Quantitative evaluation of software packages for single-molecule localization microscopy

Daniel Sage[1], Hagai Kirshner[1], Thomas Pengo[2], Nico Stuurman[3,4], Junhong Min[5], Suliana Manley[6] & Michael Unser[1]

The quality of super-resolution images obtained by single-molecule localization microscopy (SMLM) depends largely on the software used to detect and accurately localize point sources. In this work, we focus on the computational aspects of super-resolution microscopy and present a comprehensive evaluation of localization software packages. Our philosophy is to evaluate each package as a whole, thus maintaining the integrity of the software. We prepared synthetic data that represent three-dimensional structures modeled after biological components, taking excitation parameters, noise sources, point-spread functions and pixelation into account. We then asked developers to run their software on our data; most responded favorably, allowing us to present a broad picture of the methods available. We evaluated their results using quantitative and user-interpretable criteria: detection rate, accuracy, quality of image reconstruction, resolution, software usability and computational resources. These metrics reflect the various tradeoffs of SMLM software packages and help users to choose the software that fits their needs.

We have conducted a large-scale comparative study of software packages developed in the context of SMLM, including recently developed algorithms. We designed realistic data that are generic and cover a broad range of experimental conditions and compared the software packages using a multiple-criterion quantitative assessment that is based on a known ground truth.

Our study is based on the active participation of developers of SMLM software. More than 30 groups have participated so far, and the study is still under way. We provide participants access to our benchmark data as an ongoing public challenge. Participants run their own software on our data and report their list of localized particles for evaluation. The results of the challenge are accessible online and updated regularly.

SMLM was demonstrated in 2006, independently by three research groups[1–3], and has enabled subsequent breakthroughs in diverse fields[4,5]. SMLM can resolve biological structures at the nanometer scale (typically 20 nm lateral resolution), circumventing Abbe's diffraction limit. At the cost of a relatively simple setup[6,7], it has opened exciting new opportunities in life science research[8,9].

The underlying principle of SMLM is the sequential imaging of sparse subsets of fluorophores distributed over thousands of frames, to populate a high-density map of fluorophore positions. Such large data sets require automated image-analysis algorithms to detect and precisely infer the position of individual fluorophore, taking advantage of their separation in space and time.

The acquired data cannot be visualized directly; further computerized image-reconstruction methods are required. These typically comprise four steps: preprocessing, detection, localization and rendering. Preprocessing reduces the effects of the background and noise; detection identifies potential molecule candidates in each frame; localization performs a subpixel refinement of the initial position estimates, usually by fitting a point-spread function (PSF) model; and rendering turns the detected molecule positions into a high-resolution map of molecule densities. The performance of the overall processing pipeline contributes to the quality of the super-resolved image[10].

The current literature describes more than 25 image-analysis software packages that process SMLM data. Each has its own characteristics, set of parameters, accessibility and terminology[10,11]. Moreover, these packages are often validated using different data. In the absence of guidance, end users face a difficult choice in deciding which software is most suitable for them. The lack of a standardized methodology for conducting performance analysis and the need for reference benchmark data constitute the gap that we address in this work.

Our synthetic data imitate microtubule structures. The data consist of thousands of images with labeling densities that span well over an order of magnitude. The model of image formation accounts for the stochastic nature of the emission rate of the fluorophores, the characteristics of the optical setup, and various sources of noise. As in real data, it also includes inhomogeneous excitation, autofluorescence and readout electron-multiplying noise from the detector, typically an electron-multiplying charge-coupled device (EMCCD).

Our benchmark criteria were designed to objectively measure computational performance in terms of time and quality. Our evaluation effort is more comprehensive than previous work[12] in benchmarking a large number of software packages, in synthesizing

[1]Biomedical Imaging Group, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [2]Center for Genomic Regulation, Barcelona, Spain. [3]Howard Hughes Medical Institute, University of California (UCSF), San Francisco, California, USA. [4]Department of Cellular and Molecular Pharmacology, UCSF, San Francisco, California, USA. [5]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea. [6]Laboratory of Experimental Biophysics, EPFL, Lausanne, Switzerland. Correspondence should be addressed to D.S. (daniel.sage@epfl.ch).

RECEIVED 22 AUGUST 2014; ACCEPTED 17 APRIL 2015; PUBLISHED ONLINE 15 JUNE 2015; DOI:10.1038/NMETH.3442

data closer to biological reality, and in including a rich set of evaluation criteria such as detection rate, accuracy, image quality, resolution and software usability.

A byproduct of our work is an extensive and annotated list of software packages (http://bigwww.epfl.ch/smlm/software/), which should prove a resource not only to practitioners but also to developers because it helps identify which aspects of existing software may be in need of further development.

RESULTS

Bio-inspired data

We designed our synthetic data to be as similar as possible to images derived from real cellular structures. A key element is their continuous-domain description, as opposed to a spatially discrete model. For instance, we simulate microtubules by means of three-dimensional (3D) paths that are defined on the continuum (**Fig. 1a**), making it possible to render digital images at any scale. We typically choose a scale of 5 nm per pixel. Our synthetic model takes many parameters into account, among them sample thickness, random activation, laser power, variability of the excitation laser, the lifetime of the fluorophores, autofluorescence, several sources of noise, pixelation, analog-to-digital conversion and the PSF of the microscope (**Fig. 1b**). Our PSF model is made up either of classical Gaussian-based 3D functions or of the more realistic Gibson-Lanni formulation that benefits from a fast and accurate implementation[13]. Because multiple-frame events are rare in the data of interest, we tuned the lifetime model to favor single-frame events. We rely primarily on these ground-truth data for our objective evaluation of algorithms.

To accommodate the intended uses of the available software, we chose to image the same synthetic sample using different imaging modes: long sequence (LS) and high density (HD). The LS data are low-density sequences of about 10,000 frames each, and the HD data are high-density sequences of about 500 frames that include overlapping PSFs (**Fig. 1c**). Independently of the imaging mode, we changed the degree of difficulty of the data by modifying the contribution of autofluorescence, the amount of acquisition noise and the thickness of the sample (see Online Methods) to create datasets LS1-3 and HD1-3 (in order of increasing difficulty).

We generated training data and disclosed them together with the true locations of the fluorophores, allowing participants to tune their software. We also generated contest data and delivered them without ground-truth information. We assessed every algorithm on the basis of the contest data. We make these data available at http://bigwww.epfl.ch/smlm/datasets/; the collection is already used by developers[14–19].

Quantitative assessment metrics

The core task faced by participants in our study is the 2D localization of single molecules. To rate the performance of their software, we defined multiple criteria (Online Methods) that highlight different aspects of SMLM algorithms: detection rate, accuracy, image quality, resolution, usability (USA) and execution runtime (TIME). Other preprocessing or postprocessing steps, such as drift correction and rendering, were excluded from our analysis to better provide an unbiased comparison based primarily on the localization performance.

Detection rate and localization accuracy

The detection rate and localization accuracy are based on the pairing between the molecules localized by the participants and the molecules from the ground truth. These criteria do not depend on any rendering mechanism.

The detection rate quantifies the framewise fidelity and the completeness of the set of localizations with respect to the ground truth, measured in our case by a Jaccard index. We found that the detection rate (JAC) correlates with the level of difficulty.

The localization accuracy (ACC) is measured by the root-mean-square error (RMSE) of matched localizations. We found that this averaged 21.05 nm and 32.13 nm for LS1 and LS2, respectively. This is consistent with the Cramér-Rao lower bound predicted according the definition of uncertainty given by Rieger *et al.*[20]. The detection rate and localization accuracy of each software are documented in **Figure 2** and in **Supplementary Figures 1** and **2**.

Image quality and image resolution

Ultimately, the data representation favored by SMLM practitioners is not a list of localizations but a rendered image[10] (**Supplementary Data 1** and **Supplementary Videos 1–6**).

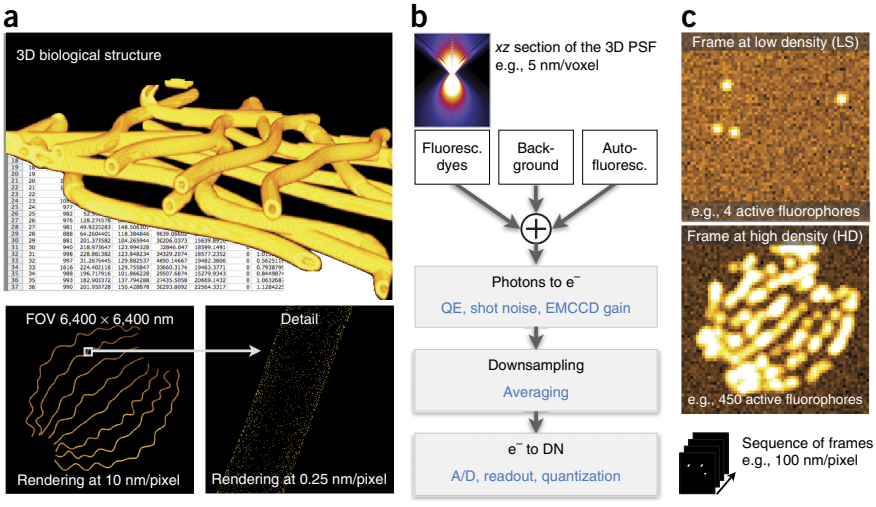

Figure 1 | Construction of the bio-inspired data. (**a**) Top, 3D structure simulating biological microtubules. Every single fluorophore event is uniquely identified and stored; collectively, they constitute the ground-truth localizations which can be rendered at any temporal and spatial scale (lower panels). (**b**) Each fluorophore is considered a point source and convolved with a 3D PSF. Combined with background and autofluorescence of the structure, the convolved image determines the number of photons at each pixel. These photons are then transformed into a number of electrons based on quantum efficiency (QE), shot noise and the EMCCD parameters. The image is reduced to the desired camera resolution, for example, 100 nm/pixel. Finally, these values are fed to an electron-to-DN converter (digital number, taking into account the readout noise and the quantization level). (**c**) These operations are repeated to obtain long sequence (LS) of low-density frames or short sequence of high-density (HD) frames.

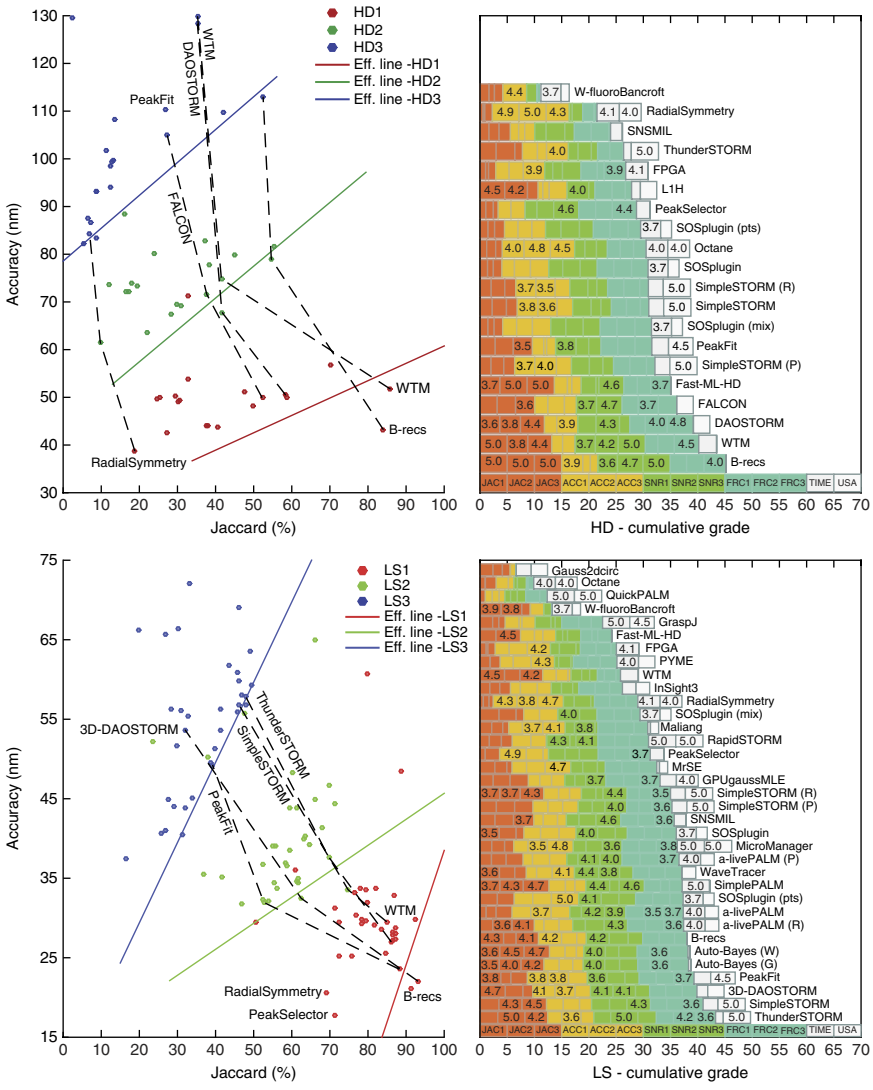

Figure 2 | Accuracy versus detection rate for each tested software. Scatter plots show high-density (HD) data above and long sequence data below. Efficiency lines (Eff. lines) are computed from the five results at the boundary of the field with high JAC and/or low ACC. The length of the bars is proportional to the grade, from 0 (poor) to 5 (good). Grades above 3.5 are written in the corresponding bar. The grades of the three data sets are given here for the detection rate, JAC1–JAC3; for the localization accuracy, ACC1–ACC3; for the image quality assessment, SNR1–SNR3; and for the image resolution, FRC1–FRC3. The grades of the computational time (TIME) and usability (USA) are reported in light gray bars.

We used two image-based criteria in our assessment: image quality (signal-to-noise ratio, SNR) and image resolution (Fourier ring correlation, FRC[21]). Methods afflicted by issues such as sampling artifacts or a low detection capacity at the image border are characterized by a low SNR. Conversely, a high SNR is often indicative of a successful tradeoff between detection rate and accuracy.

Software efficiency

In a retrospective analysis, we identified the five best methods, in terms of the tradeoff between accuracy and detection rate for each dataset. We defined a linear regression that fits the best methods in a plot of ACC versus JAC, and call it an efficiency line (**Fig. 3**). The distance of the (JAC, ACC) coordinate for each software to such a line indicates the performance of the software.

The level of difficulty increases from LS1 to LS3, as evidenced by the average performance (JAC, ACC), which was (79.58%, 29.98 nm) for LS1, (55.64%, 41.91 nm) for LS2 and (35.64%, 55.82 nm) for LS3. These findings are consistent with our engineering of the data to have increasing levels of noise, as the theory predicts that the presence of noise leads to an increase in the uncertainty of the location of a particle. Likewise, the detection rate is also affected by noise; single molecules with lower emission rate and deeper axial position are more difficult to detect.

Algorithms

Our study includes more than 30 packages (**Table 1**), covering a large proportion of the SMLM software currently available. Aside from a few that do not fit our validation framework because their SMLM reconstruction is based on deconvolution without explicit localization[22], most packages have a similar architecture. However, a detailed analysis reveals fundamental differences.

Within the detection step, methods as diverse as low-pass filtering, band-pass filtering, watershed, and wavelet transform, to name a few, are deployed. The parameters of these preprocessing operations need to be determined in an *ad hoc* fashion. In some cases, we found that they cannot be set by the user; even when they can be, often there is no calibration procedure provided. Most algorithms isolate candidate pixels by applying a threshold to identify potential local extrema, but each software uses different methods for determining the threshold value: level of noise, spot brightness, PSF size and/or particle density.

Over two-thirds of the participating packages carry out the localization step by means of a fitting with a Gaussian function. Other algorithms use an arbitrary PSF instead; DAOSTORM and SimpleSTORM use a measured PSF. Distinctively, the two packages MrSE and RadialSymmetry exploit the radial symmetry of the PSF.

We have identified three groups of localization methods and indicated their performance in **Table 2**. In Generation 1, the basic methods perform localization by means of center of mass (QuickPALM), triangularization (fluoroBancroft) or linear regression (Gauss2dcir). Although very fast, these methods often fail to reconstruct HD data. Generation 2 is the largest group of methods, including about two-thirds of all softwares submitted thus far. They are characterized by the use of iterative localization algorithms such as maximum-likelihood estimators (MLE) or least-squares minimizers (LS). Previous works compare the LS or MLE algorithm in detail[10,23]. Generation 3 comprises advanced methods, often unpublished. They improve

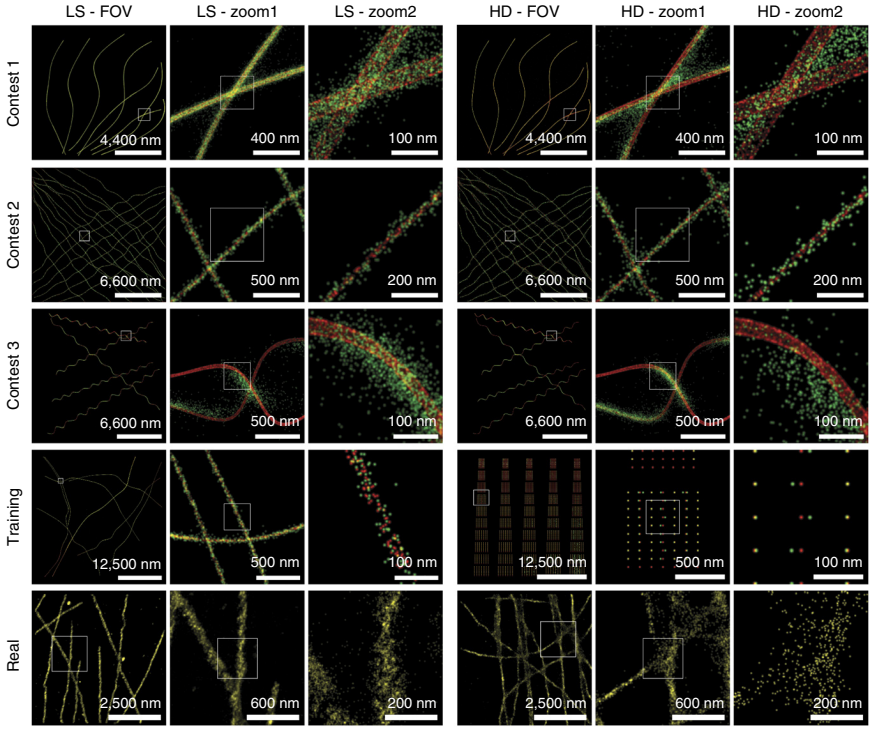

Figure 3 | Rendering of software results versus ground truth at various scales. Every participant in the challenge received a detailed report on the performance of their software, including renderings as shown; the particular instance here corresponds to the PeakFit software. Long-sequence data (LS), columns 1–3: full field of view (FOV), medium (zoom1), and high (zoom2) magnification. High-density data (HD), columns 4–6. The white frames in FOV indicate the regions displayed in zoom1, while the frames in zoom1 are themselves expanded in zoom2. Rows 1–4: simulated data. The red channel represents the rendering of the ground truth and the green channel the localizations of the tested software. Row 5: real data with unknown ground truth.

the detection rate while keeping a high localization accuracy. This group includes minimum mean squared error (MMSE)/maximum *a posteriori* probability (MAP) approaches (B-recs), a method with high-quality interpolation (simpleSTORM), a template-matching technique (WTM), a mean-shift approach (simplePALM) and packages that exploit the radial symmetry (RadialSymmetry and MrSE). Detailed information on the software packages is in **Supplementary Notes 1** and **2**.

Usability and computation time

End users require that software packages be accessible, easy to use and fast. Although these aspects are subjective, they are important enough to justify their inclusion in our study. To score them, we prepared a questionnaire for the participants. We combined the accessibility score with a usability score that covers quality of documentation and user-friendliness. The open-source software ImageJ/Fiji and the versatile platform Matlab are the most highly represented frameworks hosting SMLM packages.

Finding the accurate position of millions of fluorophores is a heavy computational task. We observed that the four packages that use specialized hardware accelerators (a graphics processor unit, GPU, or field-programmable gate array, FPGA) reduce their runtimes by an order of magnitude, sometimes reconstructing a super-resolution image in less than a minute.

Benchmarking reporting and ranking

We returned to every participant a benchmark report that includes renderings at different scales (**Fig. 3**) and quantitative measures (**Fig. 4**). In particular, the bottom left curve of **Figure 4** illustrates how the proximity of fluorophores, d_{NN}, influences the performance of the software. In this specific case, the rate of detection improves from about half—when d_{NN} is below the FWHM of the PSF—to near perfection when d_{NN} is sufficient high.

To coalesce our six criteria for a single ranking, we computed the final score as the weighted sum of relative grades from 0 to 5, as presented in **Table 2**. We gave a greater weight to the objective criteria JAC, ACC, SNR and FRC than to the subjective criteria USA and TIME. With our particular choice of weights, the ranking for the LS data is as follows, starting from the best results: ThunderSTORM, SimpleSTORM and PeakFit. For the HD data, it is B-recs, WTM and DAOSTORM.

DISCUSSION

The accuracy of single-molecule localization has a direct impact on the resolving power of the reconstructed image. We confirmed in this study of SMLM software packages that the experimental accuracy is one order of magnitude better than the classical diffraction limit, which supports theoretical findings[24,25]. This is the best one can hope for; indeed, a few software packages nearly achieve the Cramér-Rao lower bound.

Notwithstanding its popularity, the accuracy measure may still misrepresent performance. For instance, it does not capture issues related to the spread of the localizations—too few accurate ones, for example, or too many false positives. To avoid reliance on accuracy alone, we therefore considered additional criteria such as the detection rate, which describes the overlap between the set of detected molecules and the set of true molecules, along with a measure of the quality of the rendered image and a measure of its resolution.

Accuracy and detection rate tend to be in opposition—the average accuracy of localization can often be made to artificially increase just by excluding those unreliable molecules that emit a low number of photons. It is therefore enlightening to quantify the tradeoff between accuracy and detection. This idea has led us to propose the efficiency lines or curves (**Fig. 2**), which should aid microscopy practitioners in selecting software by better allowing them to judge if a particular software will help them meet their own preferred tradeoff.

We proposed a combination of six simple metrics to help users choose an SMLM software package. While no single measure of performance can capture the complexity of this choice, our goal with the combined criterion is to provide guidance to practitioners that is balanced and fair.

Although the correlation (CORR) between the number of photons of the ground truth and the number estimated by the tested software is a parameter of interest, only a few participants provided us with relevant output to obtain these correlations. We therefore decided to exclude CORR from the final score but have encouraged developers to focus their efforts on improving accessibility and usability and to provide an estimate of the number of photons or the uncertainty of measurements for future releases. Also, we did not assess the grouping of multiple-frame emission from a single molecule, as this is often carried out at the postprocessing stage.

All packages we studied require parameters from the user. Unfortunately, choosing appropriate values is by no means easy or straightforward. More often than not, the tuning of parameters requires a deep knowledge of the algorithmic pipeline; inexperienced users may find that they need to invest a lot of time before they can obtain satisfactory results. For this study, to ensure that each software was properly tuned to our simulated database,

Table 1 | Description of SMLM software

Software	Molecule detection	PSF	Method	Platform	Acc.	Affiliation
3D-DAOSTORM[28]	Adaptive threshold—update on residual images	Gauss	LS	Python	+	Harvard Univ., USA
a-livePALM[29]	Denoising, SNR threshold, adaptive histogram equalization	Gauss	MLE	Matlab	+	Karlsruhe IT, Germany
Auto-Bayes	Generalized minimum-error threshold (GMET), local maximum	Gauss, Weibull	LS	Stand-alone	+	NCNST, Beijing, China
B-recs	Detection: n/a; fit: Bayesian inference framework	Arbitrary	MMSE, MAP	Stand-alone	–	Janelia Farm, HHMI, USA
CSSTORM[30]	No explicit localization; convex optimization problem (HD)	Gauss	Compressed sensing	Matlab	+	UCSF, USA
DAOSTORM[31]	Gaussian filtering, local maximum (HD)	Measured,	LS	Python	+	Univ. Oxford, UK
FacePALM[32]	No explicit localization; background estimation	arbitrary	–	Python	–	Univ. Amsterdam, the Netherlands
FALCON[33]	Deconvolution with sparsity prior, local maximum (HD)	Taylor approx.	ADMM	Matlab	+	KAIST, Daejeon, Republic of Korea
Fast-ML-HD[34]	Sparsity constraint, concave-convex procedure (HD)	Gauss	MLE	Matlab	–	KAIST, Daejeon, Republic of Korea
FPGA[35]	Adaptive threshold	Gauss	MLE, CoMass	Stand-alone	–	Univ. Heidelberg, Germany
Gauss2DCirc[36]	Fixed SNR threshold	Gauss	REG	Matlab	+	Univ. Illinois, USA
GPUgaussMLE[37]	Simple (unspecified) methods to select subregions	Gauss	MLE	Matlab	+	TU Delft, Delft, the Netherlands
GraspJ[38]	Peak finding: fixed threshold value	Gauss	MLE	ImageJ	+	ICFO, Barcelona, Spain
Insight3	Low-pass filtering, local maximum	Arbitrary	LS	Stand-alone	–	UCSF, USA
L1H[39]	No explicit localization; L1 homotopy, FIST deconvolution	Gauss, arbitrary	Compressed sensing	Python	+	Harvard Univ., USA
M2LE[40]	Adaptive threshold	Gauss	MLE	ImageJ	+	Cal Poly Pomona, USA
Maliang[41]	Annular averaging filters, denoising by convolution	Gauss	MLE	ImageJ	+	WUST, Wuhan, China
Micro-Manager LM	Adaptive threshold	Gauss	LS	ImageJ	+	UCSF, USA
MrSE[42]	Band-pass filtering, local maximum	Radial	CoSym	Stand-alone	–	WUST, Wuhan, China
Octane[43]	Watershed maximum	Gauss	LS	ImageJ	+	Univ. Connecticut, USA
PeakFit	Band-pass filtering, local maximum	Gauss	LS	ImageJ	+	Univ. Sussex, UK
PeakSelector[44]	Time-domain filtering, adaptive threshold	Gauss	LS	IDL, Matlab	–	HHMI, USA
PYME[27]	Wiener filtering, adaptive threshold	Arbitrary	LS	Python	+	Univ. Auckland, New Zealand
QuickPALM[45]	Band-pass filtering, fixed SNR threshold	Gauss	CoMass	ImageJ	+	Institut Pasteur, France
RadialSymmetry[46]	Filtering, local max., minimal distance to gradient	Radial	CoSym	Matlab	+	Univ. Oregon, Eugene, USA
rapidSTORM[12]	Low-pass filtering, local maximum	Gauss	LS, MLE	Stand-alone	+	Univ. Würzburg, Germany
SimplePALM[47]	Variance stabilization denoising, DoG, probabilistic threshold	n/a	Mean-shift	Stand-alone	–	Molecular Genetics Center, Gif-sur-Yvette, France
simpleSTORM[14]	Self-calibration, noise normalize, background subtraction, *P* value	Gauss, measured	Interpolation	Stand-alone	+	Univ. Heidelberg, Germany
SNSMIL	Gaussian filtering, fixed contrast threshold	Gauss	LS	Stand-alone	+	NCNST, Beijing, China
SOSplugin	Wavelet transform, local maximum, Gaussian mixture	Gauss	LS	ImageJ	+	Erasmus MC, Rotterdam, the Netherlands
ThunderSTORM[15]	Extensive collection of methods, preview, filtering, local maximum	Gauss	LS, MLE	ImageJ	+	Charles Univ., Prague, Czech Republic
W-fluoroBancroft[48]	Wavelet, adaptive threshold	Gauss	fB	Matlab	+	Boston Univ., USA
WaveTracer[49]	Wavelet, watershed maximum	Gauss	LS	Metamorph	–	Univ. Bordeaux, France
WTM[50]	Wedge template matching (HD)	Wedge	Match.	Stand-alone	–	Hamamatsu Photononics, Japan

The software packages whose manufacturers participated in our study are listed. The study is ongoing, and this list will be updated at http://bigwww.epfl.ch/smlm/software/. Software marked 'ImageJ' runs on compatible products ImageJ, Fiji, Icy and ImageJ2. Abbreviations for PSF: Gauss, Gaussian, elliptical Gaussian or averaged Gaussian. Abbreviations for methods: ADMM, alternating direction method of multipliers; CoMass, center of mass; CoSym, center of symmetry; fB, fluoroBancroft; LS, least-squares; MAP, maximum a posteriori; MLE, maximum-likelihood estimator; MMSE, minimum mean-square error; REG, regression. Abbreviations regarding open access: +, available online (sometimes upon request); –, not available or included in commercial package.

we encouraged the developers themselves to run their own software, guided by the training data. To alleviate the difficulty of presetting parameters, we suggest that developers incorporate self-calibration capabilities or dynamically tune the parameters[26]. We predict that this will be one of the key factors determining the success of future software[14].

Table 2 | Quantitative comparison for long-sequence (LS) and high-density (HD) data

	Physical values					Grades							
Criteria	**JAC**	**ACC**	**SNR**	**FRC**	**CORR**	**JAC**	**ACC**	**SNR**	**FRC**	**USA**	**TIME**	**Score**	**Run by**
LS data													
Unit	%	nm	dB	nm	%	[0..5]	[0..5]	[0..5]	[0..5]	[0..5]	[0..5]	[0..5]	
Weights						1	1	1	1	0.25	0.25		
3D-DAOSTORM	62.5	36.1	6.0	71.1	n/a	3.58	3.67	4.20	3.03	3.00	1.98	3.44	Author
a-livePALM	54.5	35.6	5.1	62.6	n/a	2.02	3.74	3.13	3.71	2.50	3.97	3.09	Author
Auto-Bayes (W)	**69.1**	45.4	5.6	66.7	n/a	4.83	2.00	3.72	3.38	0.00	0.41	3.05	Author
B-recs	64.2	40.4	5.9	74.1	51.2	3.91	2.89	4.07	2.78	0.00	0.00	2.96	Author
Fast-ML-HD	52.9	44.9	3.7	83.9	60.6	1.72	2.08	1.40	2.00	0.00	0.00	1.43	Author
FPGA	47.8	36.3	4.2	81.3	62.8	0.73	3.63	2.01	2.21	0.00	4.11	2.00	Author
Gauss2dcirc	53.8	71.5	1.1	143.2	64.0	1.90	0.00	0.00	0.00	3.00	2.75	0.69	Expert
GPUgaussMLE	60.9	44.0	4.4	65.5	2.8	3.26	2.25	2.26	3.48	4.00	3.00	2.78	Author
GraspJ	51.7	47.2	4.4	77.2	n/a	1.48	1.67	2.22	2.54	4.50	5.00	2.13	Author
InSight3	53.9	41.8	4.1	74.5	53.5	1.90	2.64	1.91	2.75	2.50	2.50	2.19	Author
Maliang	53.3	35.6	4.4	69.5	60.4	1.80	3.75	2.22	3.16	1.50	0.50	2.44	Author
MicroManager	55.3	34.5	4.9	64.3	57.7	2.18	3.95	2.87	3.57	5.00	5.00	3.28	Author
MrSE	54.6	34.8	4.5	67.6	n/a	2.05	3.89	2.35	3.31	1.50	0.50	2.60	Author
Octane	42.7	53.9	3.2	114.4	n/a	0.00	0.48	0.84	0.00	4.00	3.99	0.62	Expert
PeakFit	60.0	34.9	5.5	64.6	59.5	3.09	3.87	3.61	3.55	4.50	3.07	3.51	Author
PeakSelector	49.8	40.3	5.2	66.9	59.4	1.11	2.91	3.18	3.36	0.00	2.50	2.39	Expert
PYME	48.6	36.5	3.5	73.8	n/a	0.88	3.58	1.16	2.81	3.00	3.97	2.13	Author
QuickPALM	41.9	50.6	3.5	95.4	57.7	0.00	1.05	1.21	1.08	5.00	5.00	1.14	Author
RadialSymmetry	47.3	**31.0**	4.4	74.4	n/a	0.64	4.57	2.21	2.76	4.00	4.11	2.61	Author
RapidSTORM	54.7	45.4	5.5	68.4	61.7	2.06	1.99	3.60	3.25	5.00	5.00	2.88	Author
SimplePALM	68.8	44.4	5.8	79.1	n/a	4.79	2.17	3.96	2.39	0.00	5.00	3.15	Author
SimpleSTORM	67.9	40.8	5.6	66.3	n/a	4.61	2.81	3.64	3.41	5.00	2.74	3.59	Author
SNSMIL	63.0	45.3	5.0	66.0	n/a	3.66	2.01	3.03	3.44	0.00	2.17	2.73	Author
SOSplugin	59.2	37.8	5.7	70.0	n/a	2.94	3.35	3.77	3.11	2.00	3.69	3.18	Author
ThunderSTORM	68.6	40.0	6.0	**60.8**	46.5	4.75	2.97	4.14	3.85	5.00	1.41	**3.81**	Author
W-fluoroBancroft	61.9	56.5	1.5	113.8	n/a	3.45	0.00	0.00	0.00	1.50	3.70	1.02	Author
WaveTracer	60.0	38.8	**6.1**	69.5	n/a	3.08	3.17	4.28	3.16	2.50	0.00	3.12	Author
WTM	66.0	47.6	4.0	89.7	60.5	4.24	1.61	1.83	1.54	0.00	3.28	2.08	Author
Average on LS	57.0	42.6	4.6	77.7									
Gauss		43.6	4.6	78.1									
Radial		32.9	4.4	71.0									
Generation 1		59.5	2.0	117.5									
Generation 2		40.8	4.9	72.1									
Generation 3		39.9	5.0	75.2									
HD data													
B-recs	**63.7**	78.4	**4.4**	93.2	20.2	5.00	1.76	4.44	3.86	0.00	0.00	**3.35**	Author
DAOSTORM	45.1	82.2	3.9	**78.5**	27.5	3.99	0.91	3.52	4.60	3.00	0.00	3.06	Author
FALCON[a]	39.1	75.5	3.9	99.2	17.3	3.38	2.42	3.49	3.56	3.00	0.00	3.02	Author
Fast-ML-HD	52.1	80.3	3.4	104.6	19.7	4.70	1.33	2.53	3.29	0.00	0.00	2.63	Author
FPGA	14.3	70.2	3.0	104.9	30.1	0.89	3.59	1.68	3.27	0.00	4.11	2.32	Author
L1H	42.5	76.9	3.6	136.1	n/a	3.73	2.10	2.81	1.71	3.00	1.62	2.56	Author
Octane	18.2	**62.9**	3.0	124.0	n/a	1.28	5.00	1.83	2.32	4.00	3.99	2.76	Expert
PeakFit	37.3	81.5	3.9	105.9	11.9	3.20	1.07	3.39	3.22	4.50	3.07	2.84	Author
PeakSelector	15.7	84.4	3.5	87.8	n/a	1.03	0.43	2.78	4.13	0.00	2.50	2.00	Expert
RadialSymmetry	11.8	61.5	2.0	164.2	n/a	0.64	5.00	0.00	0.30	4.00	4.11	1.77	Author
SimpleSTORM	27.9	70.5	3.6	131.8	n/a	2.26	3.54	2.90	1.92	5.00	2.74	2.79	Author
SNSMIL	23.3	80.8	3.3	143.2	n/a	1.80	1.23	2.29	1.35	0.00	2.17	1.60	Author
SOSplugin	18.2	71.8	3.6	106.2	n/a	1.28	3.23	2.85	3.21	2.00	3.69	2.67	Author
ThunderSTORM	31.7	70.9	2.9	154.2	15.7	2.64	3.44	1.67	0.80	5.00	1.41	2.26	Author
W-fluoroBancroft	19.2	80.9	0.8	203.5	n/a	1.38	1.19	0.00	0.00	1.50	3.70	0.86	Author
WTM	54.2	85.5	4.4	91.2	21.9	4.91	0.17	4.36	3.96	0.00	3.28	3.16	Author
Average on HD	32.2	75.9	3.3	120.5									

[a]Software version under development.

Performance measures for the indicated software packages are shown. JAC, Jaccard index; ACC, localization error; SNR, image quality; FRC, image resolution; USA, usability; TIME, computational time. Bold numbers indicate top scorers. The correlation of the estimated number of photons (CORR) was excluded from our analysis but is given here for the sake of completeness. The relative grades are normalized on a scale from 0 (worst) to 5 (best). The score is a weighted sum of the six criteria; here, the weights are 1 for the four quantitative criteria, 0.25 for the usability, and 0.25 for the computational time.

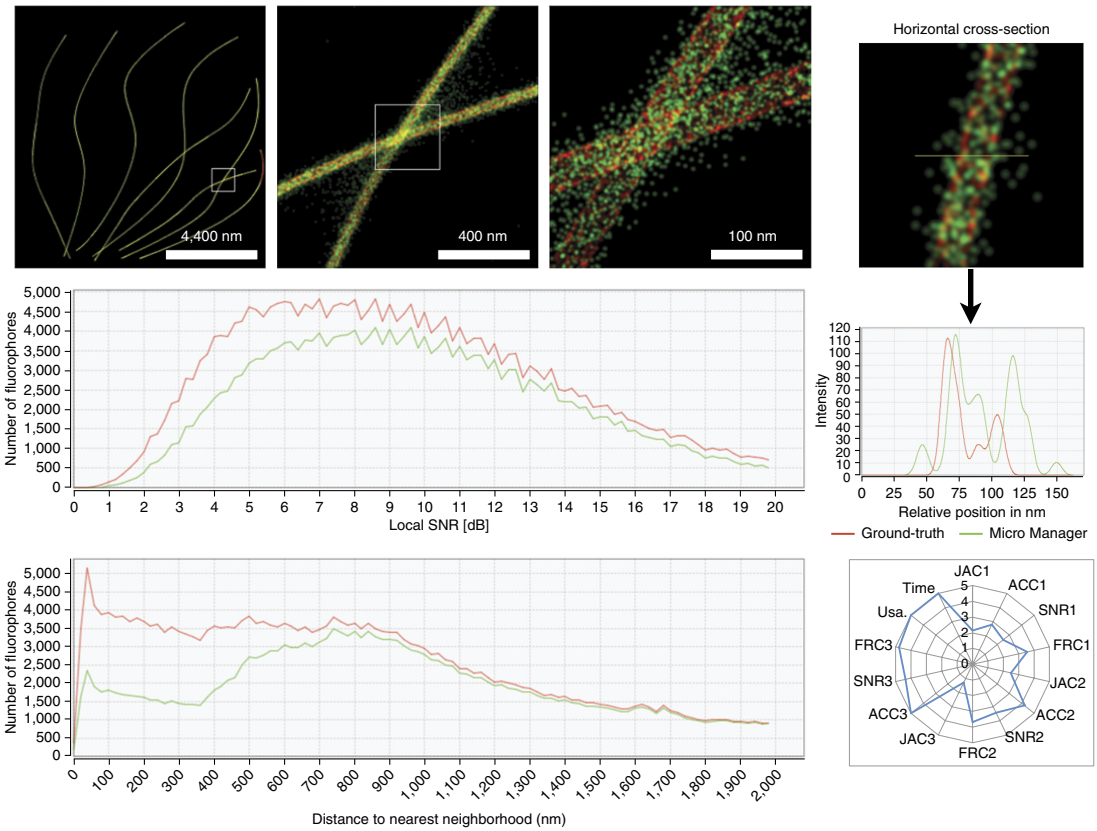

Figure 4 | Illustration of an assessment report. Every participant to the challenge received a detailed report including figures and plots as shown; the particular instance shown here corresponds to the MicroManager software. Three top left images: renderings (see the caption of **Fig. 3** for explanations). Middle left plot, distribution of the local SNR_f in the range 0 dB–20 dB; the green and the red curves correspond to the evaluated software and the ground truth, respectively. Bottom left plot, distribution of the distance to the nearest-neighbor d_{NN} in the range 0 nm–2,000 nm, same color conventions. Top right image, cross-section. Middle right plot, intensity profile along the yellow line seen in the cross-section. Lower right, radar plot of the grades.

The simulated ground-truth data used for our comparison remain accessible to future participants. We pledge to extend this study with new results as they become available and to enrich our collection of data. We plan to include additional features such as several levels of molecule density, 3D (PSF engineering and multiple planes), drift and various noise models for EMCCD cameras and sCMOS (scientific complementary metal-oxide-semiconductor) cameras.

We encouraged all participants to produce output data in common formats to facilitate interoperability and to promote independent rendering software[27] (https://github.com/PALMsiever/). A first step in this direction was taken by many participants in the IEEE International Symposium on Biomedical Imaging 2013 (ISBI 2013) challenge.

Our study has shown that a simple Gaussian PSF model is sufficiently accurate for low-density data, whereas the quality of high-density imaging depends strongly on the model of the PSF. We predict that the PSF model will have an even more significant role in 3D SMLM applications. We see great potential in a two-phase reconstruction workflow—a first reconstruction that is fast but has reduced performance, followed by a slower run that yields improved results.

METHODS

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

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

ACKNOWLEDGMENTS

We thank N. Olivier for providing the experimental data and R. Nieuwenhuizen for his technical assistance in running the Fourier ring correlation. We thank also P. Thévenaz for critical reading and for his assistance in writing the manuscript. We thank the participants in the ISBI 2013 localization microscopy challenge: S. Anthony, S. Andersson, T. Ashley, D. Baddeley, K. Bennett, J. Boulanger, N. Brede, L. Dai, L. Fiaschi, F. Gruell, G. Hagen, R. Henriques, A. Herbert, S. Holden, E. Hoogendoorn, B. Huang, Z.-L. Huang, A. Kechkar, K. Kim, M. Kirchgessner, U. Koethe, P. Krizek, M. Lakadamyali, Y. Li, K. Lidke, R. McGorty, L. Muresan, R. Parthasarathy, B. Rieger, H. Rouault, M. Sauer, J.-B. Sibarita, I. Smal, A. Small, S. Stahlheber, Y. Tang, Y. Wang, S. Watanabe, S. Wolter, J.C. Ye and C. Zimmer. This work was supported by the Biomedical Imaging Group, the School of Engineering at the Ecole Polytechnique Fédérale de Lausanne, the European Research Council (ERC) FUN-SP project (267439), the ERC Starting Grant PALMassembly (243016) and the Eurobioimaging Project (WP11).

AUTHOR CONTRIBUTIONS

D.S., H.K., T.P., J.M. and N.S. conceived the project. D.S. developed the project and organized the challenge with contribution from all authors. D.S. and H.K. wrote the code for the simulated data and analyzed the results. S.M. and M.U. directed the project. D.S. and H.K. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Betzig, E. *et al.* Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313**, 1642–1645 (2006).
2. Rust, M.J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3**, 793–795 (2006).
3. Hess, S.T., Girirajan, T.P. & Mason, M.D. Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophys. J.* **91**, 4258–4272 (2006).
4. Balzarotti, F. & Stefani, F.D. Plasmonics meets far-field optical nanoscopy. *ACS Nano* **6**, 4580–4584 (2012).
5. Walder, R., Nelson, N. & Schwartz, D.K. Super-resolution surface mapping using the trajectories of molecular probes. *Nat. Commun.* **2**, 515 (2011).
6. Manley, S., Gunzenhäuser, J. & Olivier, N. A starter kit for point-localization super-resolution imaging. *Curr. Opin. Chem. Biol.* **15**, 813–821 (2011).
7. Schermelleh, L., Heintzmann, R. & Leonhardt, H. A guide to super-resolution fluorescence microscopy. *J. Cell Biol.* **190**, 165–175 (2010).
8. Sauer, M. Localization microscopy coming of age: from concepts to biological impact. *J. Cell Sci.* **126**, 3505–3513 (2013).
9. Moerner, W.E. New directions in single-molecule imaging and analysis. *Proc. Natl. Acad. Sci.* **104**, 12596–12602 (2007).
10. Small, A. & Stahlheber, S. Fluorophore localization algorithms for super-resolution microscopy. *Nat. Methods* **11**, 267–279 (2014).
11. Endesfelder, U. & Heilemann, M. Art and artifacts in single-molecule localization microscopy: beyond attractive images. *Nat. Methods* **11**, 235–238 (2014).
12. Wolter, S., Endesfelder, U., van de Linde, S., Heilemann, M. & Sauer, M. Measuring localization performance of super-resolution algorithms on very active samples. *Opt. Express* **19**, 7020–7033 (2011).
13. Kirshner, H., Aguet, F., Sage, D. & Unser, M. 3-D PSF fitting for fluorescence microscopy: implementation and localization application. *J. Microsc.* **249**, 13–25 (2013).
14. Köthe, U., Herrmannsdoerfer, F., Kats, I. & Hamprecht, F.A. SimpleSTORM: a fast, self-calibrating reconstruction algorithm for localization microscopy. *Histochem. Cell Biol.* **141**, 613–627 (2014).
15. Ovesný, M., Krizek, P., Borkovec, J., Svindrych, Z. & Hagen, G.M. ThunderSTORM: a comprehensive ImageJ plugin for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics* **30**, 2389–2390 (2014).
16. Ovesný, M., Krízek, P., Svindrych, Z. & Hagen, G.M. High density 3D localization microscopy using sparse support recovery. *Opt. Express* **22**, 31263–31276 (2014).
17. Ma, H., Kawai, H., Toda, E., Zeng, S. & Huang, Z.-L. Localization-based super-resolution microscopy with an sCMOS camera part III: camera embedded data processing significantly reduces the challenges of massive data handling. *Opt. Lett.* **38**, 1769–1771 (2013).
18. Wang, Y. *et al.* Localization events-based sample drift correction for localization microscopy with redundant cross-correlation algorithm. *Opt. Express* **22**, 15982–15991 (2014).
19. Mandula, O., Šestak, I.Š., Heintzmann, R. & Williams, C.K. Localisation microscopy with quantum dots using non-negative matrix factorisation. *Opt. Express* **22**, 24594–24605 (2014).
20. Rieger, B. & Stallinga, S. The lateral and axial localization uncertainty in super-resolution light microscopy. *ChemPhysChem* **15**, 664–670 (2014).
21. Nieuwenhuizen, R.P.J. *et al.* Measuring image resolution in optical nanoscopy. *Nat. Methods* **10**, 557–562 (2013).
22. Mukamel, E.A., Babcock, H. & Zhuang, X. Statistical deconvolution for superresolution fluorescence microscopy. *Biophys. J.* **102**, 2391–2400 (2012).
23. Abraham, A.V., Ram, S., Chao, J., Ward, E.S. & Ober, R.J. Quantitative study of single molecule location estimation techniques. *Opt. Express* **17**, 23352–23373 (2009).
24. Thompson, R.E., Larson, D.R. & Webb, W.W. Precise nanometer localization analysis for individual fluorescent probes. *Biophys. J.* **82**, 2775–2783 (2002).
25. Ober, R.J., Ram, S. & Ward, E.S. Localization accuracy in single-molecule microscopy. *Biophys. J.* **86**, 1185–1200 (2004).
26. Holden, S.J. *et al.* High throughput 3D super-resolution microscopy reveals *Caulobacter crescentus* in vivo Z-ring organization. *Proc. Natl. Acad. Sci. USA* **111**, 4566–4571 (2014).
27. Baddeley, D., Cannell, M.B. & Soeller, C. Visualization of localization microscopy data. *Microsc. Microanal.* **16**, 64–72 (2010).
28. Babcock, H., Sigal, Y. & Zhuang, X. A high-density 3D localization algorithm for stochastic optical reconstruction microscopy. *Opt. Nanoscopy* **1**, 1–10 (2012).
29. Li, Y., Ishitsuka, Y., Hedde, P.N. & Nienhaus, G.U. Fast and efficient molecule detection in localization-based super-resolution microscopy by parallel adaptive histogram equalization. *ACS Nano* **7**, 5207–5214 (2013).
30. Zhu, L., Zhang, W., Elnatan, D. & Huang, B. Faster STORM using compressed sensing. *Nat. Methods* **9**, 721–723 (2012).
31. Holden, S.J., Uphoff, S. & Kapanidis, A.N. D.A.O.S.T.O.R.M.: an algorithm for high-density super-resolution microscopy. *Nat. Methods* **8**, 279–280 (2011).
32. Hoogendoorn, E. *et al.* in *Focus on Microscopy (FOM2013)* (Maastricht, the Netherlands, 2013).
33. Min, J. *et al.* FALCON: fast and unbiased reconstruction of high-density super-resolution microscopy data. *Sci. Rep.* **4**, 4577 (2014).
34. Kim, K.S. *et al.* in *Proceedings of the 10th International Conference on Sampling Theory and Applications (SAMPTA)* (Bremen, Germany, 2013).
35. Grüll, F., Kirchgessner, M., Kaufmann, R., Hausmann, M. & Kebschull, U. in *2011 International Conference on Field Programmable Logic and Applications (FPL)* 1–5 (2011).
36. Anthony, S.M. & Granick, S. Image analysis with rapid and accurate two-dimensional Gaussian fitting. *Langmuir* **25**, 8152–8160 (2009).
37. Smith, C.S., Joseph, N., Rieger, B. & Lidke, K.A. Fast, single-molecule localization that achieves theoretically minimum uncertainty. *Nat. Methods* **7**, 373–375 (2010).
38. Brede, N. & Lakadamyali, M. GraspJ: an open source, real-time analysis package for super-resolution imaging. *Opt. Nanoscopy* **1**, 11 (2012).
39. Babcock, H.P., Moffitt, J.R., Cao, Y. & Zhuang, X. Fast compressed sensing analysis for super-resolution imaging using L1-homotopy. *Opt. Express* **21**, 28583–28596 (2013).
40. Starr, R., Stahlheber, S. & Small, A. Fast maximum likelihood algorithm for localization of fluorescent molecules. *Opt. Lett.* **37**, 413–415 (2012).
41. Quan, T. *et al.* Ultra-fast, high-precision image analysis for localization-based super resolution microscopy. *Opt. Express* **18**, 11867–11876 (2010).
42. Ma, H., Long, F., Zeng, S. & Huang, Z.-L. Fast and precise algorithm based on maximum radial symmetry for single molecule localization. *Opt. Lett.* **37**, 2481–2483 (2012).
43. Niu, L. & Yu, J. Investigating intracellular dynamics of FtsZ cytoskeleton with photoactivation single-molecule tracking. *Biophys. J.* **95**, 2009–2016 (2008).
44. Shtengel, G. *et al.* Interferometric fluorescent super-resolution microscopy resolves 3D cellular ultrastructure. *Proc. Natl. Acad. Sci. USA* **106**, 3125–3130 (2009).
45. Henriques, R. *et al.* QuickPALM: 3D real-time photoactivation nanoscopy image processing in ImageJ. *Nat. Methods* **7**, 339–340 (2010).
46. Parthasarathy, R. Rapid, accurate particle tracking by calculation of radial symmetry centers. *Nat. Methods* **9**, 724–726 (2012).
47. Boulanger, J. *et al.* Patch-based non-local functional for denoising fluorescence microscopy image sequences. *IEEE Trans. Med. Imaging* **29**, 29 (2010).
48. Andersson, S.B. Localization of a fluorescent source without numerical fitting. *Opt. Express* **16**, 18714–18724 (2008).
49. Kechkar, A., Nair, D., Heilemann, M., Choquet, D. & Sibarita, J.-B. Real-time analysis and visualization for single-molecule based super-resolution microscopy. *PLoS ONE* **8**, e62918 (2013).
50. Watanabe, S., Bennett, K., Takahashi, T. & Takeshima, T. in *Focus on Microscopy (FOM2013)* (Maastricht, the Netherlands, 2013).

ONLINE METHODS

Data. Establishing reference data is a key point for conducting a fair evaluation of image analysis algorithms; all software packages should use the same benchmark data sets. In microscopy, biologists and practitioners prefer real experimental image sequences, while algorithm developers need simulated data sets with ground-truth information. Here, we provide both: real experimental data sets which are mostly useful for visual inspection and synthetic simulated data sets, which are intensively used for the quantitative evaluation.

Experimental data. We acquired two sequences of images of tubulins, one in low-density imaging conditions (RealLS, a long sequence of 15,000 frames of 64 × 64 pixels) and one in high-density imaging conditions (RealHD, a short sequence of 500 high-density frames of 64 × 64 pixels). The sample is Cos-7 cells fixed in methanol. The microtubules are stained with α-tubulin primary antibodies and Alexa-647–conjugated secondary antibody fragments.

Simulated data. To achieve realistic images, we defined mathematical models for biological structure. We chose microtubules because they are often used to showcase SMLM studies. They are components of most eukaryotic cells which have widths smaller than the diffraction limit of the conventional light microscope. Microtubules are defined with their central axis elongating in a 3D space having an average outer diameter of 25 nm with an inner, hollow tube of 15 nm diameter.

To obtain rendered images at all scales including very high resolution (up to 1 nm/pixel), we represent the continuous-domain 3D curve by means of a polynomial spline. The sample is imaged in a limited field of view, i.e., less than 20 × 20 μm^2, and the centerlines of the microtubules have limited variation along the z (vertical) axis, i.e., less than 1 μm. The fluorescent markers are uniformly distributed over the structure according to the required density. We randomly assign each fluorophore with a random photon emission rate and with an active time instant according to a statistical lifetime model. For the synthetic data sets, we developed a simulator that can generate realistic image sequences of thousands frames resulting from the stochastic activation of millions of fluorophores.

The sample structure, fluorophore excitation, and image formation can be completely controlled with a large number of user configurable parameters (**Supplementary Note 3**). We defined the underlying sample structure in a continuous space which allows rendering of digital images at any scale. The exact locations of all fluorophores are therefore stored at high precision, as floating point numbers expressed in nanometers. This ground-truth file is useful for conducting objective evaluations without human bias.

Photon emission model. We calculate the photon flux in the following manner[51]:

$$F = \Phi \frac{P}{e} \sigma$$

where Φ is the quantum yield of the dye, P is the excitation laser power in W/cm^2, e is the energy of a single photon, $\sigma = 1{,}000\ \varepsilon \ln(10)/\mathrm{NA}$ is the absorption cross section, ε is the absorptivity coefficient of the dye and NA is the numerical aperture of the lens. The spatial variation of the excitation laser power, P, is modulated by a unimodal function that produces higher excitation at the center, rather than the border, of the field of view. The flux F is given in photons/s, and it is used for randomly determining a flux value for every excited fluorophore under a probability density function of a Poisson random variable.

Once the flux has been determined for the excited fluorophore, we choose an active duration A and a life-time model for computing the number of photons that would be emitted during the frame acquisition time T. We proposed three different life-time models: constant, linearly decreasing and exponentially decaying. This last is motivated by photobleaching phenomenon. An additional temporal parameter is a time delay Δ between the beginning of the frame and the time the fluorophore starts emitting photons. In this study, we chose A, Δ and T to be random variables but we impose that most of the photons of a single molecule are emitted within a single frame.

Source of photons. We consider three independent sources of photons: the signal of interest (the activated molecule) modeled as described above, the background signal normally distributed, slowly changes with time, and the autofluorescent signal simulated by introducing deep clusters of intense fluorophores that are constantly in an active mode, slowly changes with time. We sum these three sources of photons to yield the image that impinges on the detector. We then generate the microscopic image of each fluorophore by simulating the image formation process (**Fig. 1b**). Noise sources and perturbations include: non-homogeneous excitation laser power; random nature of the emission process of the fluorophore; shot noise for small photons count; EMCCD and read-out noise models.

A volumetric density parameter ρ controls the number of fluorophores per μm^3 that we generate. This parameter, together with the number of frames parameter, controls the total imaging density conditions: low-density, long sequences (LS) and high-density sequences (HD).

High-density. A recent trend in the super-resolution localization microscopy is the development of methods to detect multiple overlapping emitters in high density data sets[31,41,52–55]. To benchmark these methods, our generated HD are particularly suitable. Taking advantage of the exact knowledge of the sample structure, we defined three properties to qualify the density of the sequence: the average distance d_{NN} of a fluorophore to its nearest fluorophore within the same frame, the average number of fluorophores N per frame, and the average number of fluorophores M per μm^2. We obtained the following properties: d_{NN} = 1,536.1 nm, N = 26.3 nm, M = 0.14 for LS data sets and d_{NN} = 156.9 nm, N = 562.7 nm, M = 2.19 for HD data sets. The high number of fluorophores per frame and the small d_{NN} setting induce numerous overlapping of PSFs in the HD data sets.

Image formation. We put our effort in implementing a faithful reproduction of the physical reality to produce plausible synthetic images. The individual fluorophores, which are taken to be ideal point source, are convolved with the PSF of the microscope. Here, two models of PSF were considered, the defocussed Gaussian PSF model and the Gibson and Lanni PSF model.

The xy-Gaussian and z-exponential PSF model is defined as a 2D Gaussian function in the xy plane centered at (x_c, y_c) in

$$f(x,y,z) = A(z) e^{-\frac{(x-x_c)^2+(y-y_c)^2}{2\sigma^2(z)}}$$

where the variance depends on the axial position of the particle,

$$\sigma(z) = \frac{\lambda}{2\,\mathrm{NA}} e^{\log(2)\left|\frac{z - z_{\mathrm{focus}}}{z_{\mathrm{defocus}} - z_{\mathrm{focus}}}\right|}$$

The amplitude $A(z)$ is chosen in such a way as to make $f(x,y,z)$ have a unit norm at every value of z. The value z_{focus} is the axial position of the focal plane and z_{focus} allows one to introduce defocussing effects. The xy-Gaussian and z-exponential is a common approximation of the main lobe of the Airy pattern.

The scalar Gibson-Lanni model generates a more accurate PSF by using a 3-layer model taking into account refractive index mismatch at each optical interface. We rely on our accurate and fast implementation of the Gibson and Lanni PSF model to evaluate millions of convolution operations in very high resolution (5 nm/pixel) in a reasonable amount of time.

Conversion of photons to a digital number. At this stage, the number of electrons is converted to a digital number DN[56] by a simulated A/D converter that is characterized by a linear gain and an offset. Readout noise (Gaussian distributed) and dark noise (Poisson distributed) are added, as well. Final pixel values are calculated by checking for saturation and by quantization into 14-bits. Our simulated EMCCD camera down-samples the high resolution image from 5 nm/pixel to 100 or 150 nm/pixel by means of averaging. We then simulated the electron multiplier component by multiplying every pixel value by the EM noise factor $2^{½}$.

Data delivery. The frames are stored in a standard uncompressed 16-bits TIFF format. While typical experimental data consists of 10,000 frames, we restricted the number of frames so as to have data sets of moderate size, say 300 MB, after lossless compression. Such file sizes are still convenient to download from the Internet. Our data sets are accompanied by metadata information that includes microscope and camera parameters that are usually available in real experiment.

Level of difficulty. To produce data sets with different degrees of difficulty, we modified the contribution of autofluorescence, the level of acquisition noise, and the thickness of the sample. At the end of each simulation, we calculated two measures for every fluorophores: d_{NN}, the distance of the nearest neighborhood in the same frame; SNR_f, the local signal-to-noise ratio of a fluorophore. The SNR_f is the ratio of the difference of the peak signal and the mean of the local surrounding background and the standard deviation of the local background. For the HD data, d_{NN} is smaller than the size of the PSF, so that frames contain many overlapping PSFs (**Supplementary Note 3**).

Future directions. In the future, our reference data will include more features, such as drift, 3D localization (PSF engineering[57,58] and multiple planes[59,60], additional levels of molecule density, multiple fluorescent channels, asymmetrical PSF due to dipole effect[61], scattering effects, and a richer variety of noise models associated with various types of cameras, EMCCD, sCMOS[62,63]. It will also be interesting to generate benchmarking data to test the impact of clustering (spatial aggregation) and diffusion for single-particle tracking.

Evaluation and scoring calculation. *Theoretical accuracy.* The simplest theoretical localization precision is given by $s/\sqrt{N}$, where s is the size of the PSF and N is the number of detected photons[23,25]. This Cramér-Rao lower bound (CRLB) was initially introduced as a fundamental limit of accuracy[23,24,37]. There also exist refined CRLBs that take pixelation, various sources of noise, and fluorescence background into account. A survey of localization accuracy and precision in the SMLM context can be found in Deschout *et al.*[64], while uncertainties in the lateral localization in super-resolution microscopy were also addressed in Rieger *et al.*[20].

Some of our data fail to be compatible with the restrictive assumptions needed to establish CRLBs. It is only over LS1 and LS2 that it is valid to compare the experimental accuracy of the tested software packages to the theoretical expectations. Selecting the five best algorithms, we found that the accuracies are 21.05 nm and 32.13 nm for LS1 and LS2, respectively. They are worse than predicted by Thompson's rule (13.98 nm, 15.96 nm), but it is known that Thompson's rule[24] is too optimistic. More-realistic results are obtained with a bound recently proposed that gives (19.10 nm, 25.78 nm)[20].

Matching of two sets of localization. To establish statistical measures of detection rate and the localization accuracy, a pairing must first be found between the molecules localized by the participants and the molecules from the ground-truth. For each frame f, the pairing is obtained by solving a bi-partite graph-matching problem of minimizing the sum of distances between the two elements of a pair. The matching is enabled when the distance from $P_{\mathrm{ref}}(f)$ to its closest point $P_{\mathrm{test}}(f)$ is less the full-width half-maximum (FWHM) of the PSF. We deployed two matching algorithms: the presorted nearest-neighbor search and the Hungarian algorithm (**Supplementary Software**). Both gave similar results.

Computation of detection rate using the Jaccard index (JAC). The localized molecules successfully paired with some ground-truth molecule are categorized as true positives, TP; the remaining localized molecules are farther than ρ, unpaired, and categorized as false positives, FP; finally, ground-truth molecules that are not associated with any localized molecule are categorized as false negatives, FN.

The detection rate quantifies the framewise fidelity and completeness of the set $P_{\mathrm{test}}(f)$ of localizations with respect to the ground-truth $P_{\mathrm{ref}}(f)$. It involves the positive predictive value (precision p), the sensitivity (recall r) and the Jaccard index (JAC).

$$p = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP}}, \quad r = \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{TP}}, \quad \mathrm{JAC} = \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{FP} + \mathrm{TP}}$$

In this study, we observed that the precision value p is high (average 0.956, s.d. 0.09) in comparison to the recall value r (average 0.487, s.d. 0.25). We thus believe the most relevant measure of similarity between lists of localized molecules is the Jaccard index, j (i.e., JAC, in %).

Computation of localization accuracy using the root-mean-square error (RMSE). Let $(x_n^{\mathrm{Test}}, y_n^{\mathrm{Test}})$ and $(x_n^{\mathrm{Ref}}, y_n^{\mathrm{Ref}})$ be the nth matched pair, and where the superscripts Ref and Test indicate the oracle (ground-truth) and participant's list of localizations, respectively. The root-mean-square error (RMSE) of the matched localizations is

$$a^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n^{\mathrm{Test}} - x_n^{\mathrm{Ref}})^2 + (y_n^{\mathrm{Test}} - y_n^{\mathrm{Ref}})^2$$

doi:10.1038/nmeth.3442

The expectation of a^2 is the sum of the variance (precision) and the square of the bias (accuracy).

Independent computations convinced us that the bias is always negligible (unbiased estimators). Hence, the RMSE represents essentially the standard deviation of the errors, which is truly the precision of the estimator. Confusingly enough, in the specific lingo of the SMLM community, this term is called "accuracy" instead; we shall follow this improper terminology and call RMSE a localization accuracy.

Computation of the image quality using signal-to-noise ratio (SNR). To render an image, we let the contribution of each localized molecule take the form of an additive 2D Gaussian circular function with a standard deviation ten times smaller than the standard deviation of the PSF. Correspondingly, we let the resolution of the rendered image be ten times finer than that of the simulated camera with which we collected our synthetic data.

To compare the super-resolved image I_{test} to the oracle image I_{ref}, we compute

$$\text{SNR} = 10 \log_{10} \frac{|| I_{\text{ref}} ||^2}{|| I_{\text{ref}} - I_{\text{test}} ||^2}$$

Computation of the image resolution using the Fourier ring correlation (FRC). Fourier-ring correlation (FRC) was recently introduced as a method for measuring the image resolution in the SMLM context[21,65]. In the formalism of FRC, the set of positions is partitioned in two halves to determine the resolution. In our case, we populate the first half with $P_{\text{ref}}(f)$ and the second half with $P_{\text{test}}(f)$. The resolution is determined by applying the threshold $T = 0.5$ on the spectral correlation curve which typically decays monotonically[66].

Parsing the localization files. Because every software has its own file format, unit, axis and coordinate convention, we asked the participants to report their results in a delimiter-separated values text file (typically CSV), where every localized position in a frame is stored as a single row in this file. For every software, we created a description file (XML) containing the information to univocally parse the localization. The description file contains the type of separator (comma, tab...) the unit of position (nm, pixel), the first row containing fluorophore, the x and y shift of the system coordinate (center of the pixel, top-left corner), the unit of shift, the shift in the numbering of the frame (0, 1) and finally the column numbers in which to find the information: x, y, intensity and frame. With this simple procedure, we succeeded in reading all localization files without modifying any software.

Minimal bound on performance. We developed a rudimentary ImageJ plugin in Java, called CenterOfGravity, to determine a lower performance bound. This software detects candidate molecules by thresholding the local maximum of the band-pass filter; the accurate position is simply the center of gravity of a local neighborhood window centered on the candidates. Two software packages consistently failed to meet this criteria and were not further evaluated.

Usability. For the practitioner, the usability of the software is an important aspect for the daily work. The usability (USA) evaluation cannot be carried out quantitatively because it involves human behavior and multiple interaction factors between computer, data and users. Here, we follow the strategy of Carpenter *et al.*[67] and evaluate each software using two sources of information: a questionnaire that was filled out by the software developers and some testing of the software. The maximum usability grade is 5 and such a software fulfills the following requirements.

- Accessibility: easy to find and to download from the web. Non-accessible softwares are assigned a usability grade of 0.
- Open-source: free tool or an add-on of a free software (for example, ImageJ), accessible source code, no cost.
- Installation: no dependency of specific hardware, no requirements of additional library, easy to install, binaries for multi-platforms, multiple operation systems, double-click type installation, fast learning curve.
- Usage: user-friendly interface, intuitive parameters, documentation, interoperability.
- Maintenance: continued, long-term support, feedback mechanism.

Computational time. At first sight, execution runtime would appear to be measurable objectively. Unfortunately, the participating packages all exhibit some degree of dependence upon specific hardware installations and code-development environments. This prevented us from running every software by ourselves. As a proxy, we determined the computational runtime (TIME) by analyzing the answers we received from the participants and by normalizing it by the power of their machine.

We asked the algorithm developers to report not only their own run-time values but also the main specifications of the run-time machine. There is a large variety of processors among the participants of this study. We therefore weighted the runtime by "normalized" coefficient: 0.75 for relatively slow desktop machine (e.g. 2.70 Ghz Intel Core i5), 1.25 for fast desktop machines (e.g. 3.40 Ghz Intel 4 cores).

For computers that are equipped with additional hardware, like graphical processing unit (GPU) usage, or field-programmable gate array (FPGA), we assign a penalty factor of 3.00. Another aspect we take into account is non comparable tasks between the various packages. Some software measures only the elapsed time of the localization task. Others measure the full processing task: loading frames in memory, localization and rendering. We compensate for that by introducing an advantage factor of 0.75 and we apply it to software that measured the full processing task.

The runtime of every software was normalized by the above coefficients to yield a "normalized runtime" measure, T_n, which is mapped to a grade scale and clipped to between 0 and 5. We note that our grading should be regarded as a rough indicator of the software efficiency only.

Grading and ranking. For all criteria, JAC, ACC, SNR, FRC, USA and TIME, we attributed a normalized grade between 0 (worst case) and 5 (best case); see **Supplementary Data 2**. The values of criteria were normalized to impose an average of 2.5 and a s.d. of 1.5, and they are clipped to 0 and 5.

We computed an overall performance score as a weighted sum of the criteria (average over the 3 data sets if necessary). The final rank is associated with either low-density imaging data sets or

with high-density imaging data sets. We normalize the criteria values to be in the interval [0,5] and define the final score as

$$s = \frac{\lambda_{JAC} \cdot \overline{JAC} + \lambda_{ACC} \cdot \overline{ACC} + \lambda_{SNR} \cdot \overline{SNR} + \lambda_{FRC} \cdot \overline{FRC} + \lambda_{USA} \cdot USA + \lambda_{TIME} \cdot TIME}{\lambda_{JAC} + \lambda_{ACC} + \lambda_{SNR} + \lambda_{FRC} + \lambda_{USA} + \lambda_{TIME}}$$

We ran a principal-components analysis on the 4 criteria JAC, ACC, SNR and FRC showing that all criteria have an similar importance. We chose to compute an overall performance measure, as weighted sum, namely a score *s*, by choosing the following weights (see **Supplementary Fig. 3**).

Challenge organization. The organization of a world-wide challenge was an opportunity to get the attention of a large number of the developers working in different fields, including biology, biophysics and computer science. We broadly advertized the challenge trying to ensure coverage of most representative and well-known software and also to attract newcomers to the field. The Localization Microscopy challenge (http://bigwww.epfl.ch/smlm) was presented at the IEEE ISBI conference, at San Francisco, in April 2013. The high participation rate of the developers reveals the importance of this study. The localization task was carried out by the software developers themselves with the exception of PeakSelector, Octane and CSSTORM, which were performed by experts. By having the authors or experts use their own algorithms, we believe we obtained the best performance possible. We initially provided them with training data sets that included ground-truth information, allowing them to choose the appropriate mode and to properly tune the parameters of their algorithms. We assumed that developers were at the same time most knowledgeable about their software, and keenest on cranking out the best performance, guided by the training data. We also offered the opportunity to submit three different runs for each data sets with different settings. Only four participants (a-livePALM, Auto-Bayes, SimpleSTORM and SOSplugin) have chosen this option. Finally, we observed that the results were very similar from one run to the next. This is summarized in **Table 2** and reported in **Supplementary Data 2**.

This comparative study was first released at the IEEE ISBI 2013 Symposium (http://bigwww.epfl.ch/smlm/). Ever since, it has proved to be a valuable resource to developers and end users alike. The ISBI challenge has now turned into a permanent online challenge and is referred to in the Grand Challenge in Medical Image Analysis website (http://www.grand-challenge.org/).

51. Hinterdorfer, P. & Oijen, A.V. *The Handbook of Single-Molecule Biophysics* (Springer, 2009).
52. Huang, F., Schwartz, S.L., Byars, J.M. & Lidke, K.A. Simultaneous multiple-emitter fitting for single molecule super-resolution imaging. *Biomed. Opt. Express* **2**, 1377–1393 (2011).
53. Wang, Y., Quan, T., Zeng, S. & Huang, Z.-L. PALMER: a method capable of parallel localization of multiple emitters for high-density localization microscopy. *Opt. Express* **20**, 16039–16049 (2012).
54. Babcock, H., Sigal, Y.M. & Zhuang, X. A high-density 3D localization algorithm for stochastic optical reconstruction microscopy. *Opt. Nanoscopy* **1**, 6 (2012).
55. Cox, S. *et al.* Bayesian localization microscopy reveals nanoscale podosome dynamics. *Nat. Methods* **9**, 195–200 (2012).
56. Janesick, J.R. *Photon Transfer* (SPIE Publications, 2007).
57. Huang, B., Wang, W., Bates, M. & Zhuang, X. Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. *Science* **319**, 810–813 (2008).
58. Pavani, S.R.P. *et al.* Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. *Proc. Natl. Acad. Sci. USA* **106**, 2995–2999 (2009).
59. Juette, M.F. *et al.* Three-dimensional sub-100 nm resolution fluorescence microscopy of thick samples. *Nat. Methods* **5**, 527–529 (2008).
60. Ram, S., Prabhat, P., Ward, E.S. & Ober, R.J. Improved single particle localization accuracy with dual objective multifocal plane microscopy. *Opt. Express* **17**, 6881–6898 (2009).
61. Mortensen, K.I., Churchman, L.S., Spudich, J.A. & Flyvbjerg, H. Optimized localization analysis for single-molecule tracking and super-resolution microscopy. *Nat. Methods* **7**, 377–381 (2010).
62. Bennett, K., Takahashi, T., Sage, D. & Huang, Z. in *Fourth Single Molecule Localisation Microscopy Symposium (SMLMS'14)* (London, UK, 2014).
63. Huang, F. *et al.* Video-rate nanoscopy using sCMOS camera-specific single-molecule localization algorithms. *Nat. Methods* **10**, 653–658 (2013).
64. Deschout, H. *et al.* Precisely and accurately localizing single emitters in fluorescence microscopy. *Nat. Methods* **11**, 253–266 (2014).
65. Banterle, N., Bui, K.H., Lemke, E.A. & Beck, M. Fourier ring correlation as a resolution criterion for super-resolution microscopy. *J. Struct. Biol.* **183**, 363–367 (2013).
66. Rosenthal, P.B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
67. Carpenter, A.E., Kamentsky, L. & Eliceiri, K.W. A call for bioimaging software usability. *Nat. Methods* **9**, 666–670 (2012).

doi:10.1038/nmeth.3442