Supplementary Information for:
StormGraph: An automated graph-based algorithm for quantitative clustering analysis of single-molecule localization microscopy data.

Joshua M. Scurll, Libin Abraham, Da Wei Zheng, Reza Tafteh, Keng C. Chou,
Michael R. Gold, Daniel Coombs

Supplementary Notes

Supplementary Note 1: Resolution limits of community detection algorithms

The Louvain method is subject to a resolution limit (1), whereby the minimum size of detected clusters increases as the size of the graph increases. However, each implementation of the Louvain method generates a hierarchical clustering, with the resolution limit applying to the highest (most coarse) level. To minimize the effect of the resolution limit on the preliminary clustering performed by StormGraph during node degree thresholding when localization uncertainties are unknown, the lowest (finest) level of the hierarchy is used (matching the approach taken by Emmons *et al* in their comparison of community detection algorithms (2)). The two-level Infomap algorithm is also subject to a resolution limit, but it is orders of magnitude smaller than that of the Louvain method (3) and therefore not a major concern. For best results, the lowest level of a hierarchical clustering optimized by the multi-level Infomap method, which is not subject to a resolution limit (3), could be used instead, but for computational efficiency we use the faster two-level method during the preliminary clustering step.

Supplementary Tables

Table S1

Algorithm:	SR-Tesseler	Cluster-ViSu	DBSCAN	OPTICS	FOCAL	Bayesian method	Storm-Graph
References	(4)	(5; 6)	(7)	(8)	(9)	(10; 11)	N/A
Mathematical approach	Voronoi tessellations	Voronoi tessellations	Density-based	Density-based	Pixel-based	Bayesian, model-based	Graph-based
Independent of global localization density	yes	yes	no	no*	yes	yes	yes
3D capability	no	yes	yes	yes	no	yes	yes
Adapts to axial variation in 3D SMLM data	N/A	yes	no	yes	N/A	yes**	yes
Suitable for arbitrary cluster shapes and sizes	yes	yes	yes	yes	yes***	no	yes
Accounts for localization errors	no	no	no	no	#	yes	yes
Includes colocalization analysis	no	yes	no	no	no	no	yes
Hierarchical clustering	no	no	no	yes	no	no	yes
Automatically generates single-level clustering	yes	yes	yes	no	yes	yes	yes

Table S1. Comparison of features of StormGraph and other clustering algorithms.
Uses a single global estimate of localization uncertainty to define a pixel size, but does not account for each individual localization's specific positional uncertainty.
* Requires user-specification of the same two density-based parameters as DBSCAN, but theoretically the search radius ε can be made arbitrarily large, unlike for DBSCAN. This means that, theoretically, OPTICS can operate independent of the global localization density (i.e. it can be implemented with parameters that do not depend on the density). However, this would result in all localizations being grouped together in a single cluster at the most coarse level of the cluster hierarchy and would be computationally inefficient. Therefore, in practice, the parameter ε should be set.
** Technically possible by modifying the mathematical model used by the algorithm. However, this is non-trivial and likely beyond the capabilities of biologists using the algorithm.
*** Yes, except that clusters will be composed of square pixels.

References

[1] Fortunato, S. & Barthélemy, M. Resolution limit in community detection. *Proceedings of the National Academy of Sciences* **104**, 36–41 (2007).

[2] Emmons, S., Kobourov, S., Gallant, M. & Börner, K. Analysis of network clustering algorithms and cluster quality metrics at scale. *PLoS ONE* **11**, e0159161 (2016).

[3] Kawamoto, T. & Rosvall, M. Estimating the resolution limit of the map equation in community detection. *Physical Review E* **91**, 012809 (2015).

[4] Levet, F. *et al.* SR-Tesseler: a method to segment and quantify localization-based super-resolution microscopy data. *Nature Methods* **12**, 1065–1071 (2015).

[5] Andronov, L., Orlov, I., Lutz, Y., Vonesch, J. & Klaholz, B. ClusterViSu, a method for clustering of protein complexes by Voronoi tessellation in super-resolution microscopy. *Sci. Rep.* **6** (2016).

[6] Andronov, L. *et al.* 3DClusterViSu: 3D clustering analysis of super-resolution microscopy data by 3D Voronoi tessellations. *Bioinformatics* **34**, 3004–3012 (2018).

[7] Ester, M., Kriegel, H., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.*, 226–231. KDD-96 (AAAI Press, 1996).

[8] Ankerst, M., Breunig, M. M., Kriegel, H. & Sander, J. OPTICS: Ordering Points To Identify the Clustering Structure. 49–60 (ACM Press, 1999).

[9] Mazouchi, A. & Milstein, J. Fast Optimized Cluster Algorithm for Localizations (FOCAL): a spatial cluster analysis for super-resolved microscopy. *Bioinformatics* **32**, 747–754 (2016).

[10] Rubin-Delanchy, P. *et al.* Bayesian cluster identification in single-molecule localization microscopy data. *Nature Methods* **12**, 1072–1076 (2015).

[11] Griffié, J. *et al.* 3D Bayesian cluster analysis of super-resolution data reveals LAT recruitment to the T cell synapse. *Sci. Rep.* **7** (2017).
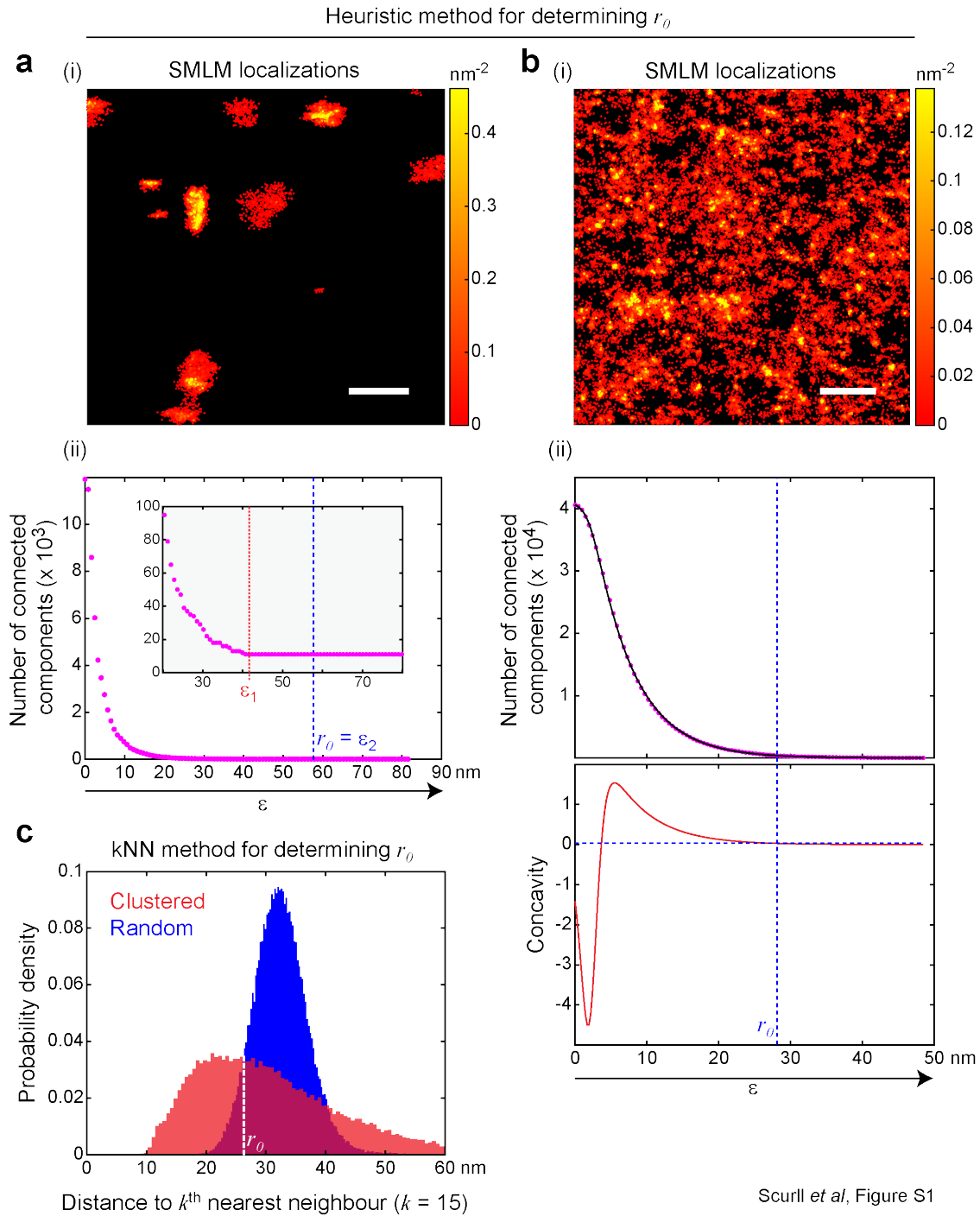
Supplementary Figures

Figure S1

Figure S1. The two available methods for setting the search radius r_0 in StormGraph. (a) The heuristic elbow method for setting r_0 when there is a natural number of clusters in the data. (i) Example

SMLM localizations with well defined clusters and minimal dispersed localizations between clusters. Color bar = density (nm^{-2}), scale bar = 500 nm. (ii) Number of connected components in the ε-neighborhood graph vs ε for the data in (i). StormGraph sets $r_0 = \varepsilon_2$, where an ε_2-neighborhood has twice the area/volume of an ε_1-neighborhood, if the number of connected components remains constant for $\varepsilon_1 \leq \varepsilon \leq \varepsilon_2$. Inset: enlarged region of plot. **(b)** (i) Example SMLM localizations with poorly defined clusters and many dispersed localizations between clusters. Color bar = density (nm^{-2}), scale bar = 500 nm. (ii) The heuristic elbow method for setting r_0 for the data in (i). A function $f(\varepsilon)$ is fit to number of connected components vs ε for the ε-neighborhood graph (top). The concavity $f''(\varepsilon)$ of this curve is then computed (bottom), and the value of r_0 (vertical dashed line) is set as the value of ε at which $f''(\varepsilon)$ returns to 2% (horizontal dashed line) of its maximum value. **(c)** The k-nearest neighbor (kNN) method for setting r_0 (illustrated for the data in panel (b)). A histogram of distances from nodes to their k^{th} nearest neighbors is obtained for both the clustered SMLM data (red) and a random distribution of points with the same average density (blue), where k is a user-specified parameter. For data with known localization uncertainties, the 95% confidence level from Monte Carlo simulations is used for the kth nearest neighbor distance for each localization. The value of r_0 (vertical white dashed line) is set as the distance at which the two histograms first intersect.
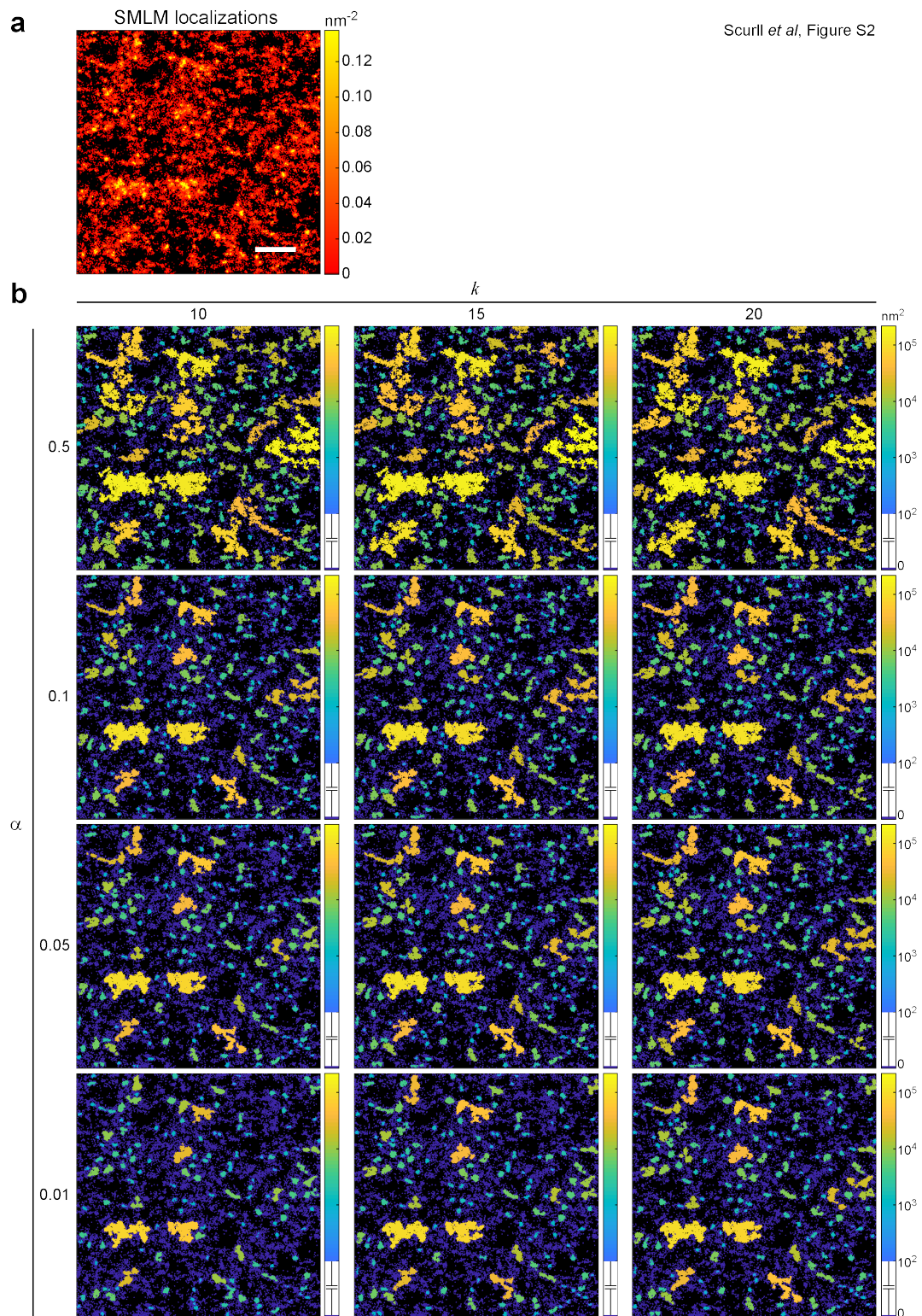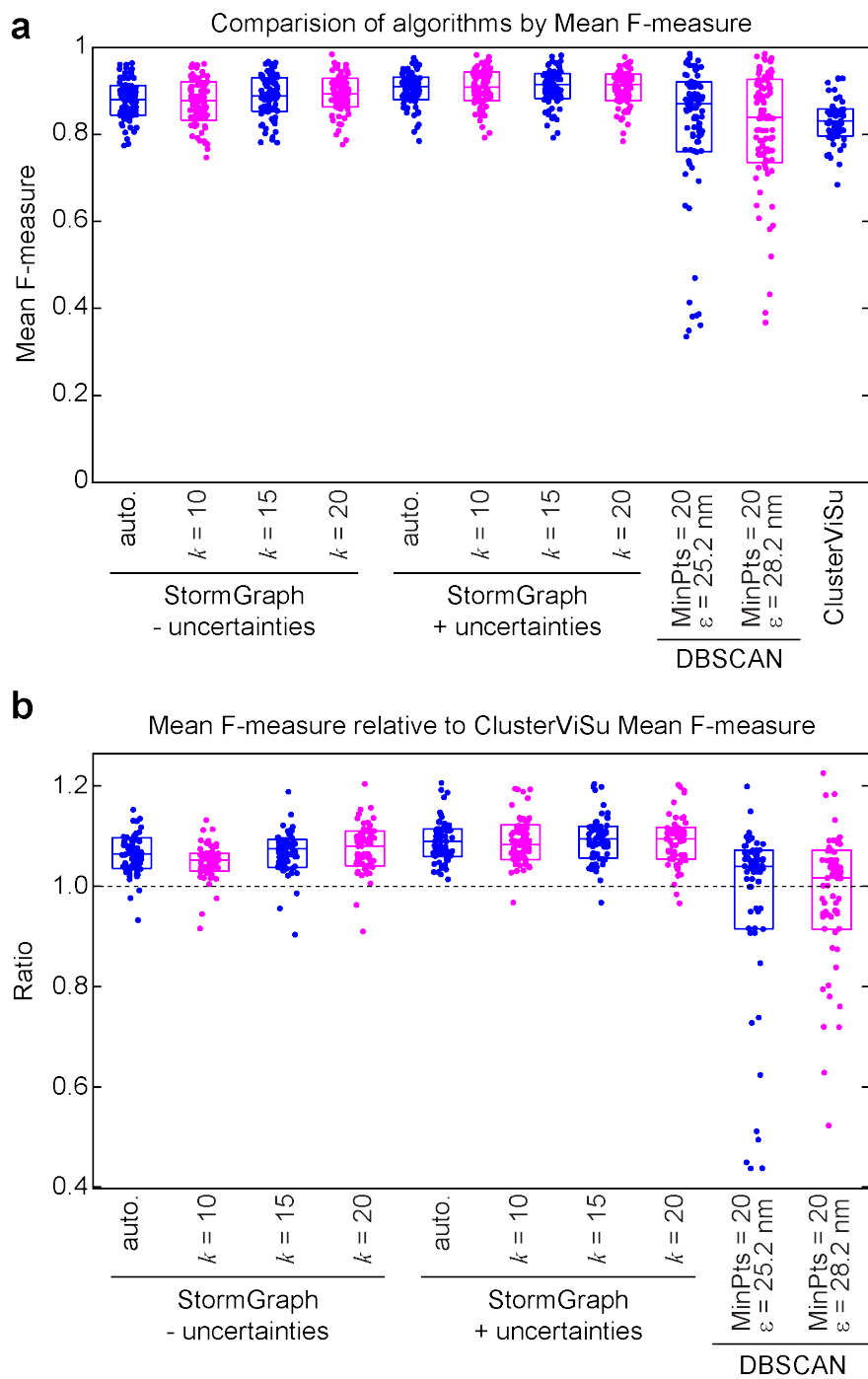
Figure S2

a

SMLM localizations

b

Figure S2. Effect of varying StormGraph's parameters k and α for an example SMLM ROI. **(a)** An ROI containing localizations of Alexa Fluor 647-labeled cell-surface B-cell antigen receptors of IgM isotype on a TMD8 diffuse large B-cell lymphoma cell imaged by dSTORM. Scale bar = 500 nm, color bar = density (nm^{-2}). **(b)** Clusters identified by StormGraph using different values of the parameters k and α. In all cases, StormGraph was implemented in 2D, using localization uncertainties, and with a minimum cluster size of 5 localizations. Clusters are colored according to their areas (nm^2) calculated by StormGraph. A cluster area of 0 (dark blue) indicates that a localization was classified as unclustered (i.e. not assigned to any cluster).
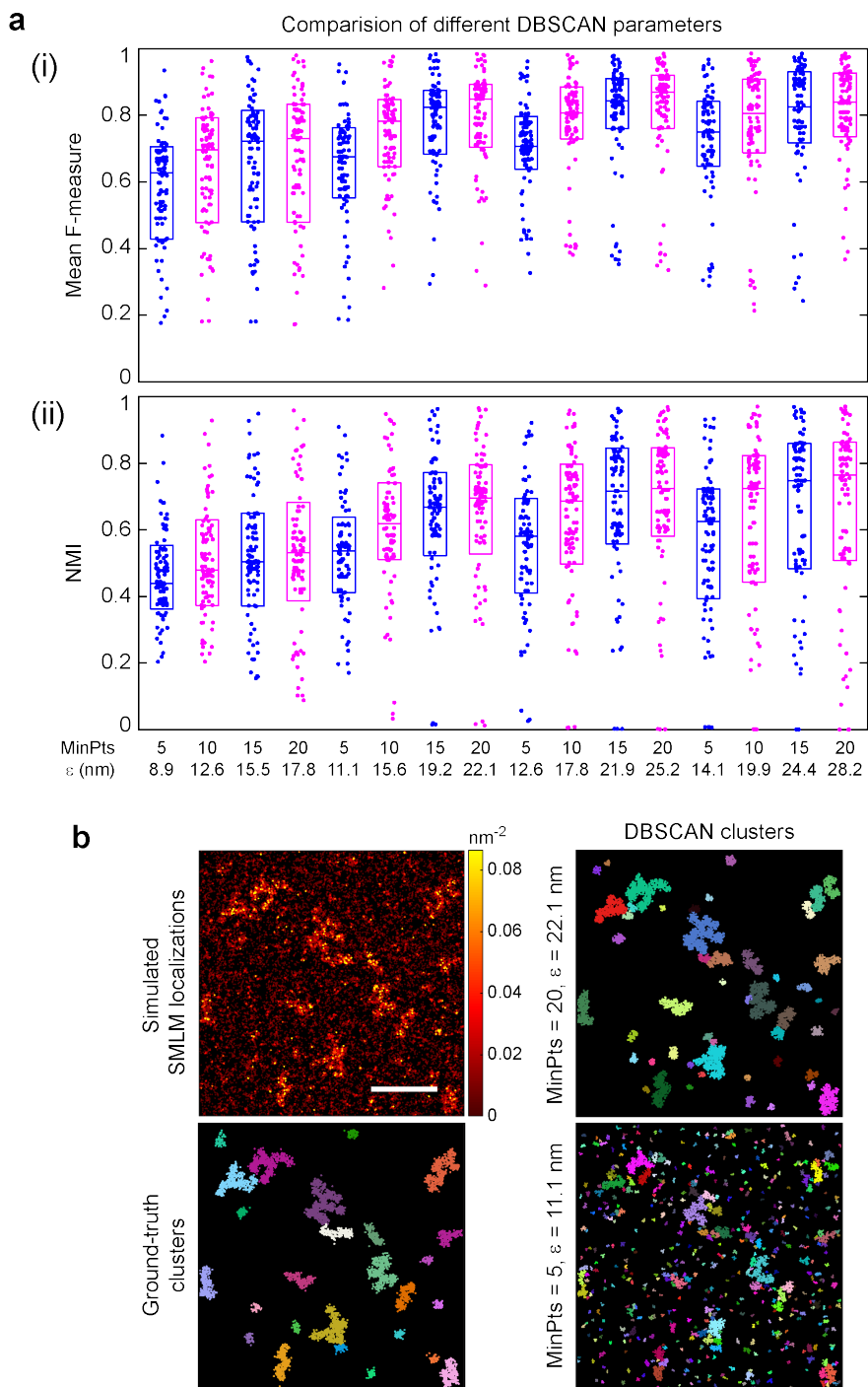
Figure S3

a Comparision of algorithms by Mean F-measure

b Mean F-measure relative to ClusterViSu Mean F-measure

Figure S3. For simulated data with known ground-truth clusters, StormGraph assigns localizations to clusters more accurately, as measured by mean F-measure, than DBSCAN and

ClusterViSu. This figure presents the exact same data as Figure 2b–c, using mean F-measure as an alternative to normalized mutual information to measure the accuracy of the cluster assignments of points. **(a)** Accuracy of assigning data points to clusters as assessed by mean F-measure (1 = perfect). A total of 64 simulated datasets were analyzed using StormGraph, ClusterViSu, and DBSCAN. StormGraph was run either with (+) or without (-) localization uncertainties. The value of r_0 used by StormGraph was determined using the heuristic method (auto.) or the k-nearest neighbor method with $k = 10$, 15 or 20. DBSCAN was implemented using 16 different selections of its two parameters, MinPts and ε, of which the two best-performing are shown here. ClusterViSu results are only shown for the 49 datasets on which the analysis was completed in under 2 h. Boxes show medians and interquartile ranges. **(b)** The same results as in panel (a) are displayed normalized to ClusterViSu's performance for each of the 49 simulated datasets for which analysis by ClusterViSu was completed in under 2 h. StormGraph was consistently more accurate than ClusterViSu at assigning points to clusters, indicated by ratios > 1.

a

Figure S4. DBSCAN is sensitive to the choice of parameters, and no single choice of parameters is suitable for batch processing cluster analysis when localization density is variable

between datasets. **(a)** Accuracy of cluster assignments by DBSCAN, using different values of the parameters MinPts and ε, of localizations in the 64 simulated datasets used for Figure 2b. Accuracy was measured by (i) mean F-measure and (ii) normalized mutual information (NMI). A total of 16 parameter pairs corresponding to four different threshold densities (0.020, 0.013, 0.010 and 0.008 nm^{-2}) were tested. Boxes show medians and interquartile ranges. **(b)** A visual demonstration that two different DBSCAN parameter choices corresponding to the same threshold density produce very different clustering results. Top left: the simulated dataset example used in Figure 2a; color bar = density (nm^{-2}); scale bar = 500 nm. Bottom left: ground-truth clusters present in the example simulated dataset; colors distinguish distinct clusters. Right panels: clusters identified by DBSCAN using two different parameter pairs, both of which correspond to a threshold density of 0.013 nm^{-2}.
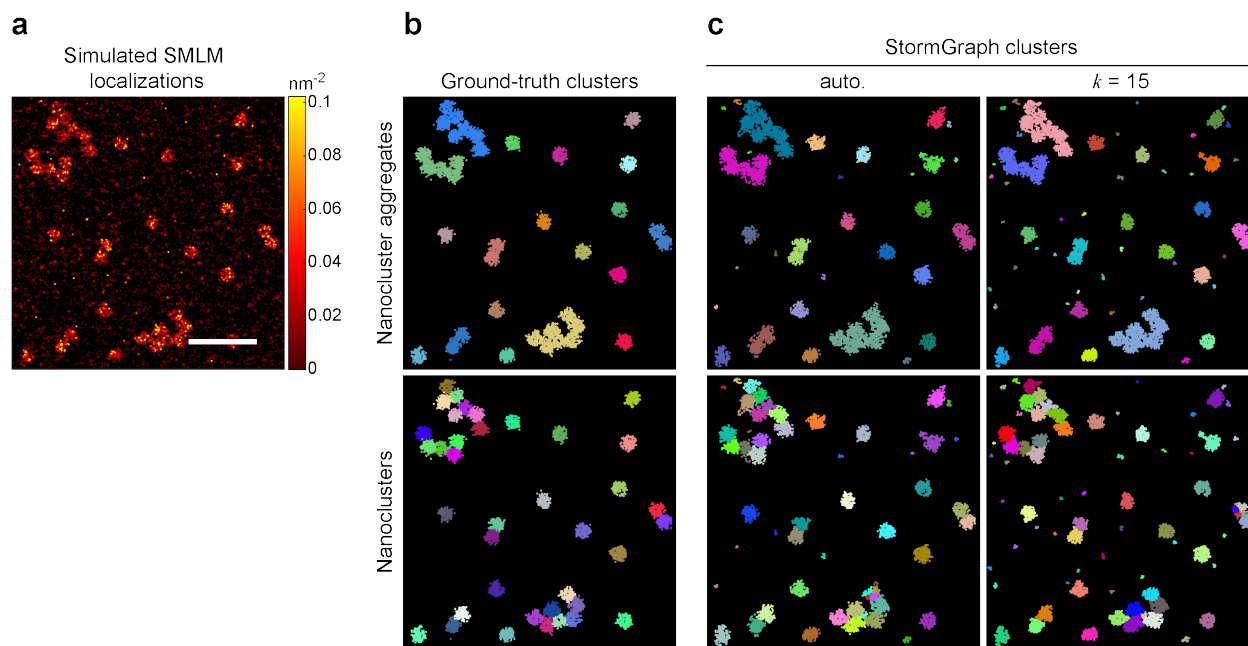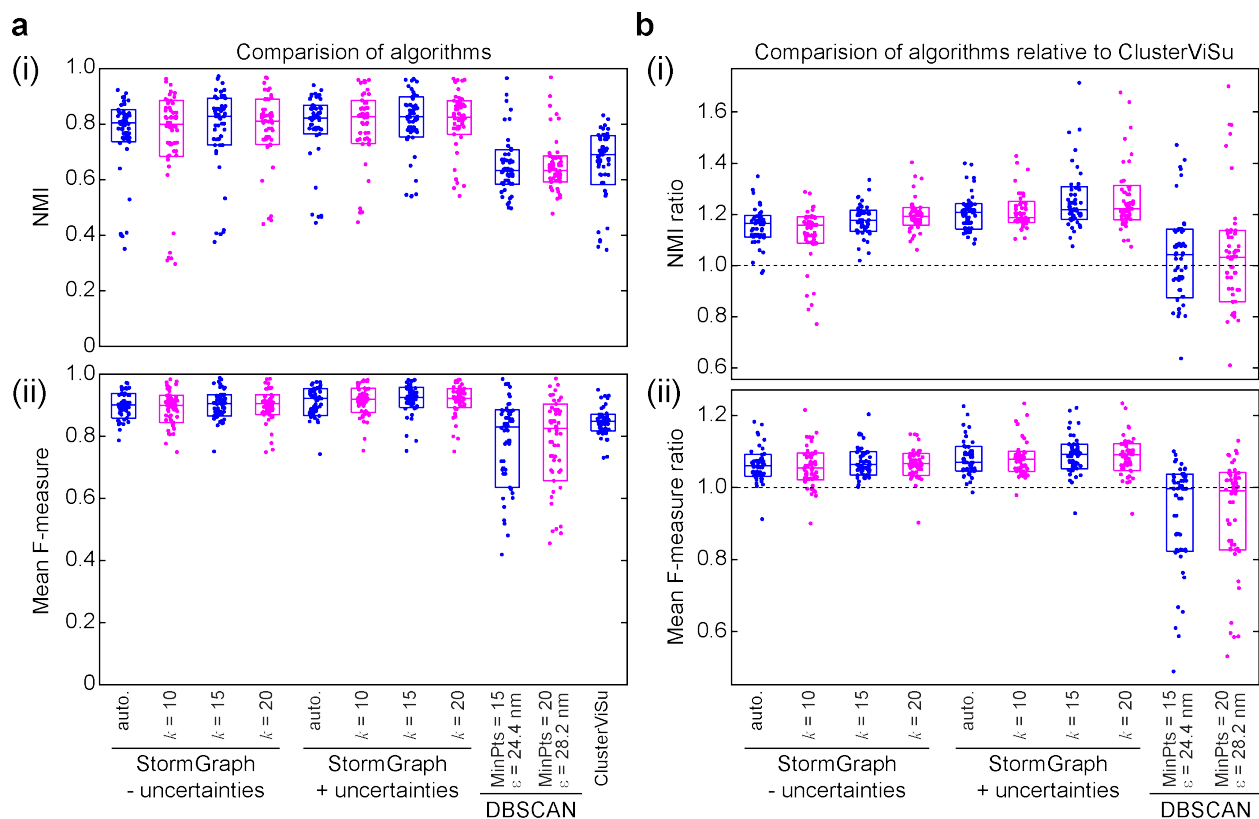
Figure S5

Figure S5. StormGraph's cluster hierarchy accurately recovers clusters at different scales in simulated data. (a) An example simulated SMLM dataset containing isolated and aggregated nanoclusters of radius 50 nm and randomly distributed molecules outside of clusters. This was one of the 64 simulated datasets used for Figure 2. Color bar = density (nm^{-2}), scale bar = 500 nm. **(b)** Ground-truth clusters at the two levels of clustering in the simulated data from (a). Colors distinguish different clusters. Top: Large-scale clustering. Some nanoclusters are aggregated into larger clusters. Bottom: Individual nanoclusters (50 nm radius) at the small scale. **(c)** Clusters from the cluster hierarchy found by StormGraph ($\alpha = 0.05$, minimum cluster size = 5 localizations) using either the heuristic (auto.) method or the kNN method with $k = 15$ to set r_0 and using localization uncertainties. Top: Colors show distinct clusters at the top (coarsest) level of the cluster hierarchy. Bottom: Colors show distinct clusters at a manually identified level of the cluster hierarchy.
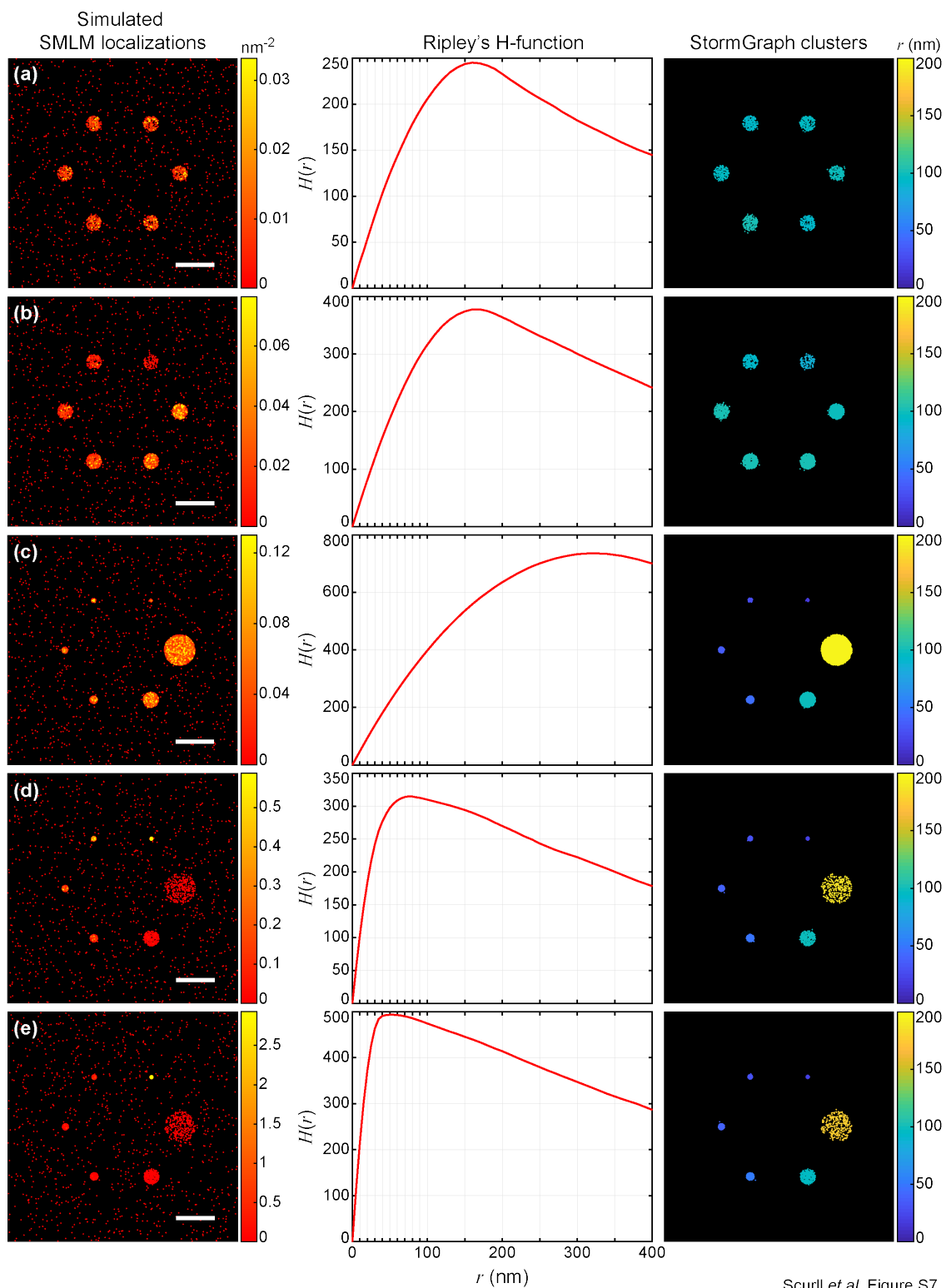
Figure S6

Figure S6. StormGraph is more accurate than ClusterViSu and DBSCAN at assigning localizations to clusters in simulated data with no overcounting of single molecules. We simulated 2D SMLM data as described for Figure 2 in Methods but with every molecule having exactly one localization (i.e. no overcounting). We then performed the same algorithm testing that we performed for Figure 2. **(a)** Accuracy of cluster assignments by StormGraph, DBSCAN, and ClusterViSu, measured by (i) Normalized Mutual Information (NMI) and (ii) mean F-measure, for 38 simulated datasets in which every molecule is localized exactly once. StormGraph was implemented either using (+) or not using (-) localization uncertainties during clustering and with r_0 set either using the heuristic method (auto.) or using $k = 10$, 15 or 20. For all runs of StormGraph, $\alpha = 0.05$ and MinCluSize = 5 were fixed. DBSCAN was tested using 24 different pairs of its two user-specified parameters MinPts and ε. Shown here are the two choices of DBSCAN parameters that yielded the best results. A minimum cluster size of 5 points was used for ClusterViSu. ClusterViSu results are not shown for one of the 38 datasets because it failed to complete analysis in under 2 h. Boxes show medians and interquartile ranges. **(b)** Same as (a), but with StormGraph's or DBSCAN's accuracy, measured by (i) NMI and (ii) mean F-measure, divided by ClusterViSu's NMI or mean F-measure for each of the 37 simulated datasets for which ClusterViSu analysis completed in under 2 h. Ratios > 1 indicate that StormGraph or DBSCAN was more accurate than ClusterViSu for the corresponding datasets.

13

Figure S7

Simulated SMLM localizations   Ripley's H-function   StormGraph clusters

Scurll *et al*, Figure S7

14

Figure S7. Ripley's H-function is unsuitable for quantifying SMLM data containing heterogeneous clusters. Left panels: Six simulated circular clusters of various radii and densities, and approximately 1000 randomly distributed points. Simulated data contained no uncertainties. Color bar = density (nm^{-2}), scale bar = 500 nm. Middle panels: Ripley's H-function vs length scale (r) for the simulated data. Deviation above 0 indicates clustering. The position (r value) of the peak of $H(r)$ is often used to estimate cluster radius. Right panels: Clusters found by StormGraph and colored by their radii estimated from the StormGraph-quantified cluster areas (radius $= \sqrt{\mathrm{area}/\pi}$). StormGraph was implemented using $k = 10$, $\alpha = 0.05$, and a minimum cluster size of 5 points. **(a)** All clusters contain approximately 200 points and have radius 100 nm. **(b)** Clusters all have radius 100 nm but contain different numbers of points (100, 200, 300, 400, 500, and 600). **(c)** Clusters all have the same density but have different radii (20 nm, 30 nm, 40 nm, 50 nm, 100 nm, and 200 nm). **(d)** Clusters all contain approximately 300 points but have different radii (20 nm, 30 nm, 40 nm, 50 nm, 100 nm, and 200 nm). **(e)** Smallest cluster (20 nm radius) contains approximately 1000 points. All other clusters (radii 30 nm, 40 nm, 50 nm, 100 nm, and 200 nm) contain approximately 300 points.
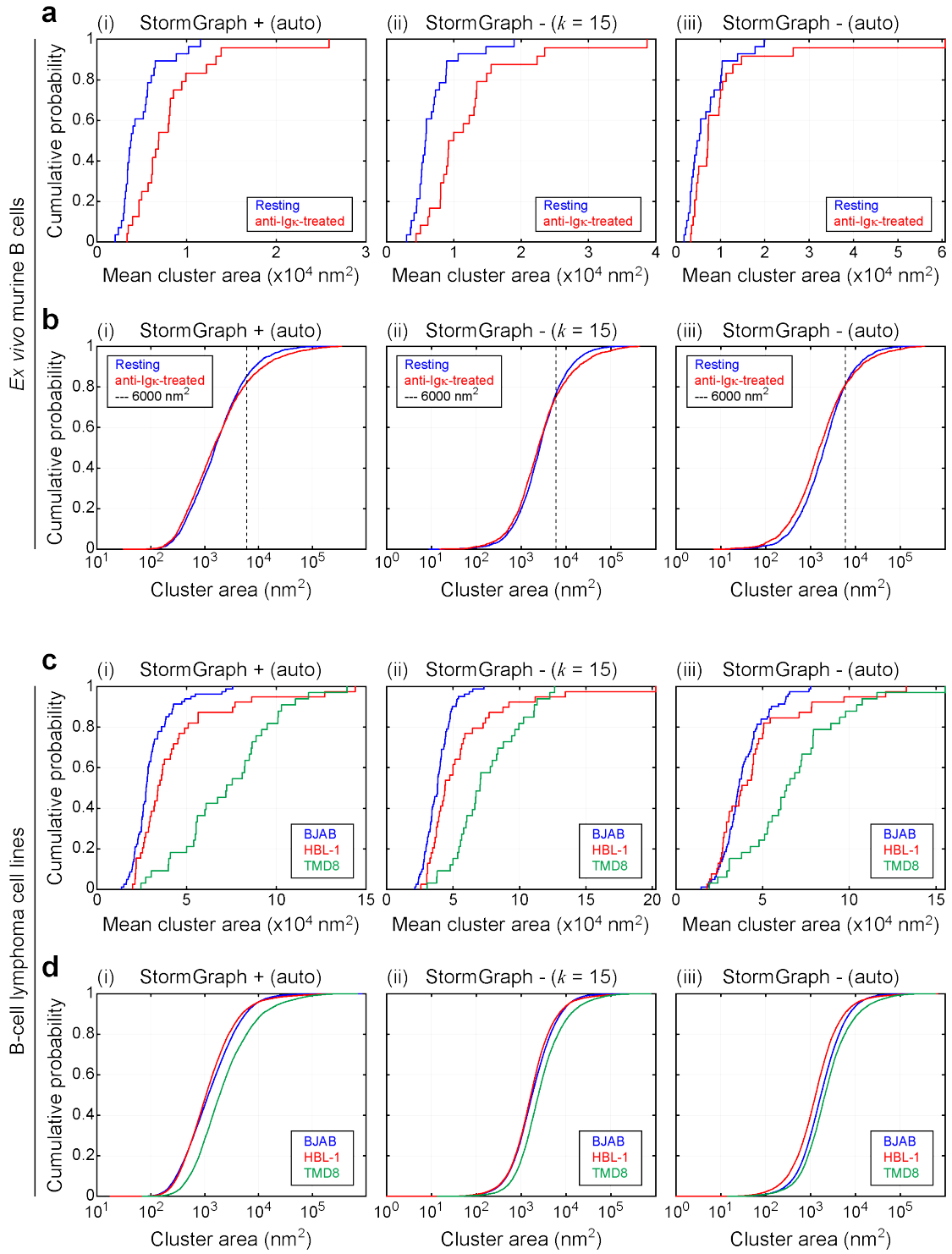
Figure S8

Figure S8. Repeated cluster analysis of dSTORM data of IgM-BCRs from Figure 3 using different StormGraph settings. This figure presents the same data as Figure 3 using different settings for StormGraph analysis. StormGraph was implemented using either the heuristic method (auto.) or the k-nearest neighbor method with $k = 15$ to determine the value of r_0. Positional uncertainties in dSTORM localizations were either used (+) or not used (-) during analysis by StormGraph. **(a–b)** Cumulative distribution function (CDF) for (a) the mean area and (b) all areas of IgM-BCR clusters detected by StormGraph, using different settings, in each dSTORM ROI from resting (blue) and anti-Igκ-treated (red) *ex vivo* murine splenic B cells. **(c–d)** CDF for (c) the mean area and (d) all areas of IgM-BCR clusters detected by Storm-Graph, using different settings, in each dSTORM ROI from BJAB (blue), HBL-1 (red) and TMD8 (green) cells.
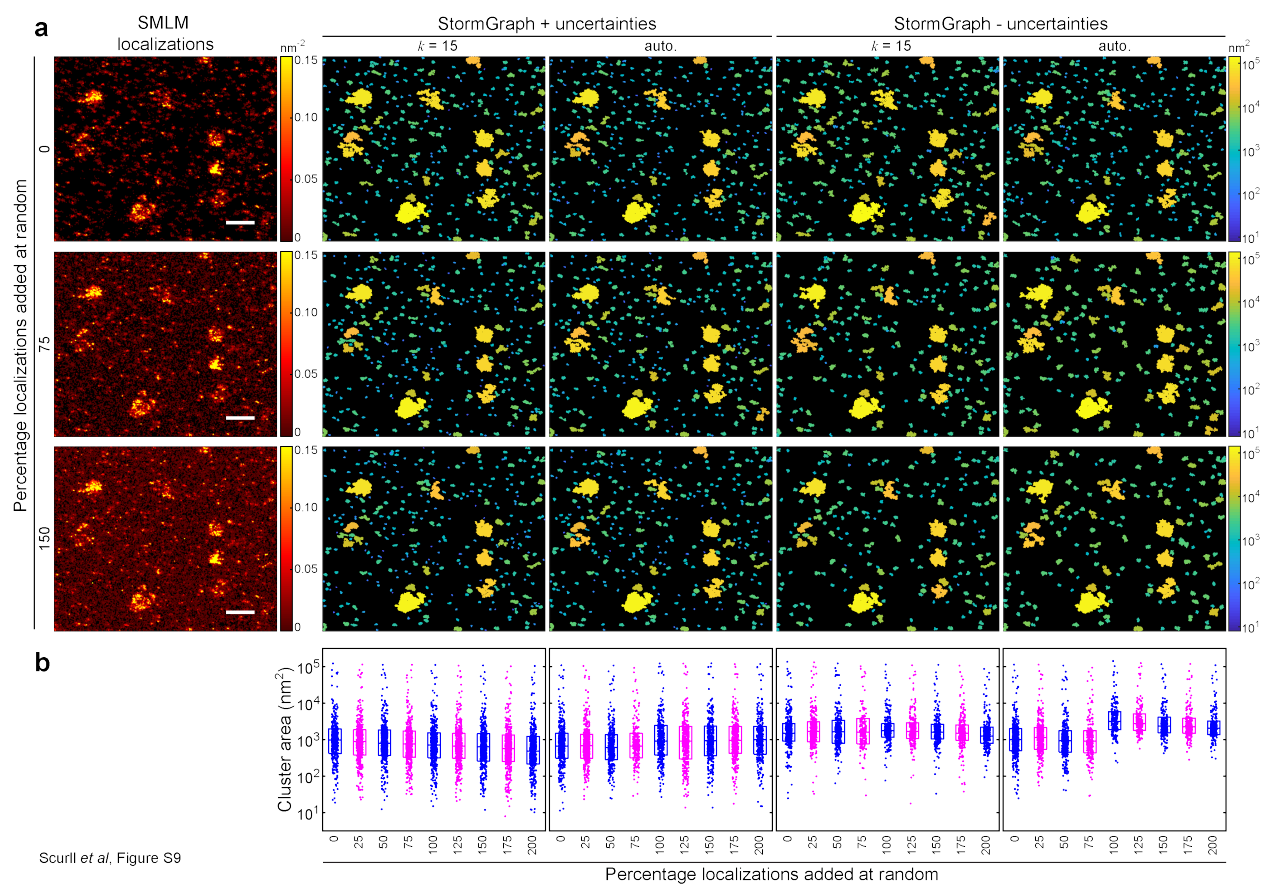
Figure S9

Figure S9. Effect on StormGraph of increasing random noise. (a) IgM-BCR dSTORM localizations in an ROI from an HBL-1 cell (top-left panel) with increasing numbers of additional randomly distributed points (left panels; color bars = density (nm^{-2})). The number of random points added are specified as a percentage of the original number of localizations: 0% (top), 75% (middle) and 150% (bottom). For each case, distinct clusters were identified by StormGraph either using (+) or not using (-) localization uncertainties and using either the heuristic method (auto.) or k-nearest neighbor ($k = 15$) method to set r_0. Localization clusters possibly arising from overcounted single molecules were not removed. Color bars = cluster area (nm^2)). Scale bar = 500 nm. **(b)** Areas of all distinct clusters identified by StormGraph in the ROI when different numbers of randomly distributed points were added (up to 200% of the original number of localizations). Results are for the four implementations of StormGraph in the corresponding columns of panel (a).
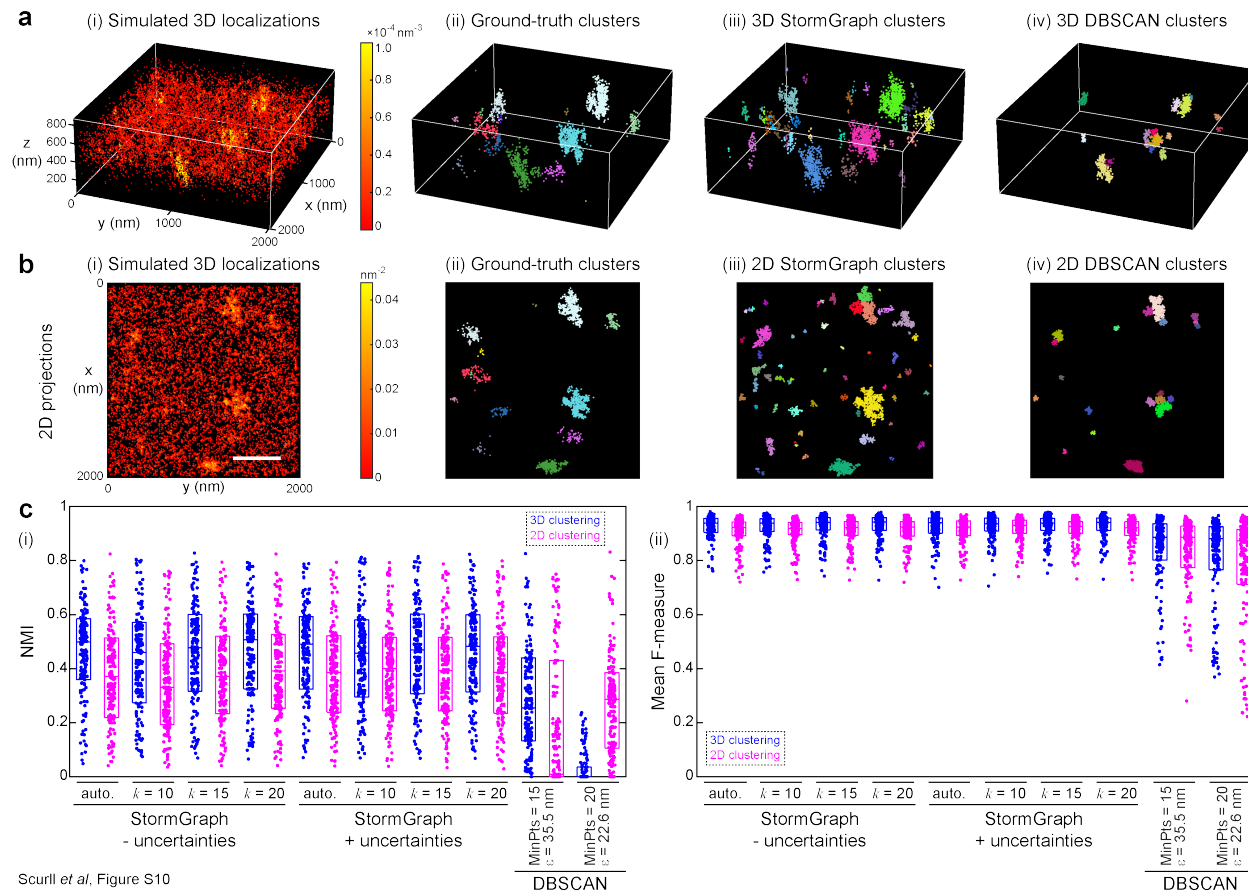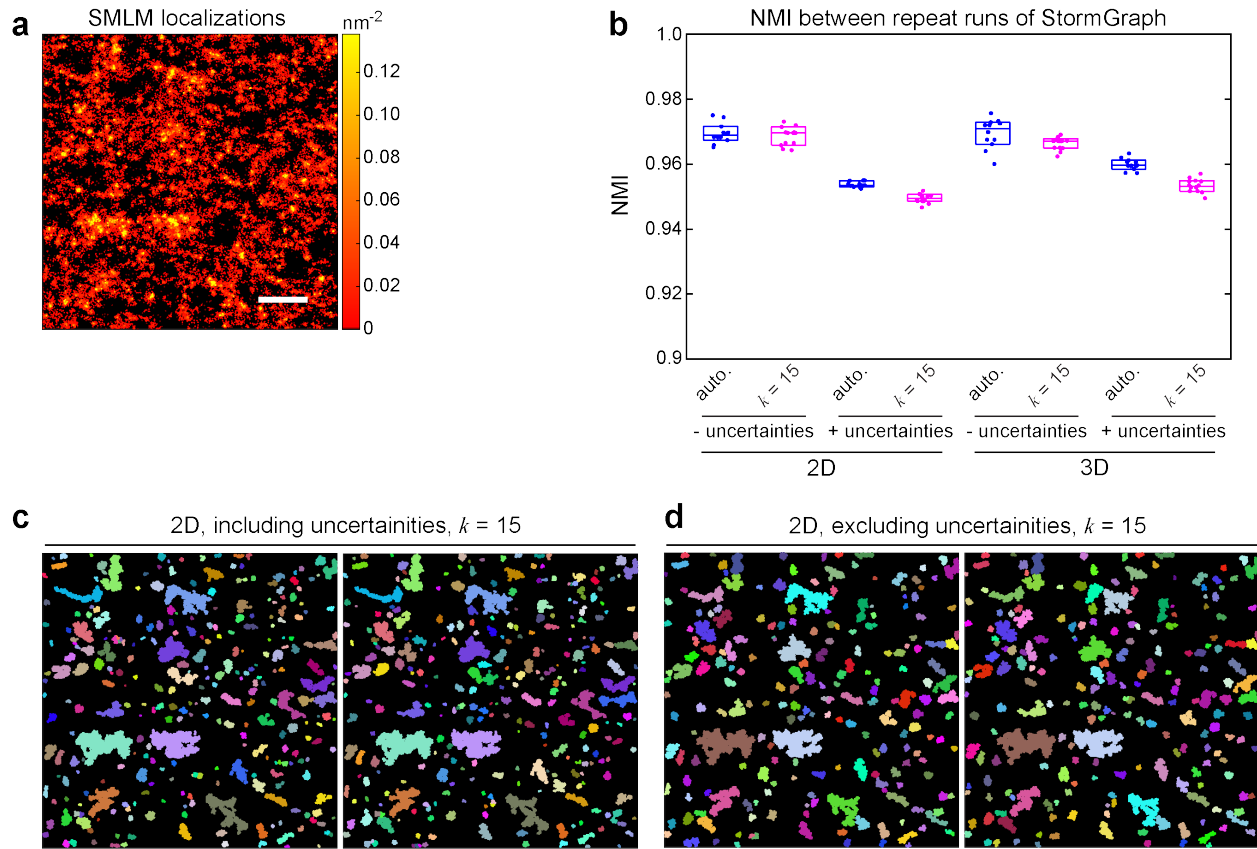
Figure S10

Figure S10. StormGraph identifies clusters in 3D simulated data more accurately than DB-SCAN. (a) (i) Simulated 3D SMLM data example. Color bar = 3D density. In this example, the density of molecules outside of clusters was 5% of the within-cluster density of molecules, which still resulted in ~85% of localizations not belonging to any true cluster. We also tested examples with up to four times this relative density of unclustered molecules. (ii) Ground-truth clusters in the simulated example. Colors distinguish distinct clusters. (iii) Clusters identified by StormGraph implemented in 3D with localization uncertainties used during clustering, r_0 determined heuristically (auto.), $\alpha = 0.05$, and MinCluSize = 5 localizations. Colors distinguish distinct clusters. (iv) Clusters identified by DBSCAN implemented in 3D using MinPts = 15 and $\varepsilon = 35.5$ nm. These were the DBSCAN parameter values that had the best average NMI scores for 3D clustering of 130 simulated datasets. **(b)** (i–ii) 2D projections of (a)(i)–(ii) respectively onto the xy-plane. (iii–iv) Clusters identified, in the 2D projection of the 3D simulated example, by 2D implementations of (iii) StormGraph, and (iv) DBSCAN. StormGraph was implemented in 2D using the same settings as the 3D implementation in (a)(iii). DBSCAN was implemented in 2D using MinPts = 20 and $\varepsilon = 22.6$ nm. These were the DBSCAN parameter values that had the best average NMI scores for 2D clustering of 130 simulated datasets. Color bar = 2D density; scale bar = 500 nm. **(c)** Accuracy of cluster assignments by 3D and 2D implementations (applied, respectively, to the full 3D data or to the 2D xy projections) of StormGraph and DBSCAN for 130 simulated 3D datasets, as assessed by (i) normalized mutual information (NMI; 1=perfect) and (ii) mean F-measure (1=perfect). Boxes show medians and interquartile ranges. StormGraph was implemented using $\alpha = 0.05$ and MinCluSize = 5 localizations. Localization uncertainties were either used (+) or not used (-) during clustering, and r_0 was determined using the heuristic method (auto.) or using the

19

indicated values of k. DBSCAN was implemented using the indicated parameters. The first set of DBSCAN parameters was optimized for 3D clustering and the second set was optimized for 2D clustering.
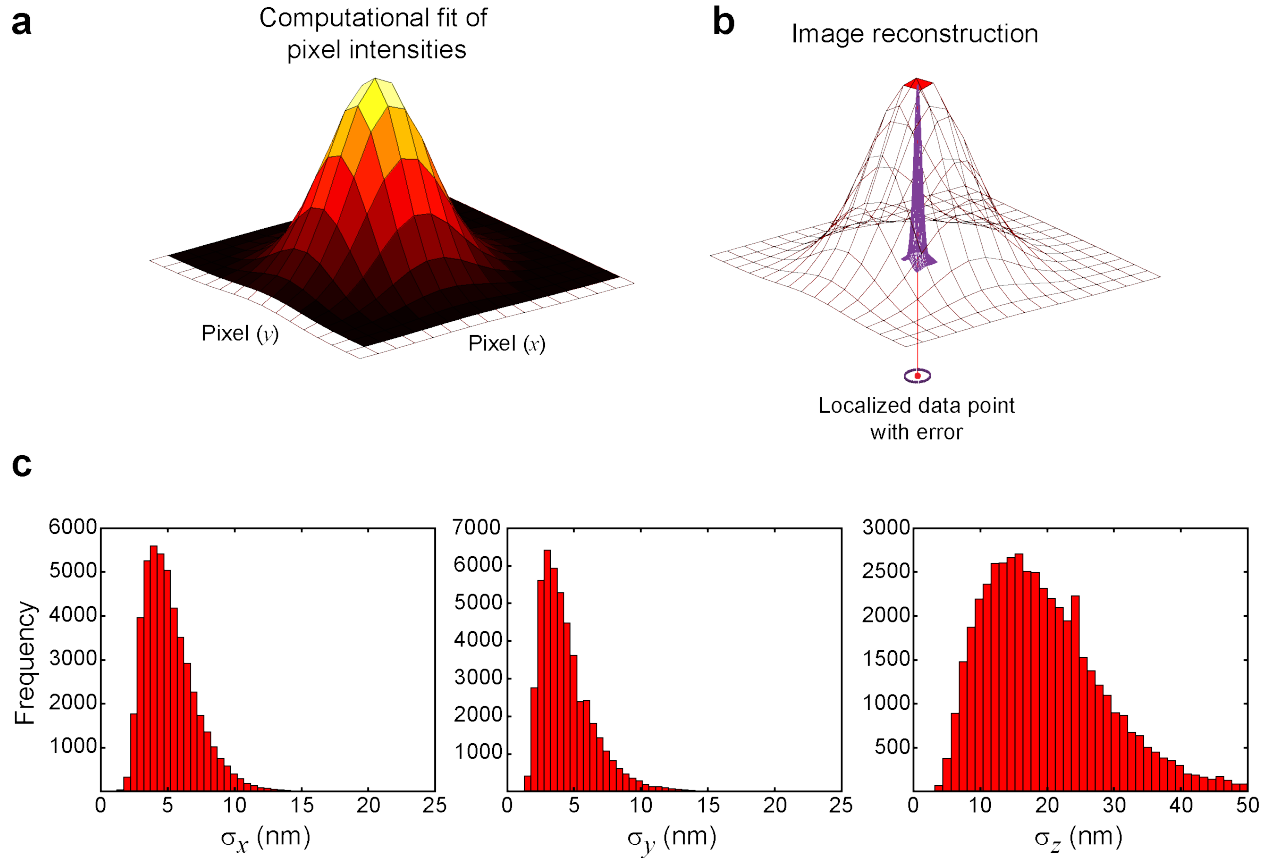
Figure S11

Figure S11. Identical repeat runs of StormGraph yield highly similar results. (a) 2D projection of a 3D ROI containing heterogeneous localizations of AF647-labeled cell-surface IgM-BCRs on a TMD8 B cell imaged by dSTORM. Scale bar = 500 nm, color bar = density (nm^{-2}). (b) Normalized mutual information (NMI) measuring the similarity between single-level cluster assignments of localizations by multiple identical runs of StormGraph. For each of the indicated StormGraph settings, StormGraph was run identically 11 times on the data in (a). The NMI values in the figure score the similarity of the cluster assignments of localizations by the first run of StormGraph to the cluster assignments of localizations by each of the 10 subsequent identical runs of StormGraph. StormGraph was implemented in 2D or 3D either not using (-) or using (+) localization uncertainties. The value of r_0 was determined using either the heuristic method (auto.) or the kNN method with $k = 15$. NMI values can range from 0 to 1, with NMI = 1 corresponding to perfect similarity. Boxes show medians and interguartile ranges. (c–d) Single-level clusters (distinguished by colors) identified by one run of StormGraph (left) and the *least similar* results of 10 subsequent identical runs of StormGraph using the indicated settings.

Figure S12

a

Computational fit of
pixel intensities

Pixel (y) Pixel (x)

b

Image reconstruction

Localized data point
with error

c

Figure S12. Computational localization and positional uncertainty estimation of dSTORM fluorescence emission events. **(a)** A point spread function (PSF) is fit to the fluorescence intensity profile of each SMLM blink. **(b)** The localization coordinates of each fluorescence emission event and their associated uncertainties (i.e. estimated measurement errors) are determined simultaneously from the computational fit of the PSF. **(c)** Distributions of the x, y and z positional uncertainties, expressed as standard deviations, in dSTORM localizations in a representative dataset.