inter
Heterogeneity
High
Medium
Low
Spatial res
Noise free
With noise
Clustering
Irregular shape
AFFECT TRACKING
t=0
t=1
t=2
Signal Decay
t=0
t=1 slow
t=1 fast
Temporal res
t=0
t=1 no sync
t=1 sync
Mitoses

Figure 2. Concept of the main parameters that determine the quality of cell images and videos. **a-f**. **Signal to Noise Ratio (*SNR*) and Contrast Ratio (*CR*).** Simulated cell with 250 intensity units (iu) and no background (0 iu) in three scenarios of increasing standard deviation (std, in iu) of background Gaussian noise: 0 (a); 50 (b); 200 (c). Simulated cell in high background (200 iu) with increasing noise std: 0 (d); 50 (e); 200 (f). The effect of decreased CR is displayed, for increasing noise levels in the following frames: 0 noise (**a** vs. **d**); 50 noise std (b vs. e); 200 noise std (c vs. f). **g-i**. **Signal heterogeneity**. Simulated cell with non-uniform distribution of the labeling marker or non-label retaining structures (g). Signal texture due to the process of image formation, in this case a simulated cell image imaged using Phase Contrast microscopy (h). Signal heterogeneity between cells, i.e., simulated cells using with different average intensities due, for instance, to different levels of protein transfection, non-uniform label uptake, or cell cycle stage or chromatin condensation, when using chromatin-labeling techniques (i). **j-l**. **Spatial resolution**. Simulated cell captured with increasing pixel size, i.e., with decreasing spatial resolution: full resolution (j); half resolution (k); one fourth of the original full resolution (l). **m-o**. **Irregular shape**. Simulated cell with highly irregular shape under two background noise std situations: 0 (m); 100 (n). Cell density within a frame. Several cells in a high-density situation, displaying frequent cell clusters (o). **p-r**. **Fluorescence temporal decay**. Simulated cell in a time series, showing increasing fluorescence decay due to bleaching or quenching of the fluorochrome, and same noise conditions (std of 50 iu): original cell at the beginning of the experiment (p); cell with 100 iu decay (q); cell with 200 iu decay (r). **s-u**. **Cell overlap between consecutive frames**. Three simulated cells at the beginning of the 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 mitoses**. Simulated 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);

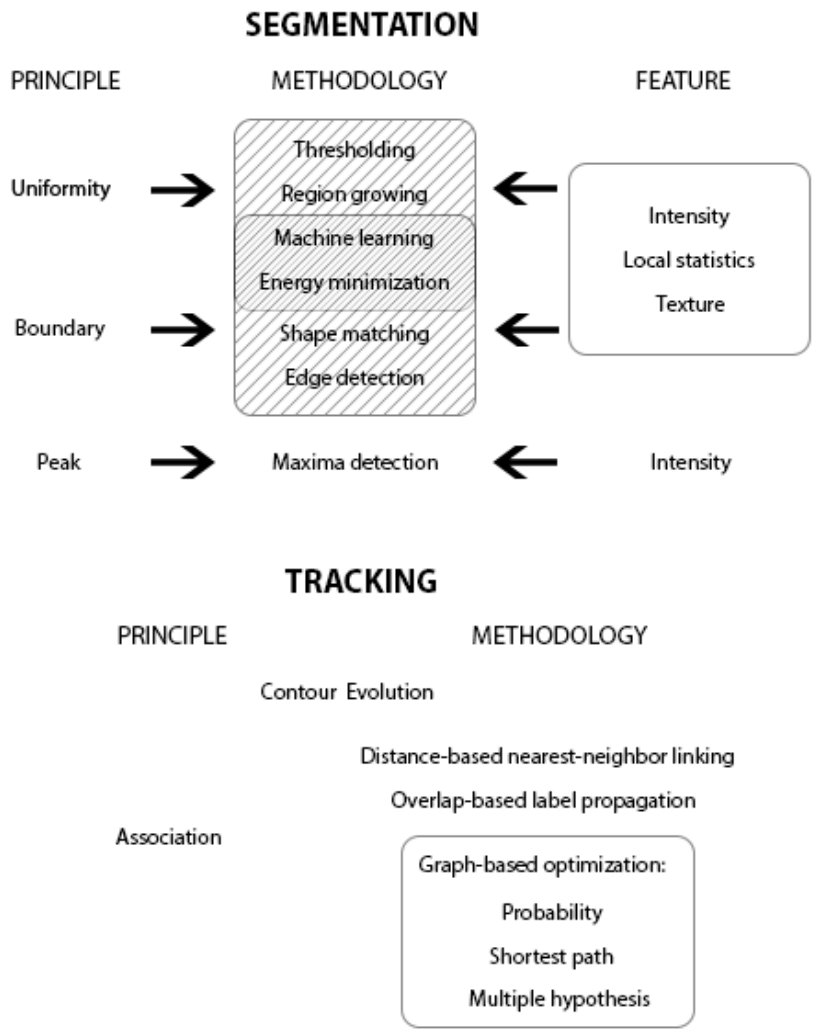

Figure 3. Taxonomy of cell segmentation and tracking methods.

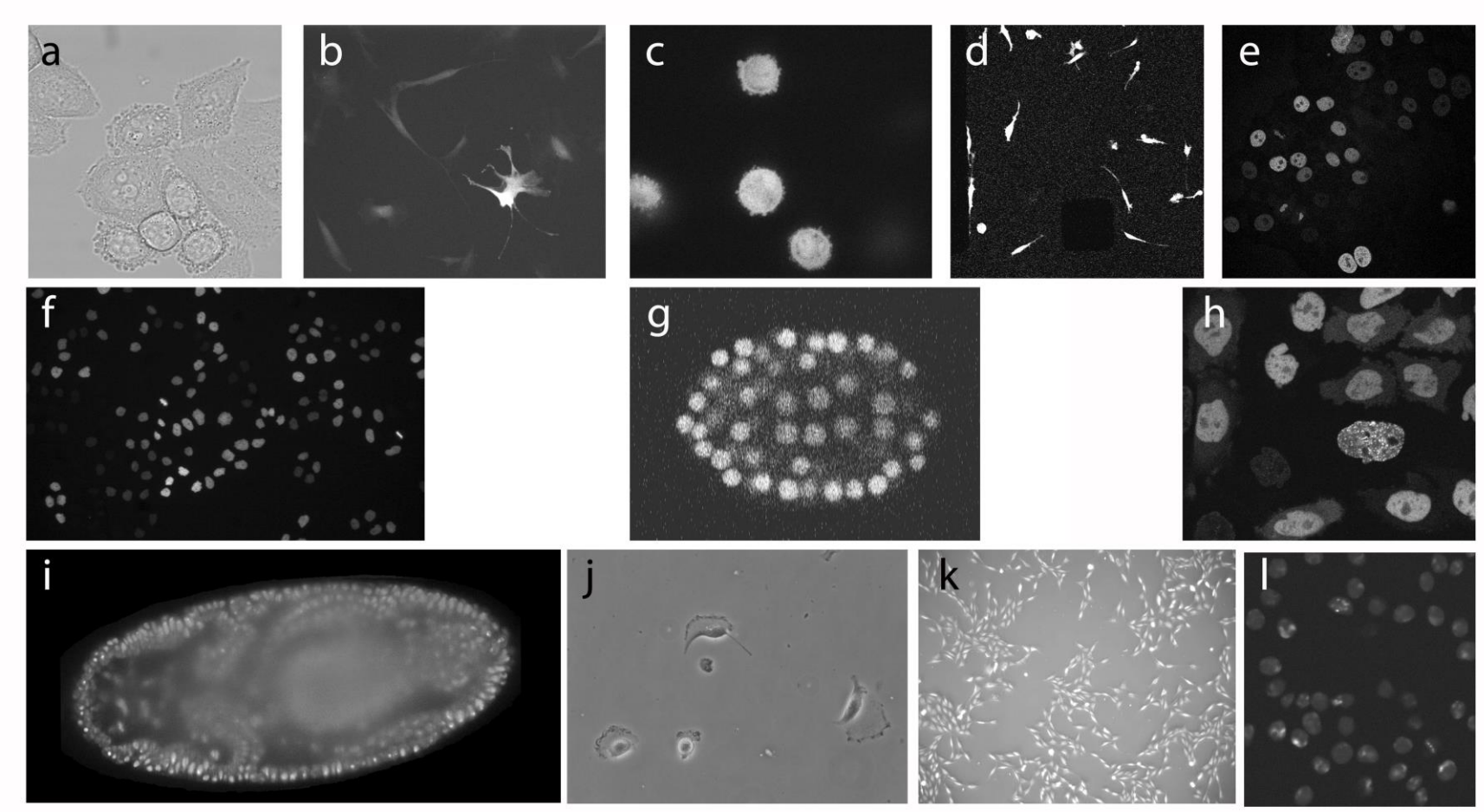

Figure 4. Sample images of the challenge datasets. (**a**) DIC-C2DH-HeLa; (**b**) Fluo-C2DL-MSC; (**c**) Fluo-C3DH-H157; (**d**) Fluo-C3DL-MDA231; (**e**) Fluo-N2DH-GOWT1; (**f**) Fluo-N2DL-HeLa; (**g**) Fluo-N3DH-CE; (**h**) Fluo-N3DH-CHO; (**i**) Fluo-N3DL-DRO; (**j**) PhC-C2DH-U373; (**k**) PhC-C2DL-PSC;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

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*	19.13*	12032	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	11787	0,32	3.8	104.78*	0.72	0.01	N	Y	N	N
Fluo-C3DH-H157	31.53	3.14	0.35	0.42	349593*	0,60	46.6	11.52	0.86	0	N	Y	N	N
Fluo-C3DL-MDA231	9.36	4.24	1.26	0.20	1696	0,60	18.5	8.86	0.71	0.17	N	Y	N	N
Fluo-N2DH-GOWT1	6.16	1131	0.83	0.81	3327	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	6001	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	14494	0,58	33.7	0.005	0.87	0.06	N	Y	Y	N
Fluo-N3DL-DRO	2.46	3.32	0.31	0.18	1188	0,65	12.3	0.98	0.68	1.05	N	N	N	N
PhC-C2DH-U373	2.88	1.10	19.30*	1.01	4287	0,58	48.8	0.04	0.91	0	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	1181	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	38285	0,73	16.2	0.14	0.86	0.49	N	Y	N	N

Table 1. Properties of the competition datasets used in the three editions of the Cell Tracking Challenge. The displayed values correspond to the image/video quality parameters mathematically described in **Online Methods** (section **Dataset quality parameters**).

Legend: ***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 average 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**: apoptotic cells; **Deb**: presence of moving debris.

Color code: For each category and dataset, the average was computed excluding outlying values (*). The background color of the cell indicates whether the highlighted value is within the categories average plus/minus one half of its standard deviation (yellow), or the value is beyond that value (green or red). A red background indicates a poor value in a given category; a green background indicates a high value for a given category. In **Sha**, the 2D and 3D datasets were treated separately because different shape descriptor was used for 2D and for 3D cases.

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 filering 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	None	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

Table 2. Segmentation strategies used by the competing methods. Principle, Feature, and Methodology used in the segmentation phase of the competing algorithms (following the taxonomy shown in **Fig. 2**) along with the preprocessing and postprocessing strategies employed.

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-SE (1)	Association	Graph-based shortest-path global optimization	All	Adjacency- and overlap-based track refinement	Inherent
KTH-SE (2)	Association	Graph-based shortest-path global optimization with detection preprocessing	All	Adjacency based track refinement	Inherent
KTH-SE (3)	Association	Graph-based shortest-path global optimization	All	Adjacency based track refinement	Inherent
KTH-SE (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

Table 3. Tracking strategies used by the competing methods. Principle and Methodology used in the tracking phase of all the competing algorithms (following the taxonomy shown in **Fig. 1**) along with postprocessing strategies employed, the temporal support given, and the scheme followed for the division detection.

	DIC-C2DH-HeLa	Fluo-C2DL-MSC	Fluo-C3DH-H157	Fluo-C3DL-MDA231	Fluo-N2DH-GOWT1	Fluo-N2DL-HeLa	Fluo-N3DH-CE	Fluo-N3DH-CHO	Fluo-N3DL-DRO	PhC-C2DH-U373	PhC-C2DL-PSC	Fluo-N2DH-SIM+	Fluo-N3DH-SIM+
OP	0.828	0.676	0.938	0.757	0.951	0.942	0.688	0.926	0.609	0.951	0.804	0.878	0.848
	0.629	0.636	0.885	0.745	0.902	0.940	0.601	0.912	0.285	0.896	0.772	0.874	0.798
	0.523	0.546	0.870	0.659	0.902	0.901	0.507	0.906	0.219	0.886	0.735	0.859	0.714
SEG_a	0.784 ± 0.066	0.769 ± 0.047	0.924 ± 0.009	0.742 ± 0.047	0.886 ± 0.062	0.904 ± 0.035	0.701 ± 0.154	0.904 ± 0.081	0.840 ± 0.035	0.836 ± 0.143	0.788 ± 0.044	NA	NA
SEG	0.776	0.590	0.888	0.631	0.927	0.903	0.479	0.917	0.561	0.920	0.665	0.791	0.746
	0.460	0.582	0.816	0.625	0.893	0.893	0.422	0.899	0.250	0.826	0.602	0.781	0.629
	0.294	0.465	0.773	0.504	0.887	0.863	0.300	0.898	0.001	0.795	0.572	0.770	0.593
TRA_a	0.965 ± 0.044	0.969 ± 0.014	0.991 ± 0.005	0.935 ± 0.010	0.995 ± 0.003	0.987 ± 0.002	NA	0.985 ± 0.011	NA	0.992 ± 0.006	0.980 ± 0.012	NA	NA
TRA	0.881	0.763	0.987	0.883	0.975	0.991	0.898	0.953	0.657	0.981	0.943	0.975	0.967
	0.797	0.691	0.976	0.865	0.925	0.986	0.781	0.935	0.438	0.978	0.942	0.957	0.950
	0.752	0.645	0.954	0.830	0.916	0.982	0.713	0.914	0.320	0.965	0.898	0.948	0.835

Table 4. Top-three technical performance values (**SEG**, **TRA**, and **OP**) obtained by the competing algorithms. Both the **SEG** and **TRA** sections start respectively with $\mathbf{SEG_a}$ and $\mathbf{TRA_a}$, which are the average values of the measures obtained by three manual annotations used to create the ground truths (**SEG-GT** and **TRA-GT**), 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, see main text) or because no manual annotation was necessary due to the existence of an absolute ground truth (simulated datasets Fluo-N2DH-SIM+ and Fluo-N3DH-SIM+).

	DIC-C2DH-HeLa	Fluo-C2DL-MSC	Fluo-C3DH-H157	Fluo-C3DL-MDA231	Fluo-N2DH-GOWT1	Fluo-N2DL-HeLa	Fluo-N3DH-CE	Fluo-N3DH-CHO	Fluo-N3DL-DRO	PhC-C2DH-U373	PhC-C2DL-PSC	Fluo-N2DH-SIM+	Fluo-N3DH-SIM+
OP	0.828	0.676 (1)	0.938 (1)	0.757 (1)	0.951 (1)	0.942 (1)	0.688 (1)	0.926 (1)	0.609 (2)	0.951	0.804	0.878	0.848 (1)
	0.629 (4)	0.636	0.885	0.745	0.902	0.940	0.601	0.912	0.285	0.896	0.772 (1)	0.874 (1)	0.798
	0.523 (1)	0.546	0.870	0.659 (2)	0.902	0.901	0.507	0.906	0.219	0.886 (3)	0.735	0.859	0.714 (2)
SEG	0.776	0.590 (1)	0.888 (1)	0.631 (1)	0.927 (1)	0.903	0.479 (1)	0.917	0.561 (2)	0.920	0.665	0.791 (1)	0.746 (1)
	0.460 (4)	0.582	0.816	0.625	0.893	0.893 (1)	0.422	0.899 (1)	0.250	0.826	0.602 (1)	0.781	0.629
	0.294 (1)	0.465	0.773	0.504 (2)	0.887	0.863	0.300	0.898	0.001	0.795 (3)	0.572	0.770	0.593 (2)
TRA	0.881	0.763 (1)	0.987 (1)	0.883 (1)	0.975 (1)	0.991 (1)	0.898 (1)	0.953 (1)	0.657 (2)	0.981	0.943	0.975	0.967
	0.797 (4)	0.691	0.976	0.865	0.925	0.986	0.781	0.935	0.438	0.978 (3)	0.942 (1)	0.957 (1)	0.950 (1)
	0.752 (1)	0.645	0.954	0.830	0.916	0.982	0.713	0.914	0.320	0.965	0.898	0.948	0.835 (2)

CUL-UK	CUNI-CZ	FR-Be-GE	FR-Ro-GE
HD-Har-GE	HD-Hau-GE	IMCB-SG (1-2)	KIT-GE
KTH-SE (1-4)	LEID-NL	MU-CZ	NOTT-UK
PAST-FR	UP-PT		UZH-CH

Table 5. Top-three performing methods. For each dataset, 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.

	DIC-C2DH-HeLa	Fluo-C2DL-MSC	Fluo-C3DH-H157	Fluo-C3DL-MDA231	Fluo-N2DH-GOWT1	Fluo-N2DL-HeLa	Fluo-N3DH-CE	Fluo-N3DH-CHO	Fluo-N3DL-DRO	PhC-C2DH-U373	PhC-C2DL-PSC	Fluo-N2DH-SIM+	Fluo-N3DH-SIM+
CT_a	0.667 ± 0.120	0.618 ± 0.034	0.806 ± 0.055	0.696 ± 0.108	0.788 ± 0.016	0.852 ± 0.052	NA	0.743 ± 0.136	NA	0.681 ± 0.250	0.203 ± 0.107	NA	NA
CT	0.017	0.235	0.625	0.354	0.366	0.580	0.257	0.513	0.265	0.573	0.168	0.406	0.456
	0.010	0.083	0.583	0.246	0.360	0.562	0.074	0.456	0.022	0.381	0.060	0.359	0.421
	0.004	0.065	0.487	0.232	0.251	0.550	0.046	0.375	0.000	0.301	0.018	0.347	0.371
$BC(i)_a$	UC	NA	NA	UC	UC	0.942 ± 0.026	NA	UC	UC	NA	0.699 ± 0.172	NA	NA
BC(i)	UC	NA	NA	UC	UC	0.814	0.568	UC	UC	NA	0.536	0.818	0.864
	UC	NA	NA	UC	UC	0.802	0.268	UC	UC	NA	0.475	0.800	0.682
	UC	NA	NA	UC	UC	0.796	0.000	UC	UC	NA	0.252	0.763	0.631
TF_a	0.969 ± 0.009	0.927 ± 0.015	0.984 ± 0.012	0.908 ± 0.051	0.982 ± 0.015	0.980 ± 0.018	NA	0.984 ± 0.012	NA	0.987 ± 0.010	0.876 ± 0.062	NA	NA
TF	0.703	0.672	0.994	0.804	0.942	0.967	0.672	0.988	0.730	0.998	0.803	0.911	0.941
	0.560	0.596	0.980	0.778	0.890	0.966	0.558	0.969	0.319	0.959	0.794	0.892	0.910
	0.395	0.586	0.962	0.717	0.859	0.956	0.531	0.955	NA	0.917	0.720	0.876	0.783
CCA_a	NA	NA	NA	NA	NA	0.963 ± 0.005	NA	NA	NA	NA	0.974 ± 0.021	NA	NA
CCA	NA	NA	NA	NA	NA	0.931	0.760	NA	NA	NA	0.636	0.899	0.929
	NA	NA	NA	NA	NA	0.880	0.579	NA	NA	NA	0.611	0.899	0.894
	NA	NA	NA	NA	NA	0.871	0.426	NA	NA	NA	0.496	0.732	0.741

Table 6. Top-three biological performance values (**CT**, **BC(i), TF** and **CCA**) measures obtained by the competing algorithms. All four **CT**, **BC(i)**, **TF** and **CCA** sections start respectively with $\mathbf{CT_a}$, $\mathbf{BC(i)_a}$, $\mathbf{TF_a}$ and $\mathbf{CCA_a}$, which are the average values of the measures obtained by three manual annotations used to create the ground truths (**SEG-GT** and **TRA-GT**), 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 datasets 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 dataset were: Fluo-N2DL-HeLa (**i**=1, corresponding to a 30-minute 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 datasets where no evidence of entire cell cycles was found (NA).

	DIC-C2DH-HeLa	Fluo-C2DL-MSC	Fluo-C3DH-H157	Fluo-C3DL-MDA231	Fluo-N2DH-GOWT1	Fluo-N2DL-HeLa	Fluo-N3DH-CE	Fluo-N3DH-CHO	Fluo-N3DL-DRO	PhC-C2DH-U373	PhC-C2DL-PSC	Fluo-N2DH-SIM+	Fluo-N3DH-SIM+
CT	0.017 [1]	0.235 [1]	0.625 [1]	0.354	0.366 [1]	0.580	0.257 [1]	0.513	0.265 [2]	0.573	0.168	0.406	0.456
	0.010 [4]	0.083	0.583	0.246	0.360	0.562 [1]	0.074	0.456	0.022	0.381 [3]	0.060 [1]	0.359	0.421 [2]
	0.004	0.065	0.487	0.232 [1]	0.251	0.550	0.046	0.375 [1]	0.000	0.301	0.018	0.347 [1]	0.371
BC(i)	UC	NA	NA	UC	UC	0.814	0.568 [1]	UC	UC	NA	0.536	0.818 [1]	0.864
	UC	NA	NA	UC	UC	0.802 [1]	0.268	UC	UC	NA	0.475 [1]	0.800	0.682 [2]
	UC	NA	NA	UC	UC	0.796	0.000	UC	UC	NA	0.252	0.763	0.631 [1]
TF	0.703 [1]	0.672 [1]	0.994 [1]	0.804	0.942 [1]	0.967 [1]	0.672 [1]	0.988	0.730 [2]	0.998	0.803	0.911 [1]	0.941 [1]
	0.560	0.596	0.980	0.778 [1]	0.890 [1]	0.966	0.558	0.969	0.319	0.959 [1]	0.794 [1]	0.892	0.910
	0.395	0.586	0.962	0.717 [2]	0.859	0.956	0.531	0.955 [1]	NA	0.917	0.720	0.876	0.783
CCA	NA	NA	NA	NA	NA	0.931	0.760 [1]	NA	NA	NA	0.636	0.899	0.929
	NA	NA	NA	NA	NA	0.880 [1]	0.579	NA	NA	NA	0.611 [1]	0.899	0.894 [1]
	NA	NA	NA	NA	NA	0.871	0.426	NA	NA	NA	0.496	0.732 [1]	0.741

CUL-UK		FR-Be-GE	FR-Ro-GE
HD-Har-GE	HD-Hau-GE	IMCB-SG (1-2)	KIT-GE
KTH-SE (1-4)	LEID-NL	MU-CZ	NOTT-UK
PAST-FR	UP-PT	UPM-ES	

Table 7. Top-three performing methods of the three challenge editions in terms of the **CT**, **BC(i)**, and **TF** 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 datasets that do not feature any division event (NA) or at least a minimum number of 50 division events in each video (UC). The dataset 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-minute 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 datasets where no evidence of entire cell cycles was found (NA).

	1st ranked			2nd ranked			3rd ranked		
	NP	**GP**	**TIM**	**NP**	**GP**	**TIM**	**NP**	**GP**	**TIM**
DIC-C2DH-HeLa	FR-Ro-GE 0.828			KTH-SE (4) 0.629			IMCB-SG (1) 0.523		
	4	0.912	4818	14	0.928	622	5	0.924	236
Fluo-C2DL-MSC	KTH-SE (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-SE (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-SE (1) 0.757			LEID-NL 0.745			IMCB-SG (2) 0.659		
	16	0.947	217	9	0.958	992	1	0.935	3506
Fluo-N2DH-GOWT1	KTH-SE (1) 0.951			LEID-NL 0.902			CUNI-CZ 0.901		
	17	0.956	632	9	0.932	1333	8	0.950	479
Fluo-N2DL-HeLa	KTH-SE (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-SE (1) 0.688			HD-Har-GE 0.601			KIT-GE 0.507		
	17	0.895	13475	9	0.889	14518	11	0.872	4258
Fluo-N3DH-CHO	KTH-SE (1) 0.926			MU-CZ 0.912			HD-Har-GE 0.906		
	17	0.954	202	8	0.936	223	10	0.933	1495
Fluo-N3DL-DRO	KTH-SE (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-SE (3) 0.886		
	5	0.965	11450	8	0.953	621	11	0.964	81
PhC-C2DL-PSC	HD-Hau-GE 0.804			KTH-SE (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-SE (1) 0.874			PAST-FR 0.858		
	3	0.979	20124	17	0.982	301	9	0.977	370
Fluo-N3DH-SIM+	KTH-SE (1) 0.848			LEID-NL 0.798			IMCB-SG (2) 0.714		
	17	0.985	13115	9	0.972	66773	9	0.988	69549

Table 8. Usability evaluation of the top-three ranked algorithms for each dataset. Legend: **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 dataset and parameter, red background means the worst value across the three methods, yellow means the intermediate value, and green means the best value out of those three listed.