Bayesian cluster identification in single-molecule localization microscopy data

Patrick Rubin-Delanchy[1], Garth L Burn[2], Juliette Griffié[2], David J Williamson[3], Nicholas A Heard[4], Andrew P Cope[5] & Dylan M Owen[2]

Single-molecule localization-based super-resolution microscopy techniques such as photoactivated localization microscopy (PALM) and stochastic optical reconstruction microscopy (STORM) produce pointillist data sets of molecular coordinates. Although many algorithms exist for the identification and localization of molecules from raw image data, methods for analyzing the resulting point patterns for properties such as clustering have remained relatively under-studied. Here we present a model-based Bayesian approach to evaluate molecular cluster assignment proposals, generated in this study by analysis based on Ripley's K function. The method takes full account of the individual localization precisions calculated for each emitter. We validate the approach using simulated data, as well as experimental data on the clustering behavior of CD3ζ, a subunit of the CD3 T cell receptor complex, in resting and activated primary human T cells.

Conventional fluorescence microscopes produce images of the distribution of fluorophores in the sample convolved with the microscope point-spread function (PSF). Owing to diffraction, this PSF typically has a width of hundreds of nanometers, meaning the resulting image has a resolution, as assessed by the Rayleigh criterion, of ~200 nm. Several strategies now exist to circumvent this resolution limit[1]. Some of these, such as stimulated emission depletion (STED) microscopy, rely on narrowing the excitation spot of a confocal microscope by means of a toroidal depletion beam and the process of stimulated emission[2,3]. Despite the increased resolution, these produce conventional fluorescence images, i.e., arrays of pixels with values representing the fluorescence intensity at those locations. Quantification can be performed in the same way as for conventional microscopes.

Another strategy is based on single-molecule localization microscopy (SMLM)[4–7]. This relies on the temporal separation of the excitation of fluorophores in the sample whose PSFs would otherwise overlap at the detector. The position of each fluorophore can then be estimated from the centers of the PSFs. Many algorithms are available to extract the x and y coordinates of the molecules[8–10]. Each emitter can be localized to a precision of 10–30 nm. Common strategies for the temporal separation of molecules involve intramolecular rearrangements to switch from dark to fluorescent states or the exploitation of non-emitting molecular radicals[11,12]. These strategies are typically pursued with photoactivatable or photoconvertible fluorescent proteins or small molecule probes coupled with a reducing buffer and immunostaining protocols[13]. We refer to all such strategies as SMLM.

Unlike non-pointillist microscopy methods, SMLM imaging does not produce a conventional image. Instead, the raw data are a list of the x and y coordinates of all the fluorophores, each with an associated, estimated localization precision. The analysis of spatial point patterns (SPPs) requires a different statistical tool kit to the analysis of pixel arrays.

Several techniques for analyzing SPPs generated from SMLM have been proposed. Ripley's K function[14–16] and pair-correlation (PC) analysis[17,18] are used widely to investigate and quantify clustering behavior. Both rely on drawing a series of concentric shapes—circles, in the case of the K function, and tori, in the case of PC—around each localization and counting the number of neighbors enclosed. These allow the degree of clustering at different spatial scales to be determined. In the case of the K function, the values at each localization can be interpolated to create cluster maps to which thresholds can then be applied[19].

These methods have several shortcomings. They often require calibration data or user-selected analysis parameters that strongly influence the output. This problem is exacerbated by batch processing, meaning that regions are often analyzed with the same suboptimal parameters. The methods also do not take any account of the individual localization precisions for each point. Finally, these are model-free methods, making their performance inherently difficult to judge and their results difficult to interpret.

Here we present a model-based, Bayesian approach to cluster analysis of SPPs generated by SMLM. The quality of a given assignment of molecules to clusters is evaluated against its (marginal) posterior probability, which is computed on the basis of a fully specified model for the data, including the localization precisions. This provides a principled mechanism for choosing between clustering proposals generated by different algorithms and settings. In this study, clustering proposals are generated through a strategy

[1]School of Mathematics, Heilbronn Institute for Mathematical Research, University of Bristol, Bristol, UK. [2]Department of Physics and Randall Division of Cell and Molecular Biophysics, King's College London, London, UK. [3]Manchester Collaborative Centre for Inflammation Research, University of Manchester, Manchester, UK. [4]Department of Mathematics, Imperial College London, London, UK. [5]Division of Immunology, Infection and Inflammatory Disease, Academic Department of Rheumatology, King's College London, London, UK. Correspondence should be addressed to patrick.rubin-delanchy@bristol.ac.uk or dylan.owen@kcl.ac.uk.

RECEIVED 5 MARCH; ACCEPTED 2 SEPTEMBER; PUBLISHED ONLINE 5 OCTOBER 2015; DOI:10.1038/NMETH.3612

based on the K function[14], with variable spatial scale and threshold. We generated several thousand candidate clustering proposals per region of interest (ROI), from which the optimum is selected according to the Bayesian model (**Supplementary Software**).

We demonstrate using simulated SPP data that we can accurately evaluate molecular clustering in a variety of conditions. We then use the technique to compare the clustering behavior of CD3ζ fused to the photoswitchable fluorescent protein mEos3.2 (CD3ζ-mEos3.2) in resting T cells to that at the T cell immunological synapse. This analysis of clustering in T cells is informative; both the K function strategy and the PC have been applied to it previously[14,15,20,21], enabling comparison with the performance of our approach.

For experimental data, it is important that artifacts caused by multiple blinking of individual fluorophores and overlapped PSFs, inherent to the methodology of SMLM, are removed (or accounted for). Our algorithm does not attempt to correct for, or be robust to, multiple blinking. Therefore, our method generates quantitatively reliable results only when multiple blinking has been corrected, as is possible with PALM data. We achieved this using the method of Annibale *et al.*[22,23], previously validated using mEos, and by fitting multiple emitters to overlapping PSFs. Our algorithm is applicable to data from other SMLM implementations, such as STORM, but because of the difficulties of correcting multiple blinking, results may be difficult to interpret.

RESULTS

Description of the algorithm

The algorithm is based on a full generative model for the data. We begin by assuming a single coordinate for each molecule in the ROI, generated by localization software (in our case, ThunderSTORM)[24]. The two-dimensional (2D) molecular positions are modeled as a set of Gaussian distributed clusters overlaid on a completely spatially random (CSR) background. These molecular coordinates are then disturbed by Gaussian distributed errors as a result of the localization process. The errors have different s.d. values, which are estimated from the raw microscopy data on the basis of the number of collected photons, PSF width, local background noise and camera pixel size[25].

The cluster centers themselves are assumed to be uniformly distributed over the ROI, and their radii (s.d.) are drawn from a user-supplied prior distribution. The algorithm is not sensitive to plausible values for this prior, and the distribution can therefore be specified from preliminary analysis, literature or visual inspection of the data. Localizations are assigned independently to the CSR background with a fixed prior probability, and the remaining localizations are clustered according to the Dirichlet process[26]. We compute the posterior probability of any given assignment of localizations to clusters (a clustering proposal) with respect to the above model. The calculation is deterministic, unlike with many Bayesian models, requiring only numerical integration over one dimension (Online Methods).

To generate clustering proposals we use a method based on Ripley's K function[16]. Every localization is allocated a clustering score L, as proposed by Getis and Franklin[27]. L is a function of the number of localizations within a distance, r, of that point, normalized by the mean molecular density of the ROI. Localizations with a value of L below a certain threshold T are assigned to the background. T can be interpreted as the minimum local density required for a point to be assigned to a cluster. Any two remaining localizations within a distance $2r$ of each other are then connected and the connected components form clusters (**Fig. 1a**). By varying r and T we generate of the order of 10,000 cluster proposals, which are then assigned a posterior probability. The algorithm identifies the highest-scoring proposal and extracts key descriptors. Although other methods for generating cluster proposals are possible, for example, K means[28], kernel density estimation (KDE) clustering[29], agglomerative clustering[30] or density-based spatial clustering of applications with noise (DBSCAN)[31], Ripley's K function is attractive because it has a straightforward geometrical interpretation and can be rapidly computed. The use of other algorithms for generating proposals would require modification of the code.

In a representative simulated data set (**Fig. 1b**), we calculated the posterior probability for a range of values of r and T (**Fig. 1c**). The dashed line indicates positions where the value of $L(r)$ for each localization is thresholded at r, i.e., where $T = r$. $L(r)$ being greater (or smaller) than r indicates that points are more (or less) clustered at that scale than would be expected under CSR. It is intuitive that a clustering model should favor thresholding $L(r)$ above r. We selected four combinations of r and T and generated clustering proposals from each (**Fig. 1d**). The highest scoring

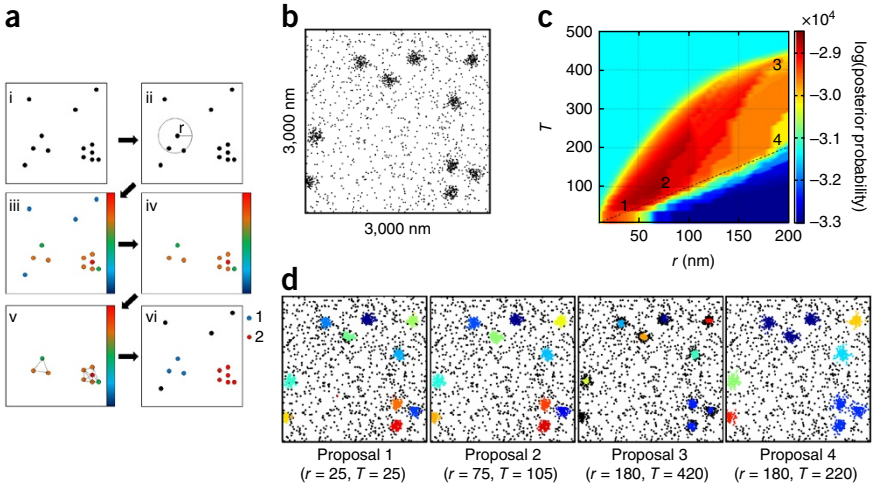

Figure 1 | Workflow of the algorithm. (**a**) The cluster proposal–generating mechanism. (**a**, **i**) A raw data set consists of molecular localizations within a square ROI. (**a**, **ii**) Getis and Franklin's method[27] counts the number of localizations within a distance r of each point. (**a**, **iii**) Each localization is assigned a score, $L(r)$, in pseudo-color. (**a**, **iv**) Values are then thresholded and (**a**, **v**) those falling above are grouped into clusters by connecting any pair whose circles intersect. (**a**, **vi**) All localizations are then given a cluster number or assigned to the background, culminating in an overall cluster proposal. The algorithm searches through many combinations of r and T to generate thousands of cluster proposals. (**b**) A representative simulated data set (from $n = 100$) from the standard conditions. (**c**) A pseudo-colored heat map showing the log(posterior probability) for a range of values of r and T. Red represents the most probable combinations according to the model. The dashed line represents $T = r$. (**d**) Four combinations of r and T and their corresponding proposals from the map in **c**. The highest-scoring combination generates proposal 2.

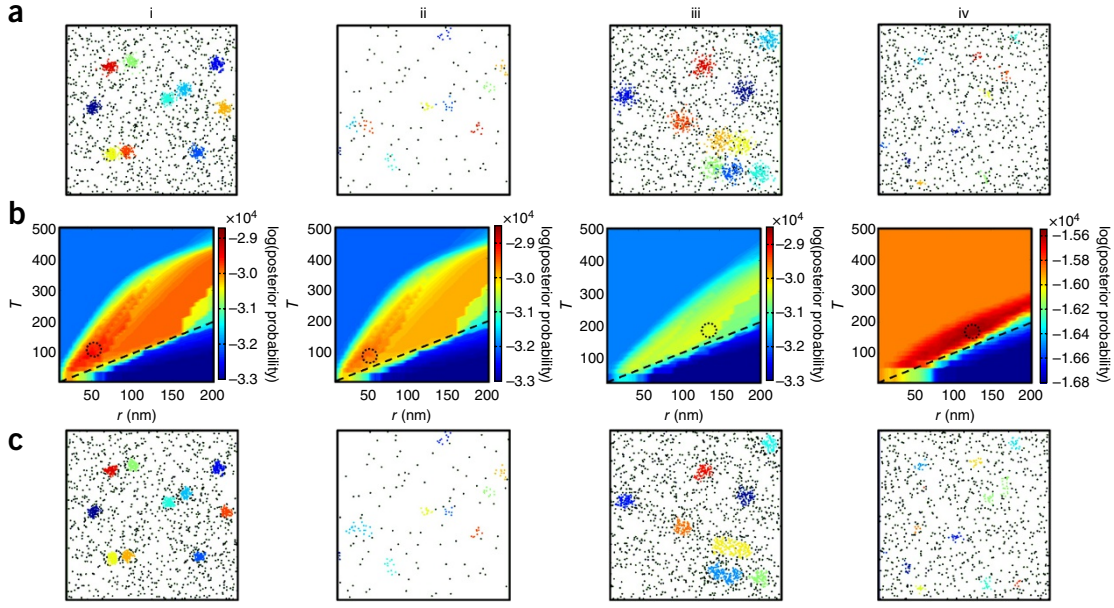

Figure 2 | Four different clustering scenarios. (**a**–**c**) Representative simulated data (**a**), log(posterior probability) heat maps showing $T = r$ (dashed lines) and the highest-scoring combination (circles) (**b**) and the highest-scoring cluster proposal (**c**) for standard conditions (**a**–**c**, **i**), a sparse data set with only 10% as many localizations as the standard conditions (**a**–**c**, **ii**), a data set where clusters are twice the size of those in the standard conditions (**a**–**c**, **iii**) and a data set with 10 localizations per cluster and 90% of localizations in the background (**a**–**c**, **iv**). Histograms of the outputted cluster descriptors for these conditions are shown in **Supplementary Figure 1**.

was proposal 2. The other proposals illustrate different manifestations of a suboptimal selection of r and T. In proposal 1, several small, spurious clusters were identified, largely owing to the small value of r used. In proposal 3, the threshold was too stringent, and localizations at the cluster extrema were assigned to the background. Finally, in proposal 4, clusters were merged owing to a large value of r.

Performance and sensitivity analysis

We tested the performance of our approach using SMLM localization data simulated under four different clustering scenarios, with 100 ROIs per scenario. In the first, the standard conditions, a 3,000 × 3,000 nm area contains 2,000 localizations. These comprise 10 Gaussian clusters with a radius of 50 nm containing 100 localizations each and 1,000 localizations (50%) in the background. Each localization is then disturbed by Gaussian noise with variance drawn from a gamma distribution of 30 ± 13 nm (mean and s.d.) (emulating the localization error of the microscope). We chose these parameters to reflect approximately the typical clustering behavior of proteins at the immunological synapse[14,15,20,21]. The three remaining scenarios have the same parameters as the standard conditions, with the following exceptions: the second scenario is a sparse data set containing only 200 localizations with 10 per cluster and 100 in the background; in the third scenario, the cluster radii are 100 nm; and the fourth scenario has 10 localizations per cluster with 900 (90%) in the background.

We generated examples of each simulated scenario (**Fig. 2a**) with corresponding heat maps showing the log(posterior probability) (**Fig. 2b**), the highest-scoring r-T combinations and the proposals generated (**Fig. 2c**). These proposals closely matched the simulated data sets, indicating accurate cluster assignments. Histograms of four key cluster descriptors—number of clusters per region, cluster radii (empirical s.d. of the localizations), number of localizations per cluster and percentage of localizations in clusters—showed close alignment to the simulated values (**Supplementary Fig. 1**).

We also tested our algorithm on simulated data generated with varying input parameters. We varied the number of localizations per cluster, cluster size and the fraction of unclustered background individually (**Supplementary Fig. 2**) while keeping all the other parameters constant. The analysis demonstrated that the algorithm is robust to variation in the number of localizations per cluster over several orders of magnitude, but we suggest six localizations per cluster as the low-end detection limit. The algorithm is also robust to background localizations up to around 80% of localizations in the background. A limitation on accurate determination of the cluster radii is the size of the selected ROI.

We compared our algorithm to other currently available methods of cluster analysis, specifically local point pattern analysis[27], DBSCAN, Ripley's K function and PC-like analysis (**Supplementary Figs. 3**–**5** and Online Methods). Ripley's K function and PC analysis provide an overview of the clustering behavior in each case and confirm the presence of clusters, but local point pattern analysis and DBSCAN produce full clustering descriptors (for example, percentage of localizations in clusters) and are therefore directly comparable to our algorithm. In each case, our algorithm was more accurate (histogram peak was closer to the expected value) than local point pattern analysis and DBSCAN.

We also tested the performance of our algorithm against more challenging conditions, including an uneven background (modeled as a beta distribution) (**Supplementary Figs. 6** and **7**), very small clusters (multimers containing two, three or six members) (**Supplementary Fig. 8**) or clusters with variable size (**Supplementary Fig. 9**). We found that the algorithm is insensitive to uneven background and correctly identifies populations of clusters with varying sizes. Dimers and trimers were not accurately identified; we therefore suggest six localizations per cluster as the low-end detectability limit of our algorithm.

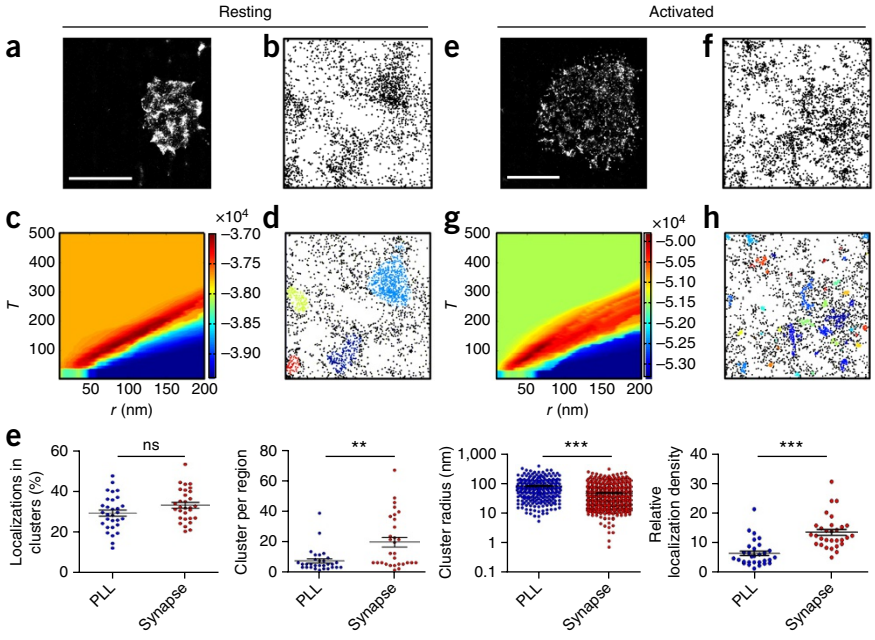

Figure 3 | Comparison of the clustering behavior of CD3ζ-mEos3.2 in primary human T cells resting on poly-L-lysine (PLL) or forming synapses (activated). (**a**,**e**) Representative SMLM images (from 19 resting cells and 10 activated cells); scale bars, 5 µm. (**b**,**f**) Examples of 3,000 × 3,000 nm regions showing localization coordinates. (**c**,**g**) log(posterior probability) heat maps showing the highest-scoring *r*-*T* combination for a representative ROI of 3,000 × 3,000 nm. (**d**,**h**) Highest-scoring cluster proposals. (**e**) Plots of the percentage of localizations in clusters (one point per ROI), number of clusters per region (one point per ROI), the cluster radii (one point per cluster) and the relative density of localizations in clusters as compared to the surrounding region (one point per ROI). $*P \leq 0.05$; $**P \leq 0.005$; $***P \leq 0.0005$; ns, not significant (n = 30 ROIs per condition). Horizontal bars indicate mean and s.e.m.

We performed a side-by-side comparison of our algorithm with local point pattern analysis and DBSCAN on three example conditions (standard conditions, larger clusters and uneven background) (**Supplementary Fig. 10**). In each case, the histograms of the cluster parameters peaked closer to the simulated values with our method than with the alternatives, demonstrating the superiority of our approach. A sensitivity analysis to prior settings is also provided (**Supplementary Fig. 11**).

Analysis of protein clustering in primary human T cells

We analyzed SMLM data of CD3ζ-mEos3.2 at the plasma membrane of fixed $CD4^+$ primary human T cells that had formed an immunological synapse on glass coverslips coated with antibodies to CD3 (anti-CD3) and anti-CD28. We used nonactivating poly-L-lysine (PLL)-coated coverslips as a control. Although mEos3.2 displays multiple blinking during PALM data acquisition, we compensated for this effect using ThunderSTORM localization software[24] and optimal settings estimated using the method of Annibale *et al.*[23] (**Supplementary Fig. 12** and Online Methods). The localization precision was calculated using the method of Quan *et al.*[25] (**Supplementary Fig. 12**).

From SMLM images of resting and activated T cells (**Fig. 3a**,**e**), we selected regions of 3,000 × 3,000 nm (n = 30 per condition) and plotted the localizations (**Fig. 3b**,**f**). For each, we generated log(posterior probability) heat maps (**Fig. 3c**,**g**) and plotted the highest-scoring proposals (**Fig. 3d**,**h**). We also generated beeswarm plots of the percentage of localizations in clusters, number of clusters per region, cluster radii and relative density of localizations inside and outside clusters (**Fig. 3e**). In agreement with previous reports[21,32], CD3ζ was clustered in stimulated and nonstimulated cells, and cluster parameters were significantly altered upon stimulation. Despite no large-scale changes in the percentage of localizations found in clusters (29% ± 2% (s.e.m.) in PLL to 33 ± 1% in activated cells, $P > 0.05$), we observed a significant increase in the number of clusters and a significant decrease in the size of clusters from resting to activated cells (8 ± 1 clusters per region on PLL versus 20 ± 3 in activated cells; $P \leq 0.005$ and 82 ± 4 nm radius versus 48 ± 2 nm after activation; $P \leq 0.0005$; **Fig. 3e**). We also observed a significant increase in the density of localizations in clusters relative to that in nonclustered regions (7 ± 1 versus 14 ± 1, $P \leq 0.0005$). The results were consistent when we divided localizations into two equally sized data sets (**Supplementary Fig. 13**).

We analyzed all experimental data using three well-established methods of cluster analysis local point pattern analysis, Ripley's K function and PC analysis) (**Supplementary Fig. 14**) to test the validity of the results obtained using our Bayesian algorithm. Ripley's K function and PC analysis also showed that CD3ζ is preclustered in resting cells. From Getis and Franklin's cluster maps, we observed that the number of clusters increases after stimulation. However, the maps were not sensitive to the change in cluster size detected by the Bayesian approach, presumably owing to the inability to optimize analysis parameters between conditions.

DISCUSSION

We have demonstrated a Bayesian cluster analysis algorithm for SMLM data. Unlike previously demonstrated methods based on the generation of cluster maps, which involve an interpolation algorithm to generate the surface[19], our approach is not prone to artifacts in sparse data sets, such as those from low-copy-number proteins. The method also has the possibility of allowing faster imaging, as fewer localizations are required to accurately identify and characterize clustering. Increasing the speed of SMLM data acquisition and processing has been one of the major goals to move the technique into the domain of live-cell imaging[33], and the algorithm presented here is compatible with live-cell SMLM data.

The algorithm is only weakly sensitive to the prior settings, offering a major advantage over previous methods, in which the initial choice of spatial scale and threshold has a large effect on the final result. In addition, in our method all ROIs are analyzed with the parameters estimated to be optimal for that specific region, rather than diverse regions being treated equally. To our knowledge, the method is the first to take full account of the localization precisions rather than treating all localizations as exact.

The detectability limit of our algorithm is around six localizations per cluster (**Supplementary Figs. 2** and **8**); therefore, the

algorithm is not suited to the quantification of small complexes such as dimers and trimers. We hypothesize that a Bayesian model that explicitly targets small features would be more successful in detecting small multimers. Indeed, an interesting avenue of future research would be to develop a number of different models to capture the diversity of point patterns observed in SMLM data, such as fibers, meshes and areas of exclusion.

It is well known that raw SMLM data can exhibit artifacts owing to the photophysics of the process[22,23,34], wherein individual molecules can be re-excited and thus generate multiple localizations. In addition, owing to the stochastic nature of the activation process, it is possible for several PSFs to overlap at the detector, causing errors in the extracted coordinates. Our algorithm does not attempt to correct or be robust to multiple-blinking effects. If there is suspicion that these have not been adequately addressed by localization software, then the output of our algorithm should be interpreted with caution. In our case, we acquired experimental data using PALM, for which multiple blinking can be corrected owing to the different timescales of molecular photoconversion and photoblinking. In other experimental conditions—for example, when small-molecule dyes are used—such corrections may not be possible. Therefore, the outputs of the algorithm may contain artifacts, particularly spurious clusters. Our algorithm remains a valuable exploratory tool for such data.

Our method allows the accurate and principled quantification of clustering behavior in SMLM data in a manner that is more automatic, robust and objective than previously possible. We have focused on a model consisting of circular, Gaussian-distributed clusters overlaid on a CSR background. In the future, it will be possible to create generative models with different clustering characteristics. Evaluation of SMLM data against such models may allow a better understanding of the biophysical principles underlying protein clustering.

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

D.M.O. acknowledges funding from the European Research Council (FP7 starter grant 337187) and Marie Curie Career Integration grant 334303. A.P.C. is funded by Arthritis Research UK grants 19652 and 20525.

AUTHOR CONTRIBUTIONS

P.R.-D., N.A.H. and D.M.O. conceived the method. P.R.-D., J.G. and D.M.O. performed the analysis. P.R.-D. and D.M.O. wrote the manuscript. G.L.B. acquired cell data. G.L.B., D.J.W. and A.P.C. provided materials.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Huang, B. Super-resolution optical microscopy: multiple choices. *Curr. Opin. Chem. Biol.* **14**, 10–14 (2010).
2. Hell, S.W. & Wichmann, J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Opt. Lett.* **19**, 780–782 (1994).
3. Chmyrov, A. *et al.* Nanoscopy with more than 100,000 'doughnuts'. *Nat. Methods* **10**, 737–740 (2013).
4. Betzig, E. *et al.* Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313**, 1642–1645 (2006).
5. Rust, M.J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3**, 793–795 (2006).
6. Heilemann, M. *et al.* Subdiffraction-resolution fluorescence imaging with conventional fluorescent probes. *Angew. Chem. Int. Ed. Engl.* **47**, 6172–6176 (2008).
7. Hess, S.T., Girirajan, T.P.K. & Mason, M.D. Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophys. J.* **91**, 4258–4272 (2006).
8. Wolter, S. *et al.* rapidSTORM: accurate, fast open-source software for localization microscopy. *Nat. Methods* **9**, 1040–1041 (2012).
9. Holden, S.J., Uphoff, S. & Kapanidis, A.N. DAOSTORM: an algorithm for high-density super-resolution microscopy. *Nat. Methods* **8**, 279–280 (2011).
10. Henriques, R. *et al.* QuickPALM: 3D real-time photoactivation nanoscopy image processing in ImageJ. *Nat. Methods* **7**, 339–340 (2010).
11. van de Linde, S. *et al.* Direct stochastic optical reconstruction microscopy with standard fluorescent probes. *Nat. Protoc.* **6**, 991–1009 (2011).
12. Heilemann, M., van de Linde, S., Mukherjee, A. & Sauer, M. Super-resolution imaging with small organic fluorophores. *Angew. Chem. Int. Ed. Engl.* **48**, 6903–6908 (2009).
13. Dempsey, G.T. *et al.* Photoswitching mechanism of cyanine dyes. *J. Am. Chem. Soc.* **131**, 18192–18193 (2009).
14. Williamson, D.J. *et al.* Pre-existing clusters of the adaptor Lat do not participate in early T cell signaling events. *Nat. Immunol.* **12**, 655–662 (2011).
15. Rossy, J., Owen, D.M., Williamson, D.J., Yang, Z. & Gaus, K. Conformational states of the kinase Lck regulate clustering in early T cell signaling. *Nat. Immunol.* **14**, 82–89 (2013).
16. Ripley, B.D. Modelling spatial patterns. *J. R. Stat. Soc. Series B Stat. Methodol.* **39**, 172–192 (1977).
17. Sengupta, P. *et al.* Probing protein heterogeneity in the plasma membrane using PALM and pair correlation analysis. *Nat. Methods* **8**, 969–975 (2011).
18. Veatch, S.L. *et al.* Correlation functions quantify super-resolution images and estimate apparent clustering due to over-counting. *PLoS ONE* **7**, e31457 (2012).
19. Owen, D.M. *et al.* PALM imaging and cluster analysis of protein heterogeneity at the cell surface. *J. Biophotonics* **3**, 446–454 (2010).
20. Sherman, E. *et al.* Functional nanoscale organization of signaling molecules downstream of the T cell antigen receptor. *Immunity* **35**, 705–720 (2011).
21. Lillemeier, B.F. *et al.* TCR and Lat are expressed on separate protein islands on T cell membranes and concatenate during activation. *Nat. Immunol.* **11**, 90–96 (2010).
22. Annibale, P., Vanni, S., Scarselli, M., Rothlisberger, U. & Radenovic, A. Identification of clustering artifacts in photoactivated localization microscopy. *Nat. Methods* **8**, 527–528 (2011).
23. Annibale, P., Vanni, S., Scarselli, M., Rothlisberger, U. & Radenovic, A. Quantitative photo activated localization microscopy: unraveling the effects of photoblinking. *PLoS ONE* **6**, e22678 (2011).
24. Ovesný, M., Křížek, P., Borkovec, J., Švindrych, Z. & Hagen, G.M. ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics* **30**, 2389–2390 (2014).
25. Quan, T., Zeng, S. & Huang, Z.-L. Localization capability and limitation of electron-multiplying charge-coupled, scientific complementary metal-oxide semiconductor, and charge-coupled devices for superresolution imaging. *J. Biomed. Opt.* **15**, 066005 (2010).
26. Ferguson, T.S. A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1**, 209–230 (1973).
27. Getis, A. & Franklin, J. Second-order neighborhood analysis of mapped point patterns. *Ecology* **68**, 473–477 (1987).
28. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–137 (2006).
29. Hinneburg, A. & Gabriel, H.-H. in *Advances in Intelligent Data Analysis VII* (eds. Berthold, M.R., Shawe-Taylor, J. & Lavrač, N.) 70–80 (Springer, 2007).
30. Johnson, S.C. Hierarchical clustering schemes. *Psychometrika* **32**, 241–254 (1967).
31. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD-96* 226–231 (1996).
32. Neve-Oz, Y., Razvag, Y., Sajman, J. & Sherman, E. Mechanisms of localized activation of the T cell antigen receptor inside clusters. *Biochim. Biophys. Acta* **1853**, 810–821 (2015).
33. Cox, S. *et al.* Bayesian localization microscopy reveals nanoscale podosome dynamics. *Nat. Methods* **9**, 195–200 (2012).
34. Lee, S.-H., Shin, J.Y., Lee, A. & Bustamante, C. Counting single photoactivatable fluorescent molecules by photoactivated localization microscopy (PALM). *Proc. Natl. Acad. Sci. USA* **109**, 17436–17441 (2012).

ONLINE METHODS

Sample preparation. Primary human T cells were isolated from peripheral blood using Lymphoprep (Stemcell) followed by a naïve CD4 T cell negative selection kit (Miltenyi). 1×10^7 naïve human CD4 T cells were transfected with 2 µg CD3ζ-mEos3.2 using an Amaxa system (Lonza). Immune synapses were formed against activating coverslips coated with anti-CD3 (2 µg/mL) and anti-CD28 (5 µg/mL) (BioLegend, catalog numbers 317315 and 102116, respectively). Immune synapses were allowed to form for 5 min and were then pH-shift fixed (3% PFA and kPIPES at 80 mM, pH 6.8 for 5 min followed by 3% PFA and borax at 100 mM for 10 min). Cells were imaged in PBS.

SMLM imaging. SMLM imaging was performed on a Nikon N-STORM microscope using a 100×/1.49 numerical aperture (NA) oil-immersion total internal reflection fluorescence (TIRF) objective. Cells were imaged under TIRF illumination with a 561-nm laser with photoactivation at 405 nm. Fluorescence was collected in the wavelength range 575–625 nm on an Andor iXon DU897U electron-multiplying charge-coupled device (EM-CCD) camera. Camera integration time was 30 ms, and a total of 20,000 frames were typically recorded. Molecular coordinates were localized using ThunderSTORM software[24]. ThunderSTORM is able to separate multiple overlapped PSFs and compensates for the multiple blinking of individual fluorophores given a user supplied merge time and distance. Up to four overlapped Gaussian PSFs were allowed. The optimal merge time was computed following the analysis method of Annibale *et al.* as described for mEos[23] (**Supplementary Fig. 12**). This has been shown to generate reliable molecular localization coordinates free from multiple blinking effects and therefore appropriate for input into our algorithm. The distance is determined by the camera pixel size, in this case 100 nm. The merge time was determined to be three frames (90 ms). The photon threshold for single-molecule identification was set at 500. ThunderSTORM corrects for sample drift during the acquisition using an autocorrelation approach. To calculate the localization precision for each emitter, we selected the method of Quan *et al.*,[25] which accounts for the specific noise statistics of EM-CCD cameras.

Permutation test. Significance for the changes in clustering properties of CD3ζ in resting and activated primary human T cells was calculated on the basis of a permutation test. This assumes only that the data are independent (in fact, exchangeable) under the null hypothesis. The test statistic T is given by the difference of the means of the two groups of values under analysis, X_1 and X_2

$$T = |\bar{X}_1 - \bar{X}_2|$$

The P value of this test is the frequency that $T^* \geq T$, where T^* is a simulated test statistic under the null hypothesis, that is

$$T^* = |\overline{X^*_1} - \overline{X^*_2}|$$

where X^*_1 and X^*_2 are two simulated groups constructed by sampling from the pooled values without replacement. Through the procedure described by Gandy[35,36], we were able to bound the probability due to simulation error of reporting a P value to be on the wrong side of the threshold (0.05, 0.005 or 0.0005) to $1/10^6$.

Bayesian cluster model for SMLM super-resolution data. *The model.* The data are a vector of 2D observed localizations $V = [V_1,\ldots,V_N]$ with associated variances $s_1^2,\ldots, s_N^2$. V_i values are assumed to be independent, unbiased and circular Gaussian observations of the true molecular locations $Z = [Z_1,\ldots,Z_N]$. Hence $V_i \sim \mathcal{N}(Z_i, s_i^2\boldsymbol{I}_2)$, where $\mathcal{N}(\mu,\Sigma)$ denotes the Gaussian distribution with mean μ and covariance Σ and $\boldsymbol{I}_d$ is the identity matrix of dimension d.

The molecular locations themselves follow a mixture of a clustered distribution (itself a mixture) and a completely spatially random background process. The membership of each molecule to a cluster is represented by a vector of integer labels, $\ell = [\ell_1,\ldots\ell_N]$, where $\ell_i = \ell_j$ indicates that Z_i and Z_j belong to the same cluster. As a convention, labels will be chosen sequentially starting from 0, so if there are $m + 1$ distinct entries in ℓ they are $0,\ldots,m$. For $k = 0,\ldots,m$, let c_k denote the set of localization indices associated with cluster k and let n_k denote the cardinality of c_k. Cluster 0 has a special role, representing localizations assigned to the background process.

Every point has a fixed, independent prior probability p_B of being generated by the background process. Non-background points are grouped using the Dirichlet process, as introduced by Ferguson[26]. This has one hyperparameter, α, called a concentration parameter. Following Green and Richardson[37], a proposed labeling ℓ therefore has the prior probability

$$p(\ell) = p_B^{n_0}(1 - p_B)^{N - n_0}.\frac{\alpha^m\Gamma(\alpha)\prod_{k=1}^{m}\Gamma(n_k)}{\Gamma(\alpha + N - n_0)}$$

Clusters are independent of each other. Points within a cluster are independent and identically distributed, conditional on any unknown parameters. Background points are uniformly distributed over the ROI. Points of a cluster $k \geq 1$ follow a circular Gaussian distribution with unknown mean and variance, $Z_i \sim \mathcal{N}(\mu_k, \sigma_k^2\boldsymbol{I}_2)$, $i \in c_k$. The parameters are *a priori* independent; μ_k is uniformly distributed over the region, whereas the prior on σ_k is supplied by the user. By default, a user-supplied histogram of s.d. is converted into a probability density by linear interpolation.

Inference. Every ℓ has a posterior probability given by

$$p(\ell\,|\,V) \propto p(\ell)\left[\prod_{i\in c_0} p_0(V_i).\prod_{k=1}^{m}\int p(\theta_k)\prod_{i\in c_k} p(V_i\,|\,\theta_k)d\theta_k\right]$$

where $\theta_k = (\mu_k, \sigma_k)$ are the parameters associated with cluster k, $p(\theta_k)$ is their prior density, $p(V_i|\theta_k)$ is the density of an observed localization given the parameters of its associated cluster and $p_0(V_i)$ is the background density.

The expression cannot be computed in closed form owing to the integral but can be obtained by numerical approximation. The solution we provide is for an arbitrary rectangular subset R of $\mathbb{R}^d$ that is aligned with the axes (for ease of notation). The corners of R are at $R_x^- \leq R_x^+$ for each coordinate x. Its volume is

$$A = \prod_{x=1}^{d}(R_x^+ - R_x^-)$$

so that $p(\theta_k) = A^{-1}p(\sigma_k)$ where, as previously mentioned, the second density is provided by the user.

We also require expressions for $p(V_i|\theta_k)$ and $p_0(V_i)$. Ignoring edge effects, we assume that if Z_i is generated by the background process, its localization precision is irrelevant, and let $p_0(V_i) = p_0(Z_i) = A^{-1}$. If Z_i is allocated to cluster k then we assume $V_i \sim \mathcal{N}(Z_i, s_i^2\boldsymbol{I}_d) \sim \mathcal{N}(\mu_k, (s_i^2+\sigma_k^2)\boldsymbol{I}_d)$.

Temporarily ignoring points outside cluster k, put $n = n_k$, $\mu = \mu_k$ and $\sigma = \sigma_k$ and, without loss of generality, assume that $c_k = \{1,\ldots,n\}$. The corresponding localizations form a vector $v = [V_1,\ldots,V_n]$. The integral therefore becomes

$$\int p(\sigma)\int p(\mu)\prod_{i=1}^{n} p(V_i\,|\,\mu,\sigma)d\mu\,d\sigma.$$

Let $\omega_i = 1/(\sigma^2 + s_i^2)$ and

$$\tilde{n} = \sum_{i=1}^{n}\omega_i$$

The likelihood of the points is

$$p(v\,|\,\mu,\sigma) = \prod_{i=1}^{n} p(V_i\,|\,\mu,\sigma) = (2\pi)^{-nd/2}\left(\prod_{i=1}^{n}\omega_i\right)^{d/2}$$
$$\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\omega_i(v_i-\mu)^T(v_i-\mu)\right\}$$

Let

$$\bar{v} = \tilde{n}^{-1}\sum_{i=1}^{n}\omega_i v_i$$

be the weighted center of the points and

$$S^2 = \sum_{i=1}^{n}\omega_i(v_i-\bar{v})^T(v_i-\bar{v})$$

The expression inside the exponential can be written as

$$-[S^2 + \tilde{n}(\mu-\bar{v})^T(\mu-\bar{v})]/2$$

Therefore, the inner integral becomes:

$$p(v\,|\,\sigma) = A^{-1}(2\pi)^{-\frac{nd}{2}}\left(\prod_{i=1}^{n}\omega_i\right)^{d/2}$$
$$e^{-\frac{S^2}{2}}\int_R \exp\{-\tilde{n}(\mu-\bar{v})^T(\mu-\bar{v})/2\}d\mu$$

The integral on the right hand side is equal to

$$\left(\frac{2\pi}{\tilde{n}}\right)^{\frac{d}{2}}\prod_{x=1}^{d}\left[\Phi\left(\sqrt{\tilde{n}}(R_x^+-\bar{v}_x)\right)-\Phi\left(\sqrt{\tilde{n}}(R_x^--\bar{v}_x)\right)\right]$$

where

$$\Phi(x) = \int_{-\infty}^{x}(2\pi)^{-1/2}\exp(-y^2/2)\,dy$$

is the cumulative distribution function of the standard normal distribution. It remains to compute the outer integral, or

$$\int p(\sigma)p(v\,|\,\sigma)d\sigma$$

This is performed by numerical integration. Because $p(v|\sigma)$ can be very peaked for moderately large data sets, we obtained the best results by evaluating $p(v|\sigma)$ at regularly interspaced points and using the midpoint rule approximation. One advantage of this approach is that it can be interpreted as using a discrete prior on σ. As is usual with Bayesian calculations, computation must be performed on the log scale.

Generating the proposals. Each localization V_i is assigned an $L(r)_i$ value given by

$$L(r)_i = \sqrt{A\sum_{j=1}^{N}\delta_{ij}/[\pi(N-1)]}$$

where $\delta_{ij} = 1$ if $i \neq j$ and the distance between V_i and $V_j \leq r$ and $\delta_{ij} = 0$ otherwise. Values of r were incremented every 5 nm from 5 to 200 nm (beyond which clusters would be resolvable by conventional microscopy). Localizations with $L(r)_i$ below a threshold T are assigned to the background process, whereas localizations above are grouped into clusters according to the connected components of a graph with edges between every two localizations that are less than $2r$ apart. The value of T was incremented from 0 to 500 in increments of 5.

Prior parameters used in experiments. The concentration parameter α was set to 20. The background probability p_B was set to 0.5. (The prior on σ_k is shown in **Supplementary Fig. 11c**,**i**).

Comparison to other methods. Other cluster analysis algorithms applied to SMLM data fall into two categories. Some, for example Ripley's K function or pair correlation, give a high-level overview of clustering statistics whereas others, such as DBSCAN and the method of Getis and Franklin (local point pattern analysis), derive a fully specified cluster assignment. Only algorithms of the second kind are directly comparable to ours. Despite this, we analyzed our simulated data using these four existing methods (**Supplementary Figs. 3–5**, **7** and **10**). For local point pattern analysis and DBSCAN, a fixed radius and threshold must be selected, which we chose (heuristically) to be optimal for standard conditions ($r = 50$, $T = 78$). The natural descriptors to use for comparison are the number of clusters per region and the percentage of localizations found in clusters.

35. Gandy, A. Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. *J. Am. Stat. Assoc.* **104**, 1504–1511 (2009).
36. Gandy, A. & Rubin-Delanchy, P. An algorithm to compute the power of Monte Carlo tests with guaranteed precision. *Ann. Stat.* **41**, 125–142 (2013).
37. Green, P.J. & Richardson, S. Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Stat.* **28**, 355–375 (2001).

doi:10.1038/nmeth.3612