An objective comparison of cell-tracking algorithms

Ulman, V.; Maska, M.; Magnusson, K.E.G.; Ronneberger, O.; Haubold, C.; Harder, N.; ... ; Ortiz-de-Solorzano, C.

HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2018 April 30.

Published in final edited form as:

Nat Methods. 2017 December ; 14(12): 1141–1152. doi:10.1038/nmeth.4473.

An Objective Comparison of Cell Tracking Algorithms

Vladimír Ulman[1,a,**], **Martin Maška**[1,**], **Klas E. G. Magnusson**[2], **Olaf Ronneberger**[3,b], **Carsten Haubold**[4], **Nathalie Harder**[5,c], **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,d], **Yuexiang Li**[11,e], **Siu-Yeung Cho**[12], **Alexandre C. Dufour**[13], **Jean-Christophe Olivo-Marin**[13], **Constantino C. Reyes-Aldasoro**[14], **Jose A. Solis-Lemus**[14], **Robert Bensch**[3], **Thomas Brox**[3], **Johannes Stegmaier**[15], **Ralf Mikut**[15], **Steffen Wolf**[4], **Fred. A. Hamprecht**[4], **Tiago Esteves**[16,17], **Pedro Quelhas**[16], **Ömer Demirel**[18], **Lars Malmström**[18], **Florian Jug**[19], **Pavel Tomancak**[19], **Erik Meijering**[6], **Arrate Muñoz-Barrutia**[20,21], **Michal Kozubek**[1], and **Carlos Ortiz-de-Solorzano**[22,23,*]

[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, Institut 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

[*]Corresponding author (codesolorzano@unav.es).
[a]Current affiliation: Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
[b]Current affiliation: DeepMind, London, UK
[c]Current affiliation: Definiens AG, Munich, Germany
[d]Current affiliation: National Heart Research Institute Singapore (NHRIS), National Heart Centre Singapore (NHCS), Singapore
[e]Current affiliation: Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[**]These authors contributed equally to this work

Competing financial interests statement

All the authors declare not to have competing financial interests.

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, IDISNA and Program of Solid Tumors and Biomarkers, Center for Applied Medical Research, University of Navarra, Pamplona, Spain [23]Bioengineering Department, TECNUN School of Engineering, University of Navarra, San Sebastián, Spain

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. With twenty-one participating algorithms and a data repository consisting of thirteen datasets of various microscopy modalities, the challenge displays today's state of the art in the field. We analyze the results using performance measures for segmentation and tracking that rank all participating methods. We also analyze the performance of all algorithms in terms of biological measures and their practical usability. Even though some methods score high in all technical aspects, not a single one obtains fully correct solutions. We show that methods that either take prior information into account using learning strategies or analyze cells in a global spatio-temporal video context perform better than other methods under the segmentation and tracking scenarios included in the 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 time-lapse microscopy recordings of cell behavior, it is often necessary to identify individual cells and follow them over time. The bioimage processing community has, since its inception, worked on extracting 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 cell movements over time, providing information about their velocities and trajectories, and detecting cell lineage changes due to cell division or cell death (Fig. 1). The level of difficulty of automatically segmenting and tracking cells depends on the quality of the recorded video

sequences. The main properties that determine the quality of time-lapse videos with respect to the subsequent segmentation and tracking analysis are graphically illustrated in Fig. 2, and expressed as a set of quantitative measures in the **Online Methods** (section **Dataset quality parameters**).

The image processing community has addressed the above-mentioned tasks using 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 *cell 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 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 segmenting the cells in all frames of a video and later, using mostly probabilistic frameworks, establish temporal associations between the segmented cells. This can be done by either using a two-frame or multi-frame sliding window, or even for all frames at once.

The diversity of imaging modalities, cell tracking tasks, and available algorithms makes it difficult for biologists to decide which algorithm to use under certain conditions. Moreover, the developers of image processing algorithms need to objectively evaluate new cell segmentation and tracking solutions by comparing their performance on standardized datasets. We addressed these problems by organizing three Cell Tracking Challenges (CTC I–III) between 2013 and 2015. For these challenges, we created a diverse repository of annotated microscopy videos, and defined quantitative evaluation measures to allow a fair comparison of the competing algorithms[30]. Here, we present a combined report on all three CTC editions. We introduce the datasets and show the results obtained by the participating algorithms. 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 (http://celltrackingchallenge.net/).

Results

Datasets and ground truth

The dataset repository (Fig. 4, Supplementary Table 1, Supplementary Videos 1–13) consists of 52 annotated videos from 13 classes, occupying 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]. Supplementary Note 1 and supporting Supplementary Figs. 1–11 provide a detailed description of the datasets. Supplementary Note 2 and supporting Supplementary Fig. 12 describe the simulator used to create the synthetic datasets, applying the parameter configuration provided in Supplementary Data 1. Finally, Table 1 provides a quantitative characterization of the quality of each dataset, based on the 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.

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

Three independent human experts created a segmentation and a tracking solution (annotation) for each non-synthetic video[30]. The final segmentation (**SEG-GT**) and tracking (**TRA-GT**) ground truths were created by combining the three annotations, following a majority-voting scheme[30]. **SEG-GT** for the datasets of *C. elegans* (Fluo-N3DH-CE) embryo and the *Drosophila melanogaster* embryo (Fluo-N3DL-DRO) embryos were generated as described above, but in the case of Fluo-N3DL-DRO, only cells of the early nervous system were annotated and used as ground truth. **TRA-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]. For the synthetic videos, **SEG-GT** and **TRA-GT** were inherently created by the cell simulator used[31].

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 respectively their segmentation and tracking strategies. Supplementary Table 2 lists affiliations of the participating teams, and Supplementary Table 3 contains links to the

executable versions of most of the submitted algorithms. Their expanded description is presented in the Supplementary Note 3 and the parameter configurations used by each algorithm are listed in the Supplementary Data 2. All submissions were received by the CTC organizers as labeled segmentation masks and structured text files containing the cell lineage graphs. The CTC organizers verified the submitted results by reproducing them on a single computer, using the executable version of each algorithm provided by the participants.

Quantitative performance criteria

In order to quantify the performance of all submitted algorithms, we developed three categories of measures that 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 detailed description of all measures can be found in the **Online Methods** (section **Performance criteria**). It is important to 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 here to analyze aspects that are of relevance from the user point of view. Supplementary Table 3 contains a link to the evaluation software used in the challenge.

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

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. **Complete Tracks** (**CT**) measures the fraction of ground truth cell tracks that a given method is capable to reconstruct in their entirety, from the frame they appear in, to the frame they disappear from. **CT** is especially relevant when a perfect reconstruction of the cell lineages is required. **Track Fractions** (**TF**) averages, for all detected tracks, the fraction of the longest continuously matching algorithm-generated tracklet with respect to the reference track. Intuitively, this can be interpreted as the fraction of an average cell's trajectory that an algorithm reconstructs correctly, once the cell has been detected. **Branching Correctness** (**BC**) measures how efficient a method is at correctly detecting division events. Finally, the **Cell Cycle Accuracy** (**CCA**) measures how accurate an

algorithm is at correctly reconstructing the length of cell cycles (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 expresses the practical usability of the submitted algorithms. The first indication of an algorithm's usability is the **number of tunable parameters** (**NP**) a user is required to manually set, excluding parameters visible only to developers. In general, a lower number of tunable parameters 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 videos acquired under otherwise unchanged imaging conditions. **GP** values are computed by comparing the results for a particular training and competition video, obtained using the same parameter configuration. This measure takes values in the interval [0,1], with higher values corresponding to better generalizability. The last value we report for each algorithm is its **execution time** (**TIM**), in seconds.

Analysis of the performance of submitted algorithms

All measures described have been computed for every dataset and competing algorithm. We first evaluated the segmentation (**SEG**) and tracking (**TRA**) accuracy measures. Top-three values and participants for each dataset are listed in Figs. 5 and 6 (see Supplementary Data 3 for the complete list of values). To determine the significance of these values, we calculated **SEG** and **TRA** values with respect to the ground truth also for the three manual annotations, 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 trying to pinpoint the features that underlie the good and not so good performance of the competing methods (Fig. 5). We observe that some algorithms reached very good values (**OP** > 0.9) for datasets 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. 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 human annotators, meaning that after automatic tracking some additional curation work is required to reach the level of the human-level solutions. The difficulty for DIC-C2DH-HeLa

and PhC-C2DL-PSC 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 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 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 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 videos, which strongly complicates tracking of the segmented cells (see Supplementary Figs. 7 and 9).

Next, we examine the results from the viewpoint of the algorithms, asking which ones show best overall performance (Fig. 6). The algorithms KTH-SE, FR-Ro-GE, and HD-Hau-GE ranked first for one or more datasets. Looking more globally at the number of top-three occurrences, 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 algorithms for all datasets. These methods rely on a simple thresholding for segmentation, the results of which are highly enriched by the use of global information in the tracking process. In some datasets, however, the *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 these methods rely 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 improve the learning capacity of the algorithms.

Interestingly, the evolution of the average of the top-three **OP** values during the three CTC editions shows progress towards the objective of reaching the level of the human expert annotators (Supplementary Fig. 13). On average across all datasets, the average top-three **OP** values rose by 0.03±0.03 (CTC II vs CTC I) and 0.05±0.07 (CTC III vs CTC I).

We studied the robustness of the **OP**-based rankings, as described in the **Online Methods** (section **Ranking robustness**) and summarized in Supplementary Fig. 14, 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 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. Fig. 7 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 ($\mathbf{CT}_a$, $\mathbf{TF}_a$, $\mathbf{BC}_a$, and $\mathbf{CCA}_a$). When looking at **CT** across datasets, 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 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 cell speeds and trajectories, but some work is still required to bring them to the level of the human annotators. Finally, Fluo-N2DL-HeLa, Fluo-N2DH-SIM+, and Fluo-N3DH-SIM+ 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, Fluo-N3DH-CE, and presumably Fluo-N3DL-DRO would benefit from improved management of division events as revealed by their low **BC** and **CCA** values.

When analyzing the performance of the individual algorithms in terms of **CT** and **TF** (Fig. 8 and Supplementary Data 4), we see similar but not completely matching pictures compared to the ranking compiled using **SEG** and **TRA** (Fig. 6). This is because **TF** and **CT** consider only tracking correctness, regardless of the accuracy of the segmentation, and have much more strict requirements on correctly reconstructed tracks. 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 two times for **TF** and four times for **CT**). This highlights that *tracking by contour evolution* methods can be superior at following cells, once a track has been initiated, if the temporal resolution of the image data permits. As a final comment, methods that inherently (KTH-SE, HD-Hau-GE, IMCB-SG) or specifically (HD-Har-GE, LEID-NL) detect cell division events 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 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, speed, and general applicability of all top-ranked algorithms. From the results shown in Table 4 (see Supplementary Data 5 for a complete list), we 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 to 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 image 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 image data at an intermediate level of complexity, for which an absolute ground truth inherently exists. Together, this represents a unique and rich resource of annotated, real and simulated image data that distinguishes our challenge from similar events that relied exclusively on simulated data[34]. Second, we developed a set of measures that quantitatively evaluate the performance of submitted solutions against the ground truth data in terms of accuracy, biological relevance of the results, and usability for biologists. Third, over the course of three challenges, we assembled a diverse collection of competing solutions that represent all main algorithmic approaches to cell segmentation and tracking problems in biology. Fourth, in this report we analyze the accumulated results 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, possibly global, 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 have been optimized, manually or using automatic supervised or unsupervised algorithms, and the software runs on decent hardware, the best methods will perform well on all similar microscopy recordings. Finally, we acknowledge that 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 videos show very high cell densities in later frames, the size of the image data itself causes very long runtimes. *Tracking by detection approaches* fail on these 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 both segmentation and tracking accuracy 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 the performance of the segmentation approach.

Although the challenge was broadly taken on by the community and many algorithms competed, it is important to stress that the voluntary nature of participation necessarily resulted in 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].

The Cell Tracking Challenge, which remains open for online submissions, is a powerful resource for algorithm developers and users alike. Along with the datasets, we offer the

evaluation suite, capable of computing the technical and biologically oriented measures as well as the dataset quality parameters we have introduced, as an open-source Fiji plugin[35], and provide executable versions of most of the participants' algorithms. Furthermore, we will encourage past and future participants to make their submitted algorithms available to biologists via easy to install and intuitive graphical user interfaces. In the future, new datasets of existing and new microscopy modalities will be incorporated to the dataset repository. It will be particularly important to collect and annotate complex tissue, organ, and whole embryo image data. Finally, we intend to add new synthetic datasets that closely mimic the variety of cell types and microscopy scenarios. These synthetic image data will model different cell labeling, cell shapes, and cell behaviors and migration patterns in 2D and 3D. Since artificially generated datasets implicitly bear absolute ground truth, they can be tuned to challenge algorithms to improve specific aspects of the problem (e.g., 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.

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 constitutes a representative amount for the evaluation of segmentation performance. To extend the manual ground truth to cover as many as possible of the cells in the videos, we first combined the manual tracking ground truth (**TRA-GT**) with the segmentation masks provided by the participants. For any marker in **TRA-GT**, we automatically merged the top-performing participants' segmentation masks that overlap the majority of this tracking marker. The number of masks used was determined manually for each video. On average, a majority of the total number of available masks were used. The process led occasionally to colliding situations, that is, when obtained segmentation masks for two different tracking markers were overlapping. If the overlap was less than 10% of the mask area/volume, the intersecting pixels/voxels were removed from both colliding masks in an expectation that 10% loss will not significantly influence the measured quantities. Otherwise, both entire masks were discarded. In this way, a rich consensus-based segmentation with reliable linking was obtained for all real challenge videos. The synthetic 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 values of the signal-to-noise ratio (*SNR*), contrast ratio (*CR*), internal signal heterogeneity of the cells (Het_i), resolution (*Res*), regularity of the cell shape (*Sha*), cell density (*Den*), and level of cell overlap in consecutive frames (*Ove*) were established as averages of $\boldsymbol{SNR_{i,t}}$, $\boldsymbol{CR_{i,t}}$, $\mathbf{HETi_{i,t}}$, $\mathbf{Res_{i,t}}$, $\mathbf{Sha_{i,t}}$, $\mathbf{Den_{i,t}}$,and $\mathbf{Ove_{i,t}}$ values, respectively, calculated for every object in every image in both competition videos:

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

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

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

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

$$Res_{i,t}=|FG_{i,t}|$$

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

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

where **|S|** is the size of the set **S** and **I(t)** is the set of indices of all cells or nuclei segmented in the **t**-th image. The heterogeneity of the signal between cells (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 absolute difference between the average object intensity at the end and the beginning of a video was divided by the number of its frames minus one and averaged over both videos in a dataset. The number of division events (*Mit*) is computed as 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 events (which are marked in the tracking ground truth

TRA-GT). The remaining qualitative parameters, synchronization of division events (*Syn*), cells entering or leaving the field of view (*Ent*/*Leav*), apoptotic cells (*Apo*), and the presence of moving debris (*Deb*), were set after manual inspection of the datasets.

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 averaged again (i.e., $\mathbf{OP} = 0.5 \cdot (\mathbf{SEG}_{\mathbf{avg}} + \mathbf{TRA}_{\mathbf{avg}})$), 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.

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

Branching Correctness (BC(i))[28,29]: **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., one provided by the method and one in the ground truth), they are allowed to be separated by no more than **i** frames. More specifically, we allowed the reconstruction of division events using a tolerance window of (**2.i+1)** frames. The tolerance value **i** used for each 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, considering only those tracks that are both initiated and terminated by a branching event. Each such track witnesses the development of a cell from its birth until its next division, and its length, therefore, corresponds to the cell cycle length of that cell. The **CCA** measure is defined as:

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

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

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

Usability Measures

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

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

Execution time (TIM): For each dataset, we accumulated the time (in seconds) that was required to analyze each competition video.

Ranking robustness

For each dataset, we ranked all methods based on their **SEG** and **TRA** scores using the formula $0.5 \cdot (a \cdot \mathbf{SEG} + b \cdot \mathbf{TRA})$, $a,b \in \{0, 0.001, 0.002, \ldots, 1\}$, and calculated the number of changes between each such ranking and the one compiled using **OP** (i.e., when a equals to b). Supplementary Fig. 14 plots the number of changes for every combination of weights. As can be seen, 45 % of the area (i.e. of possible weight configurations) causes no more than two changes in the rankings across all datasets.

Data availability

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

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

The entire set of evaluation measures obtained and used to compare the algorithms (used to produce Figs. 5–8, Table 4, Supplementary Figs. 13 and 14 and Supplementary Table 4) is

provided with this article as Supplementary Data 3 (**SEG**, **TRA**, and **OP**), **4** (**CT**, **TF**, **BC**, and **CCA**), and **5** (**NP**, **GP**, and **TIM**).

Code availability

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

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to acknowledge the following funding sources: The Spanish Ministry of Economy MINECO grants DPI2012-38090-C03-02 (C.O.d.S.) and DPI2015-64221-C2-2 (C.O.d.S.), TEC2013-48552-C2-1-R (A.M.B.), TEC2015-73064-EXP (A.M.B), and TEC2016-78052-R (A.M.B.); Netherlands Organization for Scientific Research (NWO) grants 612.001.018 (M.R., 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.); the Czech Ministry of Education, Youth and Sports grant LTC17016 in the frame of EU COST NEUBIAS project (M.M., Pa.M., Pe.M., D.S., M.K.); 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+), RNA-Code (e:Bio), and de.NBI, as well as the DFG, SFB 1129 and RTG 1653 (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 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, and 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 and Philipp Fischer.

Author’s contributions

V.U.: 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. Provided the Fiji plugin with the evaluation suite.

M.M.: 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.

K.E.G.M.: Top ranked challenge participant. Contributed to the writing of the manuscript.

O.R.: Top ranked challenge participant. Contributed to the writing of the manuscript.

C.H.: Top ranked challenge participant. Contributed to the writing of the manuscript.

N.H.: Top ranked challenge participant.

Pa.M.: 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.

Pe.M.: Actively participated in the organization of the CTC challenges: prepared data and supervised data annotation.

D.S.: 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.

M.R.: Actively participated in the organization of the CTC challenges: prepared data and supervised data annotation.

I.S.: Actively participated in the organization of the CTC challenges: prepared data and supervised data annotation.

K.R.: Challenge participant.

J.J.: Challenge participant.

H.M.B.: Challenge participant.

O.D.: Challenge participant.

B.L.: Challenge participant.

P.X.: Challenge participant.

Y.L.: Challenge participant.

S.-Y.C.: Challenge participant.

A.C.D.: Challenge participant.

J-.C.O.-M.: Challenge participant.

C.C.R.-A.: Challenge participant.

J.A.S.-L.: Challenge participant.

R.B.: Challenge participant.

T.B.: Challenge participant.

J.S.: Challenge participant.

R.M.: Challenge participant.

S.W.: Challenge participant.

F.A.H.: Challenge participant.

T.E.: Challenge participant.

P.Q.: Challenge participant.

Ö.D.: Challenge participant.

L.M.: Challenge participant.

F.J.: Contributed to the revision of the manuscript and supported V.U. with the related data processing.

P.T.: Challenge organizer. Contributed to the revision of the manuscript.

E.M.: Challenge organizer. Contributed to the writing of the manuscript.

A.M.-B.: Challenge organizer. Contributed to the writing of the manuscript.

M.K.: Challenge organizer. Contributed to the writing of the manuscript.

C.O.-de-S.: Challenge organizer. Coordinated the work of the committee that organized the challenges. Wrote the manuscript with the input from all authors.

References

1. Franz CM, Jones GE, Ridley AJ. Cell migration in development and disease. Dev Cell. 2002; 2:153–158. [PubMed: 11832241]
2. Bullen A. Microscopy imaging techniques for drug discovery. Nat Rev Drug Discov. 2007; 7:54–67.
3. Walter, RJ., Berns, MW. Digital image processing and analysis, in Video Microscopy. Inoué, S., editor. Springer Sciences; 1986. p. 327-392.
4. Schneider CA, Rasband WS, Elicieri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012; 9:671–675. [PubMed: 22930834]
5. Meijering E. Cell segmentation: 50 years down the road. IEEE Signal Proc Mag. 2012; 29:140–145.
6. Dufour AC, Liu T-Y, Ducroz C, Tournemenne R, Cummings B, Thibeaux R, Guillen N, Hero AO, Olivo-Marin JC. Signal processing challenges in quantitative 3-D cell morphology: More than meets the eye. IEEE Signal Proc Mag. 2015; 32:30–40.
7. Zimmer C, Zhang B, Dufour A, Thebaud A, Berlemont S, Meas-Yedid J, Olivo-Marin JC. On the digital trail of mobile cells. IEEE Signal Proc Mag. 2006; 23:54–62.
8. Wuttisarnwattana P, Gargesha M, van't HofW, Cooke KR, Wilson DL. Automatic stem cell detection in microscopic whole mouse cryo-imaging. IEEE Trans Med Imag. 2016; 35:819–829.

9. Lerner B, Clocksin WF, Dhanjal S, Hultén S, Bishop CM. Automatic signal classification in fluorescence in situ hybridization images. Cytometry. 2001; 43:87–93. [PubMed: 11169572]
10. Chen X, Zhou X, Wong STC. Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. IEEE Trans Biomed Eng. 2006; 53:762–766. [PubMed: 16602586]
11. Henry KM, Pase L, Ramos-Lopez CF, Lieschke GJ, Renshaw SA, Reyes-Aldasoro CC. PhagoSight: an open-source MATLAB package for the analysis of fluorescent neutrophil and macrophage migration in a zebrafish model. PloS ONE. 2013; 8:e72636. [PubMed: 24023630]
12. Wählby C, Sintorn IM, 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. 2004; 215:67–76.
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; 2013. p. 669-674.
14. Türetken E, Wang X, Becker CJ, Haubold C, Fua P. Network flow integer programming to track elliptical cells in time-lapse sequences. IEEE Trans Med Imag. 2016; 36:942–951.
15. Malpica N, Ortiz-de-Solorzano C, Vaquero JJ, Santos A, Vallcorba I, Garcia-Sagredo JM, Pozo F. Applying watershed algorithms to the segmentation of clustered nuclei. Cytometry Part A. 1997; 28:289–297.
16. Ortiz-de-Solorzano C, García-Rodríguez E, Jones A, Pinkel D, Gray JW, Sudar D, Lockett SJ. Segmentation of confocal microscopy images of cell nuclei in thick tissue sections. J Microsc-Oxford. 1999; 193:212–226.
17. Cliffe A, Doupé DP, Sung H, Lim IKH, Ong KH, Cheng L, Yu W. Quantitative 3D analysis of complex single border cell behaviors in coordinated collective cell migration. Nat Commun. 2017; 8:14905. [PubMed: 28374738]
18. Ronneberger O, Fisher P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Proc MICCAI 2015 LCNS. 2015; 9351:234–241.
19. Schiegg M, Hanslovsky P, Haubold C, Koethe U, Hufnagel L, Hampretch FA. Graphical model for joint segmentation and tracking of multiple dividing cells. Bioinformatics. 2015; 31:948–56. [PubMed: 25406328]
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. 2002; 21:1212–1221.
21. Dufour A, Thibeaux R, Labruyere E, Guillen N, Olivo-Marin JC. 3D active meshes: fast discrete deformable models for cell tracking in 3D time-lapse microscopy. IEEE Trans Image Process. 2011; 20:1925–37. [PubMed: 21193379]
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. 2013; 32:995–1005.
23. Ortiz-de-Solorzano C, Malladi R, Lelievre SA, Lockett SJ. Segmentation of nuclei and cells using membrane related protein markers. J Microsc-Oxford. 2001; 201:404–415.
24. Dzyubachyk O, van Cappellen WA, Essers J, Niessen WJ, Meijering E. Advanced level-set-based cell tracking in time-lapse fluorescence microscopy. IEEE Trans Med Imag. 2010; 29:852–867.
25. Dufour A, Shinin V, Tajbakhsh S, Guillen-Aghion N, Olivo-Marin JC, Zimmer C. Segmenting and tracking fluorescent cells in dynamic 3D microscopy with coupled active surfaces. IEEE Trans Image Process. 2005; 14:1396–1410. [PubMed: 16190474]
26. Bensch R, Ronneberger O. Cell segmentation and tracking in phase contrast images using graph cut with asymmetric boundary costs. Proc 2015 IEEE Int Symp Biomed Imaging (ISBI). 2015:1120–1123.
27. Harder N, Mora-Bermúdez F, Godinez WJ, 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. 2009; 19:2113–2124. [PubMed: 19797680]
28. Bise R, Yin Z, Kanade T. Reliable cell tracking by global data association. Proc 2011 IEEE Int Symp Biomed Imaging (ISBI). 2011:1004–1010.

29. Magnusson KEG, Jaldén J, Gilbert PM, Blau HM. Global linking of cell tracks using the Viterbi algorithm. IEEE Trans Med Imag. 2015; 34:1–19.

30. Maška M, Ulman V, Svoboda D, Matula Pt, Matula Pv, Ederra C, Urbiola A, España T, Venkatesan S, Balak DMW, Karas P, Bolcková T, Štreitová M, Carthel C, Coraluppi S, Harder N, Rohr K, Magnusson KEG, Jaldén J, Blau HM, Dzyubachyk O, Křížek P, Hagen GM, Pastor-Escuredo D, Jimenez-Carretero D, Ledesma-Carbayo MJ, Muñoz-Barrutia A, Meijering E, Kozubek M, Ortiz-de-Solorzano C. A benchmark for comparison of cell tracking algorithms. Bioinformatics. 2014; 30:1609–1617. [PubMed: 24526711]

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. 2017; 36:310–321. [PubMed: 27623575]

32. Murray JI, Bao Z, Boule TJ, Boeck ME, Mericle BL, Nicholas TJ, Zhao Z, Sandel MJ, Waterston RH. Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. Nat Methods. 2008; 5:703–9. [PubMed: 18587405]

33. Amat F, Lemon W, Mossing DP, McDole K, Wan Y, Branson K, Myers EW, Keller PJ. Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. Nat Methods. 2014; 11:951–8. [PubMed: 25042785]

34. Chenouard N, Smal I, de Chaumont F, Maška M, Sbalzarini IF, Gong Y, et al. Objective comparison of particle tracking methods. Nat Methods. 2014; 11:281–289. [PubMed: 24441936]

35. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Rueden C, Saafeld S, Schmid B, Tinevez JY, White DJ, Hartenstein V, Elliceri K, Tomancak P, Cardona A. Fiji: an open source platform for biological-image analysis. Nat Methods. 2012; 9:676–82. [PubMed: 22743772]

36. Klette, R., Zamperoni, P. Handbook of image processing operators. In: Klette, Reinhard, Zamperoni, Piero, editors. Handbook of image processing operators. Chichester; New York: Wiley; 1996.

37. Lin CL, Miller JD. 3D characterization and analysis of particle shape using X-ray microtomography (XMT). Powder Technology. 2005; 154:61–69.

38. Matula, Pa, Maška, M., Sorokin, DV., Matula, Pe, Ortiz-de-Solorzano, C., Kozubek, M. Cell Tracking Accuracy Measurement Based on Comparison of Acyclic Oriented Graphs. PLoS One. 2015; 10:e0144959. [PubMed: 26683608]

39. Li K, Miller ED, Chen M, Kanade T, Weiss LE, Campbell PG. Cell population tracking and lineage construction with spatiotemporal context. Med Image Anal. 2008; 12:546–566. [PubMed: 18656418]

40. Brown MR, Summers HD, Rees P, Smith PJ, Chappell SC, Errington RJ. Flow-based cytometric analysis of cell cycle via simulated cell populations. PLoS Comput Biol. 2010; 6:e10000741.

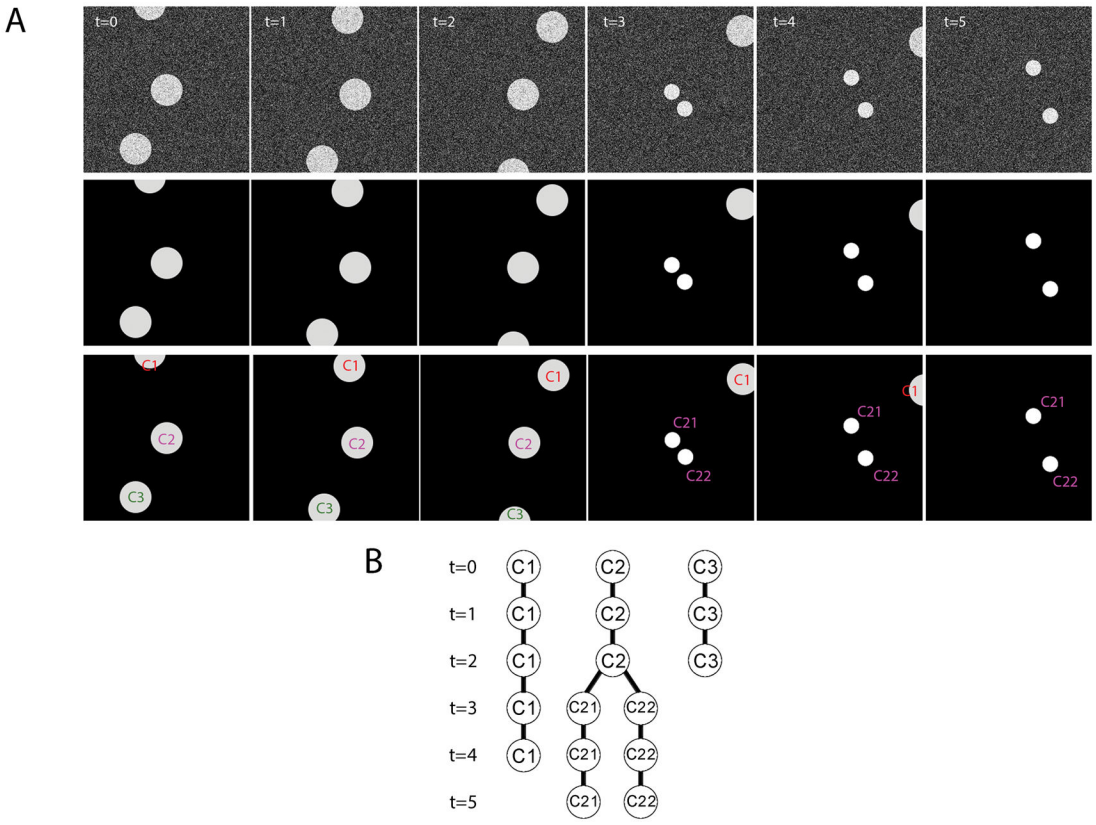


Figure 1. Concept of cell segmentation and tracking

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 toward t=5.

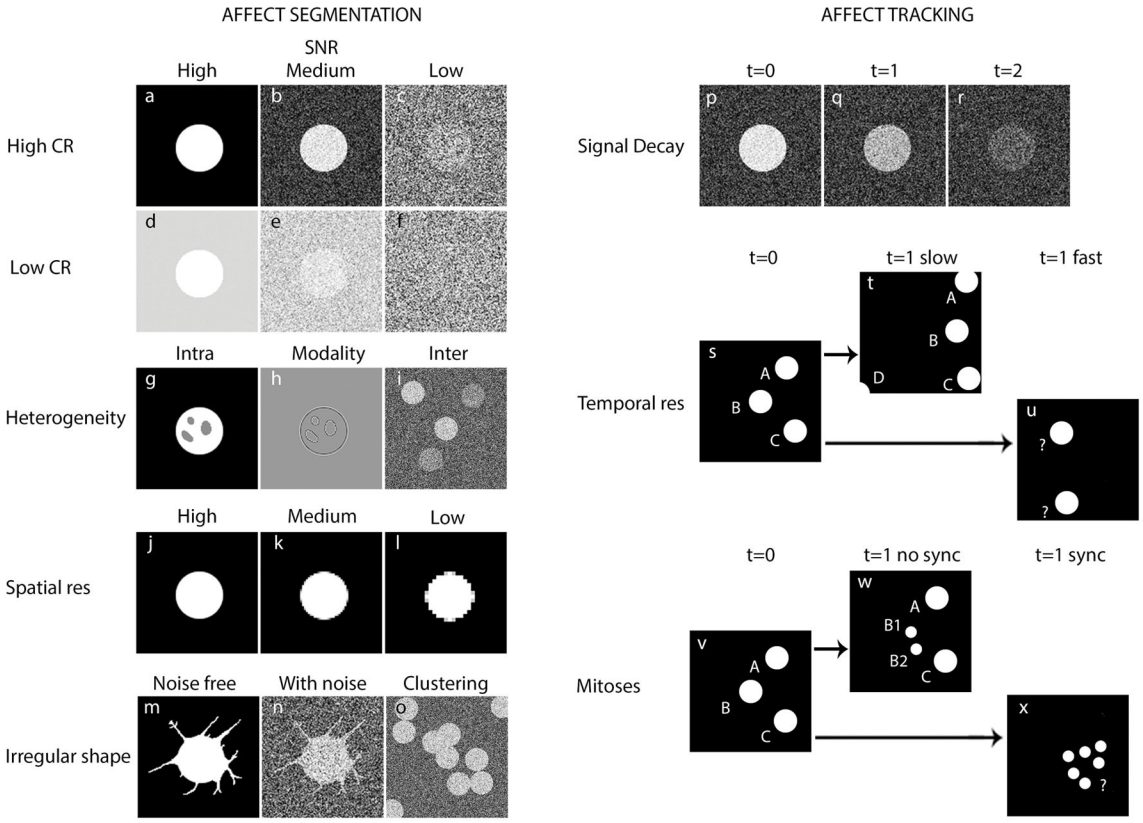


Figure 2. Concept of the main factors that determine the quality of cell images and videos

a–f. Signal to Noise Ratio (*SNR*) and Contrast Ratio (*CR*) measure the relationship between the signal captured from the cells and the unwanted noise or signal captured at the same time. Decreasing **SNR** is shown using a cell with 250 intensity units (iu) and no background (0 iu) in three scenarios of increasing standard deviation (std, in iu) of background Gaussian noise: 0 (a); 50 (b); 200 (c). The effect of decreased *CR* is displayed using a simulated cell in high background (200 iu) with increasing noise std: 0 (d); 50 (e); 200 (f). The effect is shown for three increasing noise: 0 noise (**a** vs. **d**); 50 noise std (b vs. e); 200 noise std (c vs. f). **g–h. Intra-cellular signal heterogeneity** that can lead to cell over-segmentation when the same cell yields several detections is simulated by a cell with non-uniform distribution of the labeling marker or non-label retaining structures (g). Signal texture can also be linked to the process of image formation, in this case shown using a simulated cell image imaged by Phase Contrast microscopy (h). **i. Signal heterogeneity between cells**, shown by simulated cells with different average intensities can be 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. **j–l. Spatial resolution** that can compromise the accurate detection of cell boundaries is displayed using a cell captured with increasing pixel size, i.e., with decreasing spatial resolution: full resolution (j); half resolution (k); one fourth of the original full resolution (l). **m–n. Irregular shape** that can cause over/under-segmentation, especially when the segmentation methods assume simpler, non-touching objects, is displayed using a simulated cell with highly irregular shape under two background noise std situations: 0 (m); 100 (n).This is

especially a problem in high-noise situations (n). **o. High density of cells**, also frequent cause of incorrect segmentation is shown by a cluster of simulated cells. **p–r. Fluorescence temporal decay** that can bring the SNR or CR below detection levels, thus complicating both segmentation and tracking, is simulated by a 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** is key for correctly tracking the cells since many algorithms rely on this overlap. Here it is shown using three simulated cells at the beginning of a video (t=0) (s) and two possible alternative scenarios for the following time point (t=1): t=1 in a scenario of high temporal resolution and/or low cell speed, allowing relatively simple identification of the correspondence between the cells (t); t=1 in a scenario of low temporal resolution and/or high cell speed, complicating the identification of the correspondence between the cells (u). **v–x. Number and synchronization of mitoses** also complicates cell tracking, since tracking a mitotic cell requires correctly assigning the mother to its daughter cells in consecutive frames. This is simulated by cells at the beginning of the video (t=0) (v) and two possible alternative scenarios for the following time point (t=1): t=1 in a scenario where only one of the cell divides asynchronously allowing a simple lineage assignment of mother and daughter cells (w); t=1 in a scenario of multiple, synchronized division events rendering a complicated lineage assignment of mothers and daughters (x).

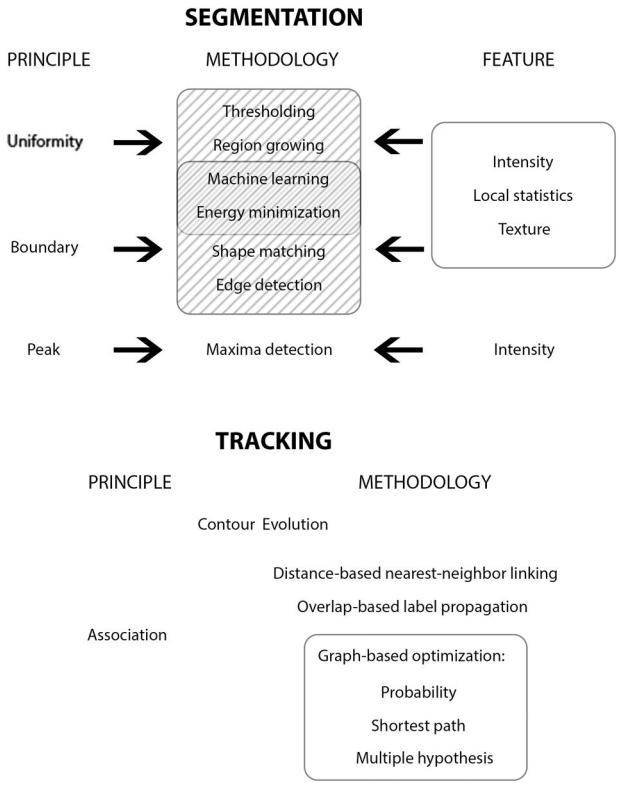


Figure 3.
Taxonomy of cell segmentation and tracking methods.

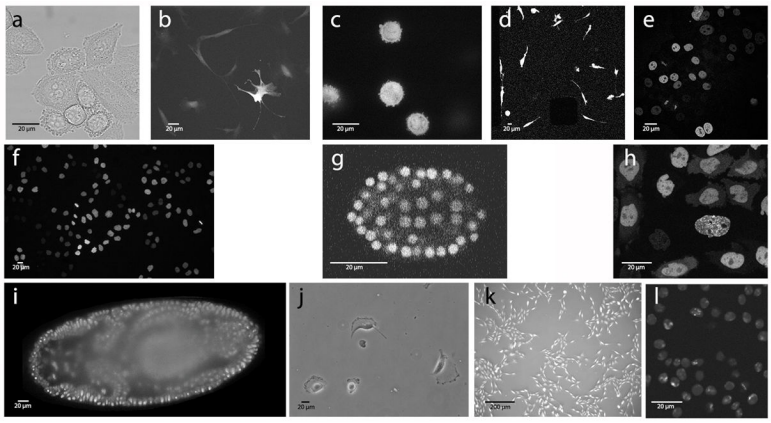


Figure 4. Sample images of the challenge datasets

(**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+.

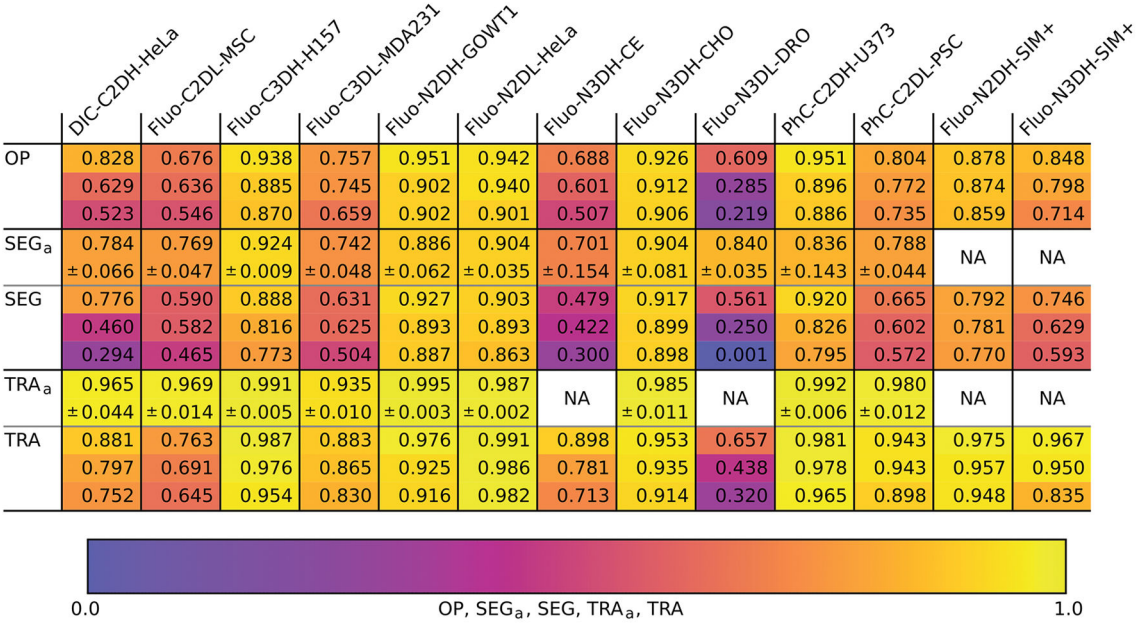
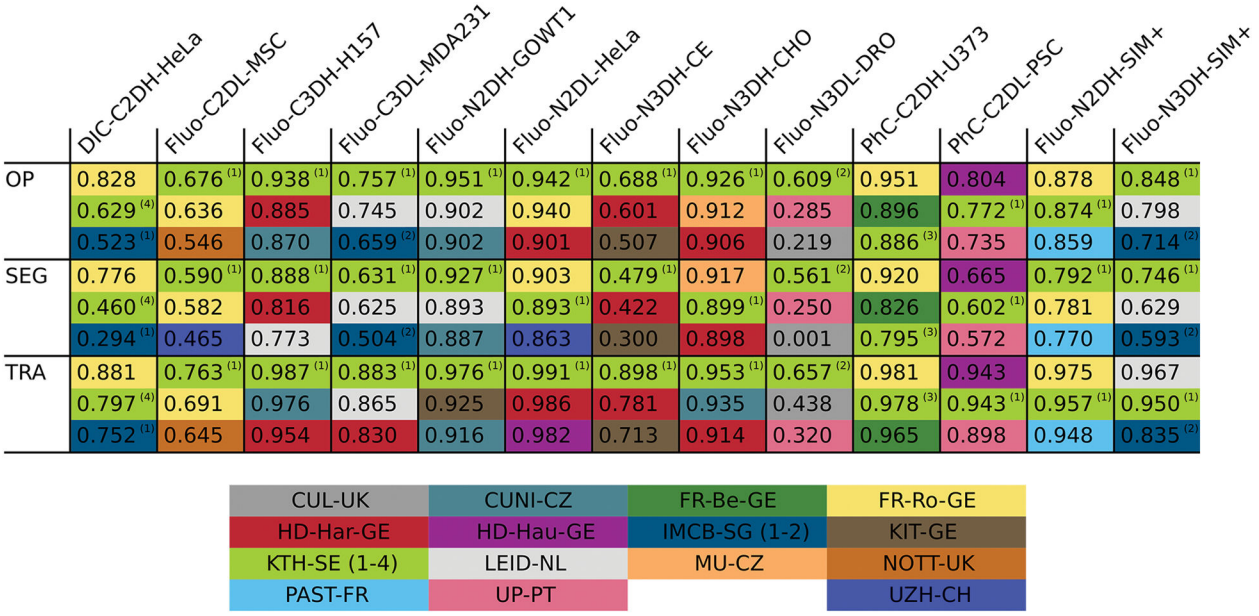


Figure 5.

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	0.792 (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)	0.976 (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)	0.943 (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)

Figure 6.

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.

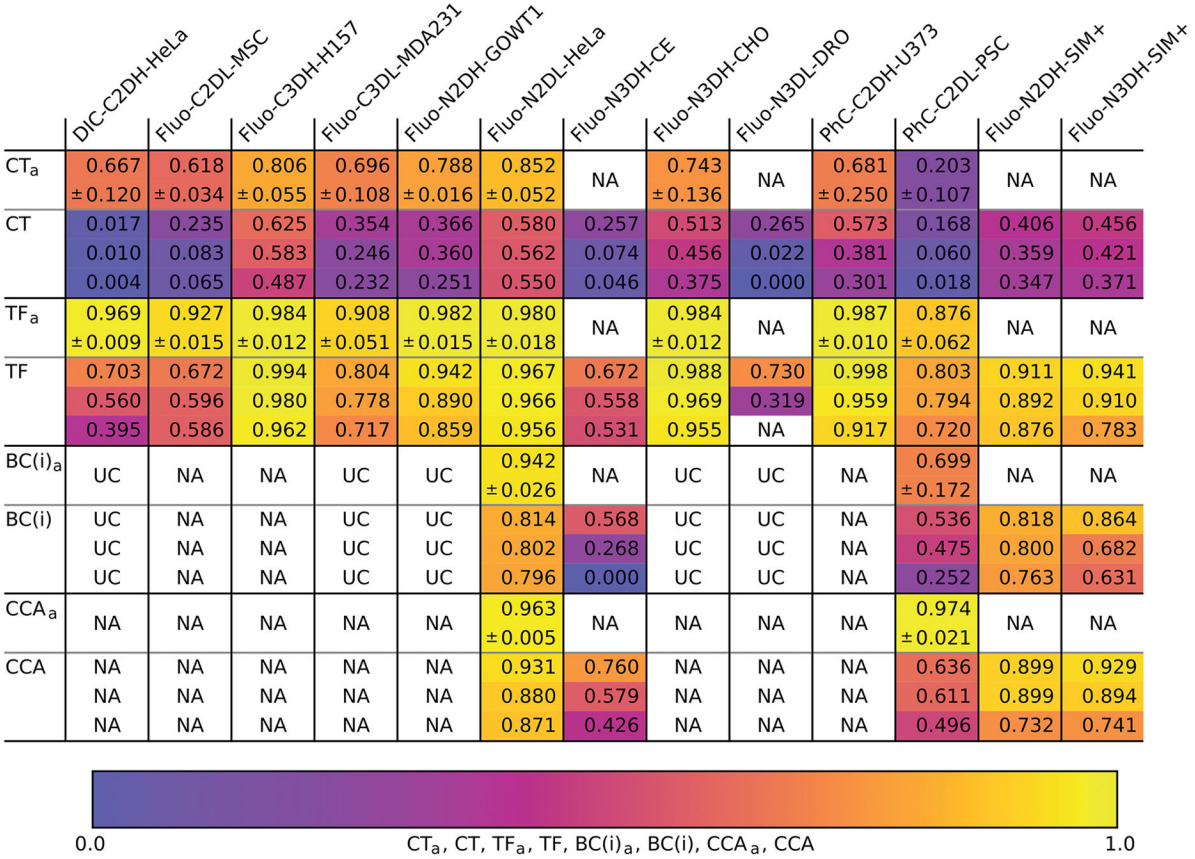


Figure 7.

Top-three biological performance values (**CT**, **TF**, **BC(i),** and **CCA**) measures obtained by the competing algorithms. All four **CT**, **TF**, **BC(i)**, and **CCA** sections start respectively with $\mathbf{CT_a}$, $\mathbf{TF_a}$, $\mathbf{BC(i)_a}$, 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).

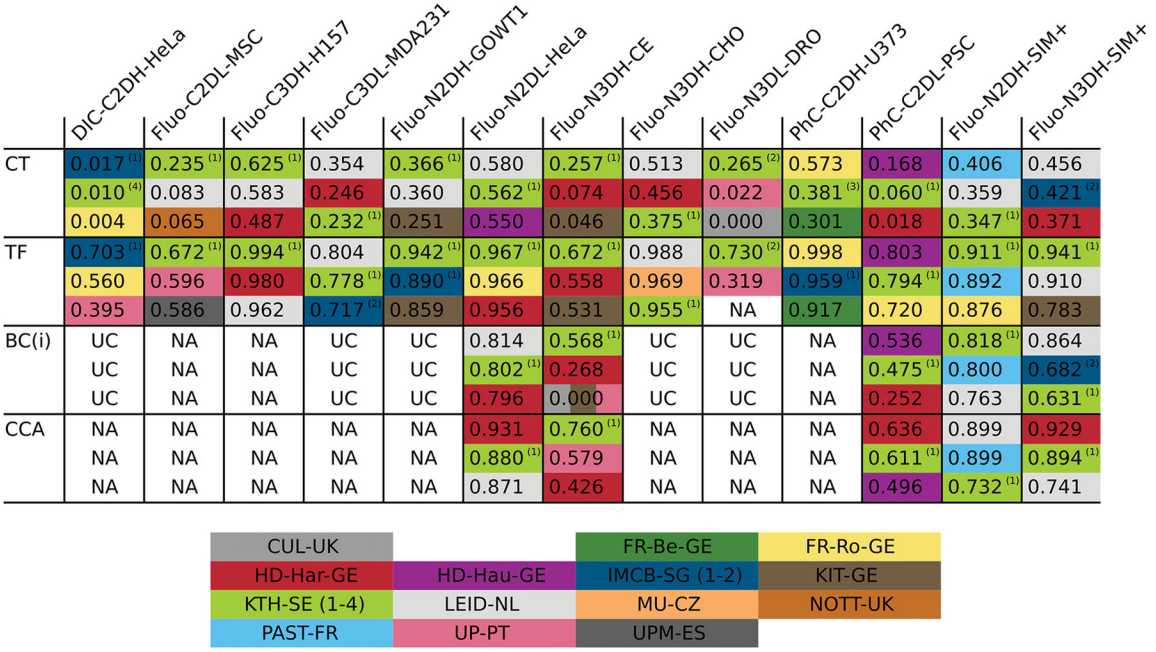


Figure 8.

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).

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.

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	32.8	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	0.00	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	11.31	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	0.01	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	0.00	N	Y	N	Y
PhC-C2DL-PSC	4.06	1.53	0.52	0.34	114	0.60	8.5	0.04	0.90	1.99	N	Y	N	Y
Fluo-N2DH-SIM+	6.30	1.23	0.95	0.48	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 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. 3) along with the preprocessing and postprocessing strategies employed.

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

Algorithm	Preprocessing	Principle	Feature	Methodology	Postprocessing
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 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. 3) along with postprocessing strategies employed, the temporal support given, and the scheme followed for the division detection.

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 4

Usability evaluation of the top-three ranked algorithms based on the overall performance measure

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 of the three methods, yellow means the intermediate value, and green means the best value out of the three listed.

	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	9	0.936	3506
Fluo-N2DH-GOWT1	KTH-SE (1) 0.951			LEID-NL 0.902			CUNI-CZ 0.902		
	17	0.955	632	9	0.932	1333	8	0.950	479
Fluo-N2DL-HeLa	KTH-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	10	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	0.923	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

	1st ranked			2nd ranked			3rd ranked		
	NP	GP	TIM	NP	GP	TIM	NP	GP	TIM
Fluo-N2DH-SIM+	FR-Ro-GE 0.878			KTH-SE (1) 0.874			PAST-FR 0.859		
	3	0.979	20124	17	0.983	301	9	0.978	370
Fluo-N3DH-SIM+	KTH-SE (1) 0.848			LEID-NL 0.798			IMCB-SG (2) 0.714		
	17	0.985	13115	9	0.973	66773	9	0.988	69549