SuperStructure: a Parameter-Free Super-Structure Identifier for SMLM Data

M. Marenda,[1,2] E. Lazarova,[2] S. van de Linde,[3] N. Gilbert,[2] and D. Michieletto[1,2]

[1]SUPA, School of Physics and Astronomy, University of Edinburgh, Peter Guthrie Road, Edinburgh, EH9 3FD, UK
[2]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine,
University of Edinburgh, Edinburgh EH4 2XU, UK
[3]SUPA, Department of Physics, University of Strathclyde, Glasgow G4 0NG, UK

Single-Molecule Localisation Microscopy (SMLM) allows the quantitative mapping of molecules at high resolution. However, understanding the non-random interaction of proteins requires the identification of more complex patterns than those typified by standard clustering tools. Here we introduce SuperStructure, a parameter-free algorithm to quantify structures made of inter-connected clusters, such as protein gels. SuperStructure works without a priori assumptions and is thus an ideal methodology for standardised analysis of SMLM data.

Single-molecule localisation microscopy [1–3] is now commonly employed for quantitative analysis of molecular interactions both in vivo [4–6] and in vitro [7, 8]. SMLM achieves resolutions far beyond the diffraction limit and its typical output is a list of 3D coordinates (or localisation events) that are naturally analysed using efficient clustering algorithms borrowed from quantitative big-data analysis and even astronomy [9–14]. However, clustering algorithms rely on user-defined parameters intrinsically intertwined with the notion of similarity used to define clusters. These parameters can be either assumed or inferred via pre-emptive analysis, yet their choice has a significant impact on the results. This hinders the portability and ease of comparison of different datasets.

At the same time, recent evidence suggest that assemblies of proteins [6, 15–20] and chromatin [21, 22] form complex structures that cannot be simply analysed with standard clustering algorithms. For example, the hnRNP protein SAF-A is suggested to form a dynamic mesh-like structure that maintains transcriptionally active genomic loci in a decompacted configuration [23, 24] whilst super-resolution techniques indicates that chromatin is organised in connected nano-scale compartments [25–27]. To understand the relationship between these structures and the dynamics and function of the genome [28–30], a more sophisticated and standardised analysis of SMLM data is required.

Motivated this, we have developed SuperStructure. This method extends the popular density-based clustering algorithm DBSCAN implementing (i) a parameter-free detection and quantification of super-structures made of connected clusters in SMLM data, (ii) a parameter-free quantification of the density of molecules within clusters. Here we show in detail that SuperStructure can be applied to discern and compare complex structures, such as nuclear networks or gels, emerging from different nuclear proteins or cellular conditions, and suggest it is well suited as tool for standardised SMLM analysis.

The working principle of SuperStructure is best ex-

plained as follows: while the classic DBSCAN algorithm detects clusters for a fixed choice of minimum number of neighbours $N_{\min}$ and neighbourhood distance ϵ [31], we argue that there is previously overlooked information in how the number of detected clusters N_c changes with ϵ for a fixed $N_{\min}$. By setting $N_{\min} = 0$, $N_c(\epsilon)$ is necessarily a monotonically decreasing function as for $\epsilon = 0$ every localisation is a cluster by itself and increasing ϵ yields fewer but larger clusters. Importantly, the rate at which N_c decays with ϵ is an indicator of how quickly localisations, and then clusters of localisations, coalesce with each other and thus of how "connected" they are.

The decay curves outputted by SuperStructure identify different clustering regimes: (i) merging localisations within clusters (intra-cluster regime); (ii) coalescing clusters into super-structures (first super-cluster regime); (iii) merging super-clusters into higher-order super-structures (second/third super-cluster regimes). The rate of decay of N_c in regime (i) is intimately related to the density of emitters ρ_{em} within the clusters (see Supplementary Note 2), while the decay in regimes (ii) and (iii) are highly dependent on the connectivity between (super-)clusters, as well as on the density of (super-)clusters (see Supplementary Note 3). These two contributions can be further discerned, as we shall see below. See Supplementary Note 1 for details of the algorithm and the pipeline.

We first evaluated the performance of SuperStructure on artificial datasets consisting of inter-connected clusters of simulated localisations (see also Supplementary Note 3). The main dataset is made by connected and spatially homogeneous clusters that are randomly positioned with a density to $\rho_{cl} = 8.2\mu m^{-2}$ and a radius $R_c \sim 40\ nm$. Pairs of clusters were connected with probability p by a sparse points distribution and only if the distance between the clusters was less than $b = 1\ \mu m$. This allowed us to readily tune the degree of "connectivity" in the system by varying a single parameter p (see also Supplementary Note 3 for another example). The length-scale associated to density of emitters inside clusters ρ_{em} and to the connections ρ_{conn} define the bound-
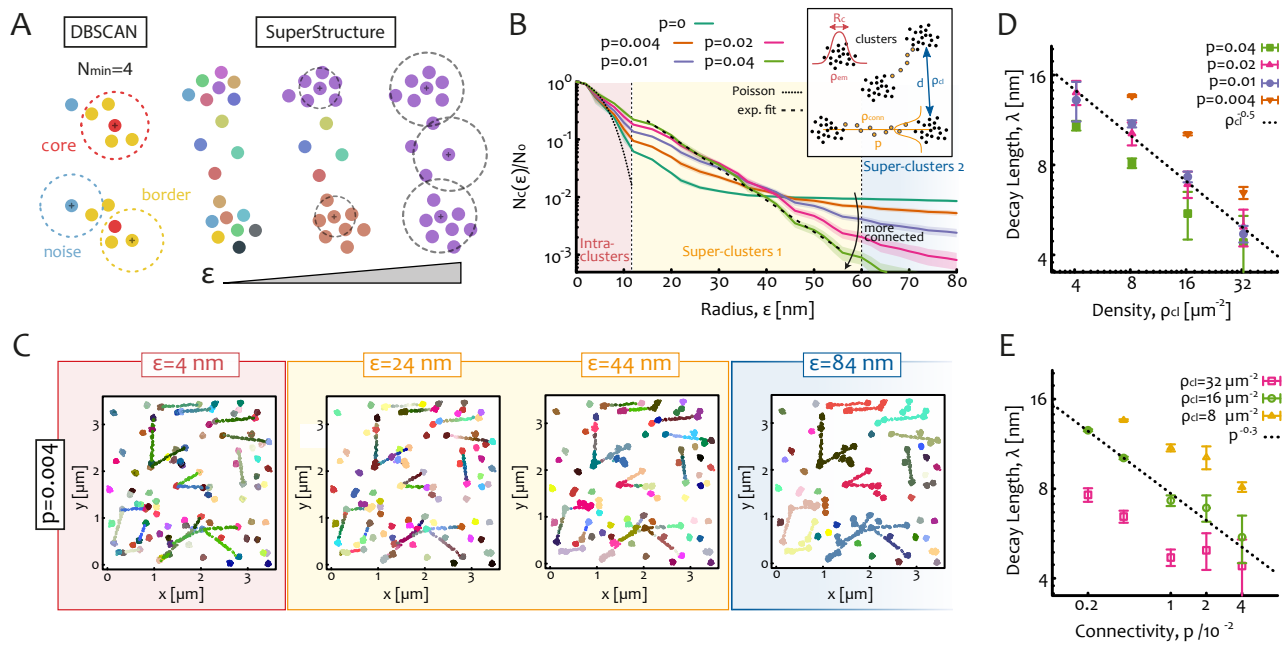
FIG. 1: **A.** Left side: sketch of DBSCAN clusters detection for a choice of ϵ and $N_{\min} = 4$; For every cluster DBSCAN identifies core (red) and border (yellow) points, as well as free noise (blue) points. Right side: sketch of SuperStructure algorithm for a pair of connected clusters. While $N_{min} = 0$ is kept fixed, the neighbour parameter ϵ is gradually increased. First points in clusters are merged, then clusters themselves are connected due to connection points. **B.** SuperStructure curves for random connected clusters for fixed average cluster radius R_c and inter-cluster distance d and different values of connectivity p. Three regimes can be distinguished: (i) intra-clusters (red), (ii) first super-clusters (yellow) and (iii) second super-clusters regime (blue). The decay in the intra-cluster regime corresponds to a Poisson Avoidance function with $\rho_{em} = 16000 \ \mu m^{-2}$ (dotted line). The first super-clusters regime can be fitted by single exponentials (dashed line) which return an effective decay length λ. **C.** Snapshots for p=0.004 for $\epsilon = 4, 24, 44, 84$ nm. **D.** λ versus ρ_{cl} has a $\rho_{cl}^{-0.5}$ dependence for different p. **E.** λ versus p has a $p^{-0.3}$a for different ρ.

aries between the three regimes of $N_c(\epsilon)$ (Fig. 1B): (i) for $\epsilon \lesssim 12 \ nm$ the intra-clusters regime follows a Poissonian decay with $\rho_{em} = N/(\pi R_c^2)$ (see Supplementary Note 2); (ii) for intermediate values of ϵ the exponential super-clusters regime dominates and the fusion of connected clusters takes place; (iii) for $\epsilon \gtrsim 60 \ nm$ we expect to observe the coalescence of super- and non-connected clusters in a second super-clusters regime; this should be captured again by a Poissonian decay for $p = 0$, and by an exponential for $p \neq 0$ (see Supplementary Note 3).

Fig. 1B shows the exponential behaviour of the super-clusters regime (ii) for different values of p. Importantly, a larger connection probability p results in an effectively smaller decay length – or larger spatial rate of merging – for the super-clusters regime (ii). This strongly suggests that the effective decay length (or rate) mirrors the connectedness of the underlying super-structures (Fig. 1C). This decay length (plotted in Fig. 1D) results from the combined contribution of clusters density ρ_{cl} and connectivity p. A larger density of clusters can impact the decay length as much as a larger connectivity, as shown by simulations at fixed p and different ρ_{cl} (Fig. 1E and Supplementary Note 3). In particular, we find that the functional form of the decay length $\lambda \sim \rho_{cl}^{-1/2}p^{-0.3}$ (Figs. 1D and E). In particular, the density contribution is $\sim \rho_{cl}^{-1/2}$ as it depends on the typical distance between clusters and is relevant when comparing datasets with different cluster

density (see below).

We then applied SuperStructure to dSTORM data acquired on three different nuclear proteins (Fig. 2A): the serine/arginine-rich splicing factor SC35, the heterogeneous nuclear RiboNuclear Protein hnRNP-C and hnRNP-U (also known as Scaffold Attachment Factor A, SAF-A). These proteins are abundantly expressed in the nucleus of human cells and are involved with RNA processing at different stages. SC35 is necessary for RNA splicing while hnRNPs are implicated in regulation and maturation of mRNA but also in chromatin structure [23, 32, 33]. In particular, SAF-A is thought to form a dynamic mesh that regulates large-scale chromatin organisation by keeping gene-rich loci in a decompacted state [23, 24]. Hence, understanding the organisation of this protein beyond the traditional single-cluster analysis is an important step towards understanding its function.

Average SuperStructure curves for these proteins are shown in Fig. 2C, where we highlighted the regimes discussed above. Both hnRNPs display a first super-clusters regime for which the curves decay as exponentials, suggesting that within this range distinct clusters tend to coalesce. Interestingly, SC35 displays exponentials with different characteristic rates in two distinct super-clusters regimes: one for intermediate ϵ, when clusters inside speckles merge (first super-clusters regime), and another one for high ϵ indicating that speckles merge together
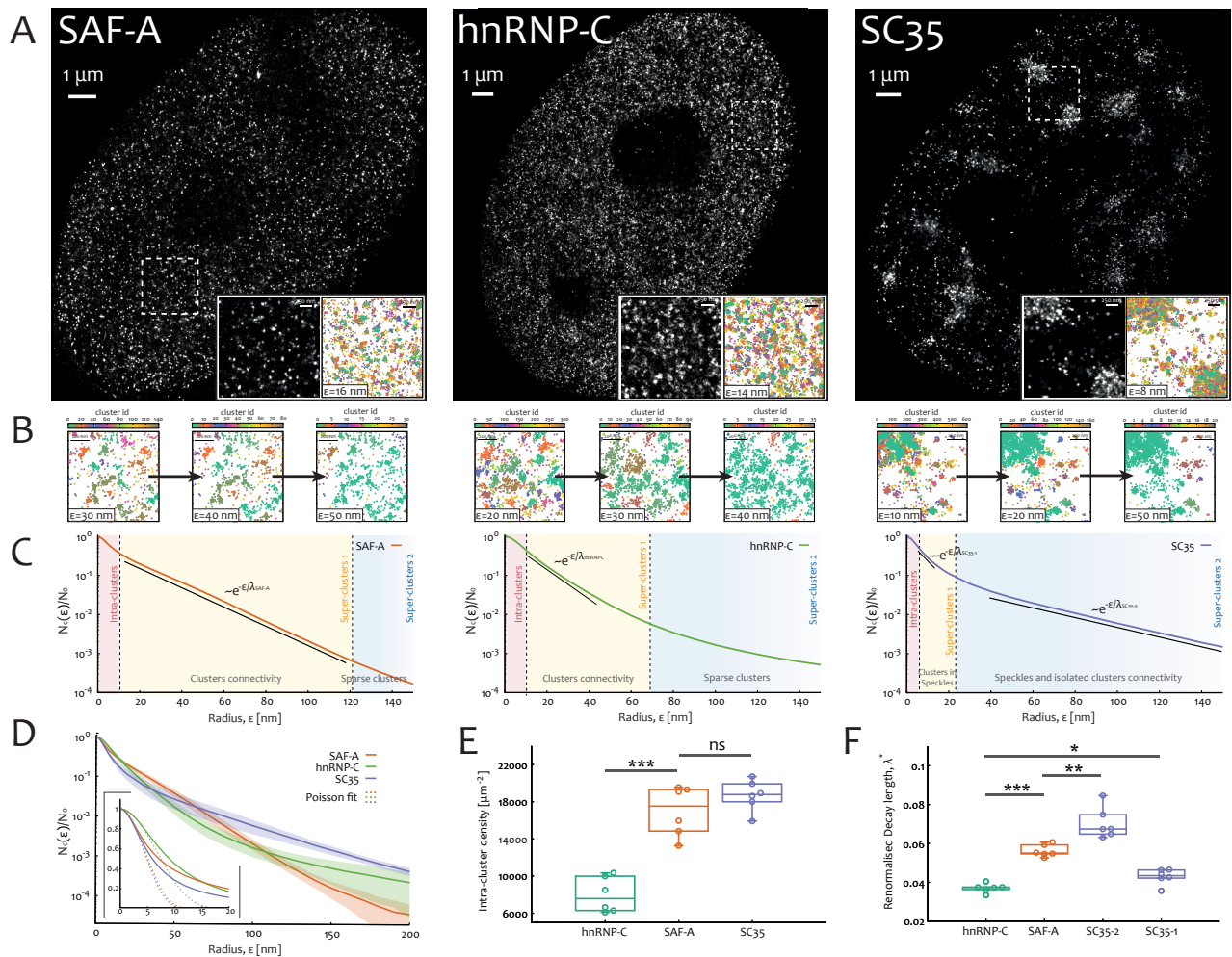
FIG. 2: Application of SuperStructure algorithm to hnRNPc, SAF-A and SC35 super-resolution data. **A.** dSTORM reconstructed images by using a pixel size of $10\ nm$. Inset of $4\ \mu m^2$ of super-resolution images and relative spatial positions of the data. Palettes represent the cluster id computed by running SuperStructure with $N_{min} = 0$ and a suitable value of ϵ for every protein ($\epsilon = 16\ \mu m$ for SAF-A, $\epsilon = 14\ \mu m$ for hnRNP-C, $\epsilon = 8\ \mu m$ for SC35). **B.** Identified clusters for increasing values of ϵ in the regimes where clusters merge. **C.** Normalised average SuperStructure curves: the number of clusters normalised with the number of total points in the system while keeping $N_{min} = 0$ and increasing ϵ in the range $[0 : 150]\ nm$. The average is computed over 6 independent dSTORM acquisitions. Vertical dashed lines highlight different merging regimes: intra-clusters, first super-clusters and second super-clusters regimes. For SAF-A and hnRNP-C the exponential regime of clusters merging (first super-clusters) is highlighted. In case of SC35, two regimes are highlighted: the merging of clusters in speckles (first super-clusters) and the merging of speckles and isolated clusters (second super-clusters). **D.** Normalised average SuperStructure curves in the range $[0 : 200]\ nm$. Shaded region represents standard deviation from the average. Poisson fits of the intra-cluster regime at small ϵ are shown in the inset. **E.** Intra-cluster density of emitters as calculate from Poisson fit at small ϵ. **F** Renormalised decay length (by cluster density) for the regimes highlighted in **C**. Note that significance is annotated as follows: $p\text{-}value \leq 0.001$ as $***$, $0.001 \leq p\text{-}value \leq 0.01$ as $**$, $0.01 \leq p\text{-}value \leq 0.05$ as $*$ and $p\text{-}value \geq 0.05$ as ns.

and with single clusters (second super-clusters regime). These connectivity properties were further confirmed by directly looking at the arrangement of identified clusters (inset of Fig. 2A and Fig. 2B).

We first fitted the intra-cluster regime with the Poisson function (see Supplementary Note 3 and 4). Interestingly, SAF-A and SC-35 form clusters with similar densities, while those of hnRNP-C are less dense (see Fig. 2D and E). Then, in order to have a quantitative description of the clusters/speckles connectivities, we calculated the decay length in the exponential regimes. However, a direct comparison is possible only by renormalising decay lengths by the cluster/speckles density (see also Supple-

mentary Note 4). Fig. 2F highlights that while hnRNP-C has a short renormalised decay length due to the highly connected clusters, SAF-A displays a slower decay (larger λ^*) due to a more weakly connected mesh. Finally, SC35 displays one (intra-speckles) very connected and fast regime of the order of that of hnRNPs (small λ^*) followed by a regime (inter-speckles) that is much slower compared to hnRNPs. In other words, our analysis revealed that the super-structures inside nuclear speckles are not more connected than those formed by hnRNPs but they are denser, as explained in Supplementary Note 4. Further, this method is also sensitive enough to distinguish the connectivity of two closely related wild-type

hnRNPs in live cells.

SuperStructure was also evaluated on published dSTORM datasets on ceramids [34] (Supplementary Note 5). In this case we were able to confirm the quantitative results of the authors, but moreover we found a notable absence of connections between clusters, further supporting that the connections detected in hnRNP-U/C and SC35 are significant and potentially functional.

In summary, we have developed a novel analysis method that yields the degree of connectivity between clusters in SMLM data. Because parameter-free and based on a well established and highly popular clustering software (DBSCAN), we believe that our method will find broad use as standardised tool in SMLM data analysis and to tackle challenging datasets. We stress that the proposed method is ideal to quantify and compare different datasets without a priori assumptions. Additionally, we predict this method could be also applied for technical purpose, such as evaluation of fluorophores blinking quality.

Methods

SuperStructure algorithm and pipeline.

The algorithm details and the complete pipeline are exhaustively discussed in Supplementary Note 1.

Experimental details for generating experimental dSTORM dataset for SAF-A, hnRNP-C and SC-35.

Cells Preparation for dSTORM imaging. RPE1 cells were grown overnight in a 8-well Nunc Lab-Tek II Chambered Coverglass (1.5 borosilicate glass) at 37 degrees at initial concentration of 10^5 $cells/ml$. We fixed the cells for 10 minutes in 4% PFA, followed by washing in PBS, permeabilisation with 0.2% Triton-X for 10 minutes, washed in PBS again and blocked with 1% BSA for 10 minutes.
Immuno-fluorescence labelling was done by exposing the cells for 2 hours to (i) hnRNP-U polyclonal rabbit antibody A300-690A from *Bethyl Laboratories* at 10 $\mu g/ml$ or (ii) hnRNP-C1/C2 (4F4) mouse monoclonal antibody sc-32308 from *Santa Cruz Biotechnology* at 0.2 $\mu g/ml$ or (iii) SC-35 mouse monoclonal antibody ab11826 from *abcam* at 2 $\mu g/ml$ and then washed. Then, cells were exposed for 1 hour to secondary antibody. The secondary antibody was made by anti-rabbit or anti-mouse F(ab) fragments fused to organic fluorophore CF647 at a stechiometric ratio of about 1.
Oxygen scavenger imaging buffer for dSTORM was prepared fresh on the day and the recipe employed was similar to that of Ref. [35]. We mixed (i) 5.3 ml of 200 mM Tris and 50 mM NaCl solution with (ii) 2 ml of 40% glucose solution, (iii) 200 μl of GLOX, (iv) 1.32 ml of 1M β-mercaptoethanol and (v) 100 μl of 50 $\mu g/ml$ DAPI solution. The GLOX solution was made by mixing 160 μl of 200 mM Tris and 50mM NaCl with 40 μl of bovine liver catalase and 18 mg of glucose oxidase. The 8.8 ml final solution was suitable to fill the chambers of the 8-well dish and a cover glass was sealed at the top of the dish to prevent inflow of oxygen.

dSTORM Acquisition. We performed 3D STORM acquisitions using a Nikon N-STORM system with Eclipse Ti-E inverted microscope with laser TIRFilluminator (Nikon UK Ltd, Kingston Upon Thames, UK). We equipped the microscope with a CFI SR HP Apo TIRF 100x objective lens and applied a 1.5X additional optical zoom. The Z position was acquired by using a cylindrical astigmatic lens. Laser light was provided via a Nikon LU-NV laser bed with 405, 488, 561, 640nm laser lines. In particular, fluorophores were stochastically excited using the 640 nm laser beam. Images were acquired with an Andor iXon 897 EMCCD camera (Andor technologies, Belfast UK). The Z position was stabilised during the entire acquisition by the integrated perfect focus system (PFS).
For every nucleus, we acquired a stack of 20000 (256 pixels x 256 pixels) frames with 19 ms exposure time. Acquired images had a pixel resolution of 106 nm/pixel. For every condition (SAF-A, hnRNP-C, SC35 antibody labelling) we acquired 6 nuclei, i.e. 6 independent datasets.

Raw images and post-processing analysis. The raw stack of frames was analysed with the FIJI Thunderstorm plugin [36]. Frames were initially filtered by using Wavelet functions to separate signal from noise. The B-Spline order was set to 3 and the B-Spline scale to 2.0 as suggested in Ref. [36]. In order to localise the emitters centroids, we thresholded filtered images (threshold values was set 1.2 times the standard deviation of the 1st Wavelet function) and calculated the local maximum relative to the 8 nearest neighbours. Finally, we fitted the emitters signal distribution with elliptical gaussians (ellipses are necessary for z-position localisation) using the weighted least square method and by setting 3 $pixels$ as initial fitting radius and 1.6 $pixels$ as initial sigma.
Localised data was then post-processed using the same plugin. (i) We corrected the XY drift using a pair correlation analysis, (ii) filtered data with a position uncertainty < 40, (iii) restricted the z-position to the interval $[-100 : 100]$ nm and projected the data in a 2-dimensional plane, as the z-axis precision is around 100 nm.
Reconstructed images shown in the main text were created by using the average shifted histograms method with a 10X magnification (10.6 nm/pixel).

Data Availability

The simulated and experimental datasets that support the findings of this study are available from the corresponding authors upon request.

Author Contribution

M.M., D.M. and N.G. conceived the project. M.M. and D.M. analysed both simulated and experimental datasets. M.M., S.v.d.L. and D.M. generated the simulated dataset. M.M., E.L. and D.M. performed super-resolution experiments and localisation analysis. M.M., D.M., S.v.d.L. and N.G. wrote the manuscript with input from all the authors.

Code Availability

The code for the generation of SuperStructure curves is available from *https://git.ecdf.ed.ac.uk/dmichiel/superstructure*.

Additional Information

Supplementary Notes with the algorithm pipeline and the SuperStructure analysis of the simulated/experimental datasets are available.
Correspondence and requests for materials should be addressed to D.M. and N.G.

Competing Interests

The authors declare there are not competing interests.

[1] Van De Linde, S. *et al.* Direct stochastic optical reconstruction microscopy with standard fluorescent probes. *Nature Protocols* **6**, 991–1009 (2011).

[2] Schermelleh, L., Heintzmann, R. & Leonhardt, H. A guide to super-resolution fluorescence microscopy. *Journal of Cell Biology* **190**, 165–175 (2010).

[3] Henriques, R., Griffiths, C., Rego, E. H. & Mhlanga, M. M. PALM and STORM: Unlocking live-cell super-resolution. *Biopolymers* **95**, 322–331 (2011).

[4] Cisse, I. I. *et al.* Real-Time Dynamics of RNA Polymerase II Clustering in Live Human Cells. *Science* **341**, 664–667 (2013).

[5] Kapanidis, A. N., Lepore, A. & El Karoui, M. Rediscovering Bacteria through Single-Molecule Imaging in Living Cells. *Biophysical Journal* **115**, 190–202 (2018).

[6] Chong, S. *et al.* Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* **361** (2018).

[7] Revyakin, A., Liu, C., Ebright, R. H. & Strick, T. R. Abortive Initiation and Productive Initiation by RNA Polymerase Involve DNA Scrunching Andrey. *Science* **314**, 1139–1144 (2006).

[8] Deniz, A. A., Mukhopadhyay, S. & Lemke, E. A. Single-molecule biophysics: At the interface of biology, physics and chemistry. *Journal of the Royal Society Interface* **5**, 15–45 (2008).

[9] Owen, D. M. *et al.* PALM imaging and cluster analysis of protein heterogeneity at the cell surface. *Journal of Biophotonics* **3**, 446–454 (2010).

[10] Sengupta, P. *et al.* Probing protein heterogeneity in the plasma membrane using PALM and pair correlation analysis. *Nature Methods* **8**, 969–975 (2011).

[11] Garcia-Parajo, M. F., Cambi, A., Torreno-Pina, J. A., Thompson, N. & Jacobson, K. Nanoclustering as a dominant feature of plasma membrane organization. *Journal of Cell Science* **127**, 4995–5005 (2014).

[12] Baumgart, F. *et al.* Varying label density allows artifact-free analysis of membrane-protein nanoclusters. *Nature Methods* **13**, 661–664 (2016).

[13] Spahn, C., Herrmannsdörfer, F., Kuner, T. & Heilemann, M. Temporal accumulation analysis provides simplified artifact-free analysis of membrane-protein nanoclusters. *Nature Methods* **13**, 963–964 (2016).

[14] Griffié, J. *et al.* A Bayesian cluster analysis method for single-molecule localization microscopy data. *Nature Protocols* **11**, 2499–2514 (2016).

[15] Brangwynne, C. P., Tompa, P. & Pappu, R. V. Polymer physics of intracellular phase transitions. *Nature Phys.* **11**, 899–904 (2015).

[16] Larson, A. G. *et al.* Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature* **547**, 236–240 (2017).

[17] Strom, A. R. *et al.* Phase separation drives heterochromatin domain formation. *Nature* **547**, 241–245 (2017).

[18] Sabari, B. R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361** (2018).

[19] Cho, W.-K. *et al.* Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361**, 412–415 (2018).

[20] Maharana, S. *et al.* RNA buffers the phase separation behavior of prion-like RNA binding proteins. *Science* **360**, 918–921 (2018).

[21] Bintu, B. *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362** (2018).

[22] Boettiger, A. N. *et al.* Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature* **529**, 418–422 (2016).

[23] Nozawa, R.-s. *et al.* SAF-A regulates interphase chromosome structure through oligomerisation with chromatin-associated RNAs. *Cell* **169**, 1214–1227.e18 (2017).

[24] Michieletto, D. & Gilbert, N. Role of nuclear RNA in regulating chromatin structure and transcription. *Current Opinion in Cell Biology* **58**, 120–125 (2019).

[25] Prakash, K. *et al.* Superresolution imaging reveals structurally distinct periodic patterns of chromatin along pachytene chromosomes. *Proceedings of the National Academy of Science of the United States of America* **112**,

14635–40 (2015).

[26] Szabo, Q. *et al.* TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Science Advances* **4**, 1–14 (2018).

[27] Nir, G. *et al.* Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genetics* **14**, 1–35 (2018).

[28] Bronshtein, I. *et al.* Loss of lamin A function increases chromatin dynamics in the nuclear interior. *Nature Communications* **6**, 1–9 (2015).

[29] Khanna, N., Zhang, Y., Lucas, J. S., Dudko, O. K. & Murre, C. Chromosome dynamics near the sol-gel phase transition dictate the timing of remote genomic interactions. *Nature Communications* **10**, 1–13 (2019).

[30] Leidescher, S. *et al.* Spatial Organization of Transcribed Eukaryotic Genes. *bioRxiv* 2020.05.20.106591 (2020).

[31] Ester, M., Kriegel, H.-P., 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*, KDD'96, 226–231 (AAAI Press, 1996).

[32] Xiao, R. *et al.* Nuclear Matrix Factor hnRNP U/SAF-A Exerts a Global Control of Alternative Splicing by Regulating U2 snRNP Maturation Rui. *Mol. Cell* **45**, 656–668 (2012).

[33] Caudron-Herger, M. *et al.* Coding RNAs with a non-coding function: Maintenance of open chromatin structure. *Nucleus* **2** (2011).

[34] Burgert, A. *et al.* Characterization of Plasma Membrane Ceramides by Super-Resolution Microscopy. *Angewandte Chemie* **129**, 6227–6231 (2017).

[35] McSwiggen, D. T. *et al.* Evidence for dna-mediated nuclear compartmentalization distinct from phase separation. *eLife* **8**, e47098 (2019).

[36] 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).