Robust nonparametric descriptors for clustering quantification in single-molecule localization microscopy

Shenghang Jiang[1], **Sai Divya Challapalli**[2], **and Yong Wang**[1,2,3,*]

[1]Department of Physics, University of Arkansas, Fayetteville, Arkansas, 72701, United States
[2]Microelectronics and Photonics Graduate Program, University of Arkansas, Fayetteville, Arkansas, 72701, United States
[3]Cell and Molecular Biology Program, University of Arkansas, Fayetteville, Arkansas, 72701, United States
[*]To whom correspondence should be addressed: yongwang@uark.edu (Y.W.)

ABSTRACT

We report a robust nonparametric descriptor, $J'(r)$, for quantifying the spatial organization of molecules in single-molecule localization microscopy. $J'(r)$, based on nearest neighbor distribution functions, does not require any parameter as an input for analyzing point patterns. We show that $J'(r)$ displays a valley shape in the presence of clusters of molecules, and the characteristics of the valley reliably report the clustering features in the data. More importantly, the position of the $J'(r)$ valley ($r_{J'_m}$) depends exclusively on the density of clustering molecules (ρ_c). Therefore, it is ideal for direct measurements of clustering density of molecules in single-molecule localization microscopy. We demonstrate that the $r_{J'_m}$–ρ_c relation is robust against noises at high levels. We exploited this relation to characterize the clustering of H-NS proteins in *E. coli* bacteria, and measured that density of H-NS clusters increases by 27% upon cold shock.

Introduction

Single-molecule localization microscopy (SMLM) has been utilized broadly in imaging biological molecules – proteins, DNA, and RNA – in various biological systems.[1–5] More importantly, by localizing individual molecules, SMLM has allowed quantitative analyses on the spatial organizations and patterns of these molecules, and produced new, quantitative and crucial information that was not accessible previously. New mechanisms of various cellular and molecular organizations and activities at the single-cell level have been unraveled using SMLM.[6–15]

Many algorithms have been adopted, utilized, or developed, in the field of SMLM for analyzing localization data of molecules and quantifying inter-molecular organizations.[13,14,16–23] These methods provide means to distinguish single molecules from clusters of molecules, to examine complex patterns of molecular organization, and to quantify features of spatial organizations. For example, pair-correlation analysis has been applied to SMLM data on membrane proteins to detect the presence of clusters, as well as various cluster features, such as the density of molecules in a cluster and overall size of a cluster.[16] In addition, density-based algorithms such as DBSCAN (density-based spatial clustering of applications with noise)[24,25] and OPTICS (ordering points to identify the clustering structure)[26,27] have been exploited to identify clusters of proteins and nucleic acids, as well as to probe the clustering structures, in both bacteria and animal cells.[13,14,17–19] Another method that has been used for analyzing SMLM data is Ripley's K function and its derivatives.[20,21] More recently, Bayesian analysis and Voronoï diagrams have been utilized to identify and analyze the clustering of biological molecules.[22,23]

However, a general problem in some of these methods or algorithms is the need of input parameters. For example, DBSCAN and OPTICS need two parameters (a radius, eps, and the minimum number of points in the neighborhood for a point to be considered as a core point, minPts),[24–27] and they are known to be sensitive to the chosen parameters.[18,28] The identification of clusters in the Voronoï diagram based method also requires a density threshold to determine whether points form clusters.[23] Although various techniques have been proposed to

determine "appropriate" parameters for use,[23,25,27,29] bias is inevitably introduced by the choice of parameters in these algorithms. As a result, a robust nonparametric method is desired for the quantification of clustering features of molecules.

Here we present a descriptor based on nearest neighbor distribution functions for quantifying the spatial pattern of molecules in SMLM. We examined nearest neighbor function,[30] $G(r)$, spherical contact distribution function,[30] $F(r)$, and the J-function,[31,32] $J(r) = (1-G(r))/(1-F(r))$, and found that the associated derivative functions, $G'(r)$ and $J'(r)$, reliably report the clustering features of points. In the presence of clusters, $G'(r)$ and $J'(r)$ are peak/valley shaped. More importantly, we observed that $r_{J'_m}$, the position of the $J'(r)$ valley, depends exclusively on the density of clustering points (ρ_c). Therefore, it is ideal for direct measurements of the clustering density of molecules in SMLM. In addition, it turns out that the $r_{J'_m}$–ρ_c relation is robust in the presence of up to ~ 10 times more noise points. We applied the $r_{J'_m}$–ρ_c relation to quantify the clustering of H-NS proteins in *E. coli*, and found that the density of H-NS clusters increases by 27% upon cold shock (from 37°C to room temperature). We expect that this nonparametric descriptor $J'(r)$ is useful in a broad range of applications in SMLM.

Results

$G(r)$, $F(r)$ and $J(r)$, and their derivatives

When quantifying the spatial organization of biological molecules in SMLM data, of particular interest in certain situations is the clustering or aggregation of molecules,[33–36] which is featured by an enhancement in the local density of molecules. This enhancement in density has been used to identify clusters methods such as DBSCAN, OPTICS, and Voronoï tessellation.[13,14,17–23] On the other hand, the enhancement in the molecular density is also accompanied by the decrease of intermolecular distances, which could be described by functions based on nearest neighbor distances, such as pair-wise correlation function,[16] nearest neighbor function $G(r)$, and spherical contact distribution function $F(r)$.[30] The nearest neighbor function $G(r)$ is the distribution function of the distance r of a point (existing in the data) to the nearest other point, while the spherical contact distribution $F(r)$ is the distribution function of the distance r of an arbitrary point in the space (not necessarily existing in the data) to the nearest point in the data.[30] In addition, another function, $J(r)$, has been suggested by van Lieshout and Baddeley in 1996,[31] $J(r) = \frac{1-G(r)}{1-F(r)}$, as a better nonparametric test to determine whether data were from a Poisson process.

We first explored how $G(r)$, $F(r)$ and $J(r)$ functions depends on the clustering features of points using numerical simulations. Briefly, we generated points forming various clusters in the presence of noises (i.e., Poisson random points) in a region of interest, and computed these three functions. In a two-dimensional Poisson random process where points were not forming clusters (Fig. 1A), the nearest neighbor functions gave the expected curves, $G_p(r) = F_p(r) = 1-\exp(-\lambda\pi r^2)$ and $J_p(r) = 1$ (Fig. 1C). However, when points aggregated into clusters (Fig. 1B), both $G(r)$ and $J(r)$ deviated significantly from random points, while $F(r)$ became only slightly different (Fig. 1D). We observed that $J(r)$ drops quickly from 1 to ~ 0.4 when r increases from 0 to 5 nm, while $G(r)$ raises quickly in the same r-range (0–5 nm). This observation indicates that $G(r)$ and $J(r)$ could be used for detection of clusters.

Furthermore, to remove accumulative effects, and inspired by Kiskowski et.al.,[37] we calculated the derivatives of these functions ($G'(r)$, $F'(r)$ and $J'(r)$). Striking peaks or valleys appeared in $G'(r)$ and $J'(r)$ if points formed clusters (Fig. 1F). In contrast, these derivative functions remain essentially flat for random points (Fig. 1E). On the other hand, $F'(r)$'s were very similar in the two cases (Fig. 1E–F).

Dependence of $G'(r)$ and $J'(r)$ functions on clustering features.

To explore quantitative applications of $G'(r)$ and $J'(r)$, we examined how they differ when the clustering features of points vary. Here we focus on the following features: the radius of clusters, R_c, the density of points in clusters, ρ_c, the number of clusters, n_c, the density of (random) noise points, ρ_r, and the width (W) and height (H) of the region of interest. The first three features, R_c, ρ_c and n_c, are directly related to the properties of clusters in the data, while ρ_r is an indicator of the noise level. By varying one feature at a time, we observed that changes in ρ_c, ρ_r, or R_c resulted in horizontal shifting or vertical scaling of both $G'(r)$ and $J'(r)$ (Fig. 2A–C). For example, both $G'(r)$ and $J'(r)$ shift to the left and scale up when the clusters become denser (ρ_c increases). If the clusters become bigger

(R_c increases) while keeping the clustering density constant, little horizontal translation was observed (Fig. 2C), although both $G'(r)$ and $J'(r)$ scale up too. In contrast, $G'(r)$ and $J'(r)$ are insensitive to the number of clusters (N_c) or the size of the region of interest, W and H (Fig. 2D–F).

We further quantified the dependence of $G'(r)$ and $J'(r)$ on the clustering features. By fitting $G'(r)$ and $J'(r)$ with polynomials, both the amplitude (height G'_m for $G'(r)$, or depth J'_m for $J'(r)$) and the positions ($r_{G'_m}$ and $r_{J'_m}$) of the peaks and valleys were determined. The dependence of these values on the clustering features are shown in Fig. 3, S1, S2 and S3. We observed that both G'_m and J'_m depend on all the clustering features (Fig. S1 and S3), but $r_{G'_m}$ and $r_{J'_m}$ are most sensitive to the density of clustering points ρ_c (Fig. 3 and S2). Most interestingly, $r_{J'_m}$ is essentially independent on all the clustering features except the density of clustering points ρ_c (Fig. 3), providing a way to correlate $r_{J'_m}$ with directly measuring the clustering densities of molecules, as shown below.

Direct measurement of the density of clustering points by $r_{J'_m}$

To explore the possibility of using $r_{J'_m}$ for direct measurements of clustering densities of molecules, we first confirm that the $r_{J'_m}$–ρ_c relation is independent on other clustering features even if we simultaneously vary both ρ_c and R_c, or N_c, or $\rho_r \cdots$. We found that the the $r_{J'_m}$–ρ_c relation from all the simulations collapse onto a single curve (Fig. 4A). Fitting all the data with a power-law function $r_{J'_m} = A \cdot \rho_c^{-\alpha} + b$ gives $\alpha = 0.76 \pm 0.03$, yielding a "calibration" curve for measuring clustering densities in any point patterns. We note that the $r_{J'_m}$–ρ_c relation is monotonic over a larger range of ρ_c (Fig. 3A); however, more complicated mathematical forms other than a single power-law are needed to fit the "calibration" curve.

To verify the capability of using the $r_{J'_m}$–ρ_c "calibration" curve to directly measure the clustering density of molecules in SMLM data, we exploited numerical simulations again, for which we know the ground-truth clustering density. Briefly, we generated 50 simulated data with a "unknown" clustering density ($\rho_c^t = 0.0103$ nm^{-2}), and computed $J'(r)$ and $r_{J'_m}$ for each simulation. The "measured" clustering density ρ_c^m is then obtained from the $r_{J'_m}$–ρ_c "calibration" curve, $\rho_c = \left((r_{J'_m} - b)/A\right)^{-1/\alpha}$. We found that the "measured" clustering densities ρ_c^m are very close to the the ground-truth density ρ_c^t: the relative error, $\delta_{\rho_c} = |\rho_c^m - \rho_c^t|/\rho_c^t \times 100\%$, is below 4.6% for all the 50 simulations, and the average relative error is 2.5%.

Robustness of the $r_{J'_m}$–ρ_c relation

"Noises" are almost always present in SMLM data, due to individual molecules not forming clusters, non-specific labeling, and/or false-positive localizations. Amazingly, we found that $r_{J'_m}$ is independent on the density of random points in the data, as shown in Fig. 3B and 4, strongly suggesting that the use of $r_{J'_m}$–ρ_c relation for measuring clustering densities is robust.

To further assess the robustness of the $r_{J'_m}$–ρ_c relation, we systematically investigated, given a density of clustering point (ρ_c), how $r_{J'_m}$ deviates in the presence of various amount of noises. The noise level is defined as $\beta = n_{rp}/n_{cp}$, where n_{rp} is the number of (random) noise points, and n_{cp} is the number of clustering points, in the whole region of interest. We found that $r_{J'_m}$ remains constant (with relative errors $\delta_{r_{J'_m}}$ below 5%) even if there are $\approx$ 10 times more noise points than clustering points (Fig. 5), demonstrating that the $r_{J'_m}$–ρ_c relation is extremely robust.

$J'(r)$ as a global descriptor for heterogeneous clusters.

It's known that, in certain applications, molecules of interest might form heterogeneous clusters.[16,23] We examined heterogeneity arising from either clustering radius (R_c) or clustering density (ρ_c). Briefly, simulations were run for clusters with two different clustering radii (R_{c1} and R_{c2}), or two different clustering densities (ρ_{c1} and ρ_{c2}), in the presence of random noises. We noticed that $J'_{(\rho_{c1},\rho_{c2})}(r)$ from heterogeneous clusters with different clustering densities shifted both horizontally and vertically, and fell between the two curves from homogeneous clusters, $J'_{\rho_{c1}}(r)$ and $J'_{\rho_{c2}}(r)$ (Fig. 6). In addition, we observed that $J'_{(\rho_{c1},\rho_{c2})}(r)$ overlapped very well with $J'_{\bar{\rho}_c}(r)$ from a homogeneous sample with a clustering density equal to the algebraic mean, $\bar{\rho}_c = (\rho_{c1} + \rho_{c2})/2$ (Fig. 6). It is noted that $G'(r)$ shows a similar behavior. In contrast, for heterogeneous clusters with different radii, $J'_{(R_{c1},R_{c2})}(r)$ shifted only in the vertical direction. $r_{J'_m}$ does not change for heterogeneous clusters with different radii (Fig. S4), which is expected because the $r_{J'_m}$–ρ_c relation does not depend on R_c anyway. In addition, we found that $J'_{(R_{c1},R_{c2})}(r)$ is equivalent to

$J'_{\bar{R}_c}(r)$ from homogeneous clusters with a radius of $\bar{R}_c = \sqrt{\left(R_{c1}^2 + R_{c2}^2\right)/2}$ (Fig. S4). In summary, we conclude that $G'(r)$ and $J'(r)$ are descriptors reporting the "global" clustering density throughout the region of interest.

Application of the $r_{J'_m}$–ρ_c relation to bacterial H-NS response to cold shock

We applied this method of measuring the clustering density using the $r_{J'_m}$–ρ_c relation to probe bacterial response to cold shocks in terms of clustering of H-NS proteins. The H-NS protein is a neutral, heat-stable, dimeric protein that displays high affinity for curved DNA.[38–40] It has been shown that its expression in *E. coli* enhances when shifting from 37°C to a lower temperature.[41] We have been using SMLM to investigate the spatial organization of H-NS proteins in K12 *E. coli*, as well as their responses to cold shock.[1] The data from SMLM (i.e., localizations) of two example cells before and after the cold shock, from 37°C (WT) to room temperature $\sim$ 22°C (RT), are shown in the Fig. 7A and B, respectively. It's clear that the expression of H-NS proteins became higher after cold shock. Simply counting the number of localizations showed that the overall expression level was enhanced by 40% (1137 ± 413 *vs.* 1596 ± 517). A two-sample *t*-test showed that this enhancement in the overal expression level is significant (*p*-value $= 0.0017$). Our observation is qualitatively consistent with previous ensemble observations by La Teana et. al.[41]

Further analysis shows that the enhancement in the expression level of H-NS protein upon cold shock results in denser clusters of the proteins, as shown by the super-resolved images from SMLM (Fig. 7C and D). The super-resolved images were produced by rendering each localization as a 2D Gaussian peak with a standard deviation of 10 nm. A higher density of H-NS proteins results in overlapping of the Gaussian peaks and therefore areas of higher intensities. We observed that H-NS proteins formed small clusters. Upon cold shock from 37°C to room temperature (22°C), the clusters displayed higher intensities, indicating that the H-NS proteins became denser.

To quantify the increase in the clustering density of H-NS proteins upon cold shock, we applied our method based on the $r_{J'_m}$–ρ_c relation (Fig. 7E–H). We found that, compared to the WT cells (before cold shock), $J'(r)$ of the RT cells (after cold shock) shifted to the left, with $\Delta r_{J'_m} = -0.75$ nm, and scaled deeper, both indicating that the clustering of H-NS proteins became denser. Application of the "calibration curve" in Fig. 4A) showed that the density of the H-NS clusters increased from 4.16×10^{-3} nm^{-2} to 5.27×10^{-3} nm^{-2}, or by $\sim$ 27%, when the temperature dropped from 37°C to 22°C.

Discussion

To conclude, we explored the possibility of utilizing nearest neighbor functions to quantify spatial patterns of molecules in single-molecule localization microscopy. We observed that the associated derivative functions, $G'(r)$ and $J'(r)$, could reliably report the clustering features of point patterns. We found that $J'(r)$ is particularly useful because its position, $r_{J'_m}$, relies exclusively on the density of clustering points (ρ_c). More importantly, we showed that this $r_{J'_m}$–ρ_c relation is very robust in the presence of up to $\sim$ 10 times more noise points than clustering points. We successfully applied the $r_{J'_m}$–ρ_c relation to quantify the reorganization of H-NS proteins in *E. coli*, and measured that the density of H-NS clusters increases by 27% upon cold shock (from 37°C to room temperature).

In the current study, we chose not to exploit any border correction when computing the nearest neighbor functions. A simplest approach for border correction is the "reduced sample" method,[42] which focuses on the points lying more than *r* away from the boundary of the region of interest. However, the "reduced sample" method discards much of the data, and therefore unacceptably wasteful. In addition, it's particularly inappropriate in certain applications where points are preferentially located at the boundary, an example of which is the spatial organization of high-copy number plasmids in bacteria.[14] We note that more sophisticated methods for border correction are available, including the Kaplan-Meier correction[43] and the Hanisch correction,[44] both are provided in the *spatstat* R-package.[45,46] These edge corrections can readily used in our method. However, for the sake of simplicity, uncorrected estimators for the nearest neighbor functions have been used in the current study.

We would like to emphasize that the current method based on nearest neighbor functions is nonparametric and robust. Computing the nearest neighbor functions and their derivatives does not require any parameters as human inputs, eliminating possible subjective biases that might exist in other algorithms such as DBSCAN and OPTICS. In addition, the performance of this method is robust, even in the presence of 10 times more noise points than clustering

points. The nonparametric nature and robustness of the current method would allow broad applications in the field of single-molecule localization microscopy.

We expect several types of applications of our method in the field of SMLM. First, it can be used as a direct quantification of the clustering density (ρ_c) of molecules in biological samples. Second, although our method does not identify clusters by itself, it provides objective means to determine parameters (i.e., clustering density) that could be used in other clustering-identification algorithm such as DBSCAN and Voronoï tessellation. In addition, in the current work, we focused on the $r_{J'_m} - \rho_c$ relation for non-parametric measurement of the clustering density of molecules; however, it's possible to design a way to figure out other cluster features (such as R_c and ρ_r) by taking advantage of the dependence of G'_m and J'_m on those features (Fig. S1 and S3), together with the information of ρ_c.

Finally, we would like to mention that, although the current study shows that the nearest neighbor functions are global descriptors, which cannot dinstinguish heterogeneous clusters, more sophisticated algorithms based on the nearest neighbor functions might be developed in order to detect heterogeneities in the data. For example, it's possible to divide the region of interest into sub-regions, perform computations on $J'(r)$, and report heterogeneous features on different sub-regions.

Methods

Spherical contact distribution function $F(r)$, nearest-neighbor distribution function $G(r)$, and the J function $J(r)$

In a set of points, X, in the k-dimensional space, the spherical contact distribution function, or sometimes referred to as the empty space function, $F(r)$, of X is defined as $F(r) = P\{d(y,X) \leq r\}$, where $d(y,X) = \min\{|y-x| : x \in X\}$ is the distance from an arbitrary point, y, to the nearest point of the point process, X.[30] For a Poisson process in the k-d space, $F_p(r) = 1 - \exp\left(-\lambda_k \pi r^k\right)$.[30] The nearest-neighbor distribution function $G(r)$ is very similar to $F(r)$: $G(r) = P^y\{d(y,X) \leq r\}$ where P^y is the Palm distribution, which is the conditional distribution of the entire process given that y is one point in X.[30] Therefore, $G(r)$ is the distribution function of the distance from a point of the process to the nearest other point of the process, i.e., the "nearest-neighbor". For a Poisson process in the k-d space, $G_p(r) = 1 - \exp\left(-\lambda \pi r^k\right) = F_p(r)$.[30] In 1996, van Lieshout and Baddeley suggested using the quotient $J(r) = \frac{1-G(r)}{1-F(r)}$ to characterize a point process.[31] For a Poisson process, $J_p(r) = 1$.

Simulation and computation of $G(r)$, $F(r)$, $J(r)$ and their derivatives.

Sets of points were generated in R programing language.[47] In a region of interest with a width (W) and a height (H), n_c circular clusters with radii of R_c were randomly distributed. Each cluster contains random points at a density of ρ_c. Poisson noise points were added randomly to the whole region of interest, with a density ρ_r. The total number of clustering points ($n_{cp} = n_c \cdot \rho_c \cdot \pi R_c^2$) and the total number of noise points ($n_{rp} = \rho_r \cdot WH$) define the noise level $\beta = n_{rp}/n_{cp}$.

Simulations were run using various sets of cluster features (W, H, ρ_r, ρ_c, n_c, R_c). For each set of features, 50–200 trials were run. The $G(r)$, $F(r)$, $J(r)$ functions and their derivatives were computed using the *spatstat* package,[45,46] without applying any edge corrections.

Bacterial strain, growth, fixation and harvest

An E. coli bacterial strain with *hns* gene deleted,[48] K12ΔHNS, was transformed with a plasmid, pHNS-mEos3.2C1, which was constructed by fusing the mEos3.2 protein (from plasmid mEos3.2-C1, Addgene #54550) to the *hns* gene. The bacterial strain (K12ΔHNS/pHNS-mEos3.2C1) was grown at 37°C overnight in defined M9 minimal medium, supplemented with 1% glucose, 0.1% casamino acids, 0.01% thiamine and appropriate antibiotics (kanamycin + chloramphenicol). On the second day, the overnight culture was diluted by 50 to 100 times into fresh medium so that the OD600 was 0.05. The fresh cultures were again grown at 37°C. When the OD600 reaches 0.2–0.4, the bacteria were either fixed directly (wild type) or incubated at room temperature (20°C) for 30 minutes and then fixed. Fixation was done by adding 37% formaldehyde directly to the cell cultures such that the final concentration is 3.7%, followed by incubation at room temperature for 30 minutes, with orbital shaking. The bacterial cells were harvested

by centrifugation at 1000*g* for 15 minutes at room temperature, followed by removal of the supernatant. The sample was then washed for three times by resuspension in PBS, centrifugation, supernatant removal, and stored in PBS at 4°C on a nutator.

Sample mounting and imaging

The sample prepared above was diluted in PBS by 5-fold before mounting. 10 uL of the diluted sample was loaded on a 5mm x 5mm 3% agarose pad, left at room temperature for 30 minutes, allowing the bacterial cells absorbed into the agarose pad. The agarose pad was flipped and attached to a clean coverslip. A chamber was then constructed by sandwiching a hollow square PDMS pad between the coverslip and a microscope slide. The chamber was sealed using epoxy glue and incubated at room temperature for 1-2 hours in dark before imaging, to prevent water evaporation and shrinkage of the agarose pad during data acquisition (Fig. S5).

The microscope was home-built on an Olympus IX-81 inverted microscope with an Olympus TIRF 60X N.A.=1.49 oil immersion objective. The microscope and data acquisition were controlled by Micro-Manager.[49] A 405 nm laser (Coherent, CA) and a 561 nm laser (CrystaLaser, NV) were used to "activate" and excite mEos3.2-HNS fusion proteins in bacteria. Emissions from the fluorescent proteins were collected by the objective and imaged on an EMCCD camera (Andor, MA). Before the camera, an additional telescope was inserted such that the total magnification reached 240X, and the effective pixel size of acquired images was 67 nm. The resulting movies (20 fps, 20,000 frames) were analyzed with RapidStorm.[50] The localizations that appear in adjacent frames and within 10 nm to each other were regrouped as a single molecule. The spots that were too dim, too wide, or too narrow, were rejected. The spots that survived the criteria were further corrected for drift using a mean cross-correlation algorithm,[51] and then used for generating reconstructed super-resolved images and for quantitative analysis using $G(r)$, $F(r)$, and $J(r)$, as well as their derivatives.

References

1. Betzig, E. *et al.* Imaging intracellular fluorescent proteins at nanometer resolution. *Science (New York, N.Y.)* **313**, 1642–1645 (2006).
2. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature methods* **3**, 793–795 (2006).
3. Huang, B., Wang, W., Bates, M. & Zhuang, X. Three-Dimensional Super-Resolution Reconstruction Microscopy. *Health San Francisco* **319**, 810–813 (2008).
4. Heilemann, M. *et al.* Subdiffraction-resolution fluorescence imaging with conventional fluorescent probes. *Angewandte Chemie - International Edition* **47**, 6172–6176 (2008).
5. Klein, T., Proppert, S. & Sauer, M. Eight years of single-molecule localization microscopy. *Histochemistry and Cell Biology* **141**, 561–575 (2014).
6. Kopek, B. G., Shtengel, G., Xu, C. S., Clayton, D. A. & Hess, H. F. Correlative 3D superresolution fluorescence and electron microscopy reveal the relationship of mitochondrial nucleoids to membranes. *Proceedings of the National Academy of Sciences* **109**, 6136–6141 (2012).
7. Doksani, Y., Wu, J. Y., de Lange, T. & Zhuang, X. Super-Resolution Fluorescence Imaging of Telomeres Reveals TRF2-Dependent T-loop Formation. *Cell* **155**, 345–356 (2013).
8. Xu, K., Zhong, G. & Zhuang, X. Actin, spectrin, and associated proteins form a periodic cytoskeletal structure in axons. *Science (New York, N.Y.)* **339**, 452–6 (2013).
9. Ricci, M. A., Manzo, C., García-Parajo, M. F., Lakadamyali, M. & Cosma, M. P. Chromatin Fibers Are Formed by Heterogeneous Groups of Nucleosomes In Vivo. *Cell* **160**, 1145–1158 (2015).
10. Stracy, M. *et al.* Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid. *Proceedings of the National Academy of Sciences* **112**, E4390–E4399 (2015).

11. Haas, B. L., Matson, J. S., DiRita, V. J. & Biteen, J. S. Single-molecule tracking in live V ibrio cholerae reveals that ToxR recruits the membrane-bound virulence regulator TcpP to the toxT promoter. *Molecular Microbiology* **96**, 4–13 (2015).

12. Buss, J. *et al.* A Multi-layered Protein Network Stabilizes the Escherichia coli FtsZ-ring and Modulates Constriction Dynamics. *PLOS Genetics* **11**, e1005128 (2015).

13. Fei, J. *et al.* Determination of in vivo target search kinetics of regulatory noncoding RNA. *Science* **347**, 1371–1374 (2015).

14. Wang, Y., Penkul, P. & Milstein, J. N. Quantitative Localization Microscopy Combined with DNA smFISH Reveals New Features of the Organization of High-Copy Number Plasmids in Bacteria. *Biophysical Journal* **In Press** (2016).

15. Boettiger, A. N. *et al.* Super-resolution imaging reveals distinct chromatin folding for different epigenetic states - nature16496.pdf. *Nature* **529**, 418–422 (2016).

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

17. Endesfelder, U. *et al.* Multiscale spatial organization of RNA polymerase in escherichia coli. *Biophysical Journal* **105**, 172–181 (2013).

18. Nan, X. *et al.* Single-molecule superresolution imaging allows quantitative analysis of RAF multimer formation and signaling. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 18519–24 (2013).

19. Caetano, F. A. *et al.* MIiSR: Molecular Interactions in Super-Resolution Imaging Enables the Analysis of Protein Interactions, Dynamics and Formation of Multi-protein Structures. *PLoS Computational Biology* **11**, 1–30 (2015).

20. Scarselli, M., Annibale, P. & Radenovic, A. Cell type-specific ??2-adrenergic receptor clusters identified using photoactivated localization microscopy are not lipid raft related, but depend on actin cytoskeleton integrity. *Journal of Biological Chemistry* **287**, 16768–16780 (2012).

21. Muranyi, W., Malkusch, S., Müller, B., Heilemann, M. & Kräusslich, H. G. Super-Resolution Microscopy Reveals Specific Recruitment of HIV-1 Envelope Proteins to Viral Assembly Sites Dependent on the Envelope C-Terminal Tail. *PLoS Pathogens* **9** (2013).

22. Rubin-Delanchy, P. *et al.* Bayesian cluster identification in single-molecule localization microscopy data. *Nature Methods* **12** (2015).

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

24. Ester, M., Kriegel, H. P., Sander, J. & Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Second International Conference on Knowledge Discovery and Data Mining* 226–231 (1996).

25. Daszykowski, M., Walczak, B. & Massart, D. Looking for natural patterns in data. Part 1. Density-based approach. *Chemometrics and Intelligent Laboratory Systems* **56**, 83–92 (2001).

26. Ankerst, M., Breunig, M. M., Kriegel, H.-p. & Sander, J. OPTICS : Ordering Points To Identify the Clustering Structure. *SIGMOD '99 Proceedings of the 1999 ACM SIGMOD international conference on Management of data* **28**, 49–60 (1999).

27. Daszykowski, M., Walczak, B. & Massart, D. L. Looking for natural patterns in analytical data. 2. Tracing local density with OPTICS. *Journal of Chemical Information and Computer Sciences* **42**, 500–507 (2002).

28. Deschout, H., Shivanandan, A., Annibale, P., Scarselli, M. & Radenovic, A. Progress in quantitative single-molecule localization microscopy. *Histochemistry and Cell Biology* **142**, 5–17 (2014).

29. Andronov, L., Orlov, I., Lutz, Y., Vonesch, J.-L. & Klaholz, B. P. ClusterViSu, a method for clustering of protein complexes by Voronoi tessellation in super-resolution microscopy. *Scientific reports* **6**, 24084 (2016).

30. Diggle, P. J. *Statistical Analysis of Spatial Point Patterns* (Hodder Education Publishers, London; New York, 2003), 2 edition edn.

31. van Lieshout, M. N. M. & Baddeley, A. A non-parametric measure of spatial interaction in point patterns (1996).

32. Kerscher, M. *et al.* A Global Descriptor of Spatial Pattern Interaction in the Galaxy Distribution. *The Astrophysical Journal* **513**, 543–548 (1999).

33. Sherman, E. *et al.* Functional nanoscale organization of signaling molecules downstream of the T cell antigen receptor. *Immunity* **35**, 705–720 (2011).

34. 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. *Nature Immunology* **14**, 82–89 (2012).

35. 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).

36. Ehmann, N. *et al.* Quantitative super-resolution imaging of Bruchpilot distinguishes active zone states. *Nature communications* **5**, 4650 (2014).

37. Kiskowski, M. A., Hancock, J. F. & Kenworthy, A. K. On the Use of Ripley's K-Function and Its Derivatives to Analyze Domain Size. *Biophysical Journal* **97**, 1095–1103 (2009).

38. Higgins, C. F. *et al.* Protein H1: a role for chromatin structure in the regulation of bacterial gene expression and virulence? *Molecular Microbiology* **4**, 2007–2012 (1990).

39. Ussery, D. *et al.* The chromatin-associated protein H-NS. *Biochimie* **76**, 968–980 (1994).

40. Williams, R. M. & Rimsky, S. Molecular aspects of the E. coli nucleoid protein, H-NS: A central controller of gene regulatory networks. *FEMS Microbiology Letters* **156**, 175–185 (1997).

41. La Teana, A. *et al.* Identification of a cold shock transcriptional enhancer of the Escherichia coli gene encoding nucleoid protein H-NS. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 10907–11 (1991).

42. Ripley, B. D. *Statistical Inference for Spatial Processes* (Cambridge University Press, Cambridge, 1988).

43. Baddeley, A. & Gill, R. D. Kaplan-Meier estimators of distance distributions for spatial point processes. *Annals of Statistics* **25**, 263–292 (1997).

44. Hajstcsch, K. Some remarks on estimators of the distribution function of nearest neighbour distance in stationary spatial point processes. *Series Statistics* **15**, 409–412 (1984).

45. Baddeley, A. & Turner, R. spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal Of Statistical Software* **12**, 1–42 (2005).

46. Baddeley, A., Rubak, E. & Turner, R. *Spatial Point Patterns: Methodology and Applications with* $\{R\}$ (Chapman and Hall/CRC Press, London, 2015).

47. R Core Team. R: A Language and Environment for Statistical Computing (2016).

48. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology* **2**, 2006.0008 (2006).

49. Edelstein, A., Amodaj, N., Hoover, K., Vale, R. & Stuurman, N. Computer Control of Microscopes Using Micro-Manager. In *Current Protocols in Molecular Biology* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2010).

50. Wolter, S. *et al.* rapidSTORM : accurate , fast open-source software for localization microscopy orcae : online resource for community annotation of eukaryotes. *Nature methods* **9**, 1040–1 (2012).

51. Wang, Y. *et al.* Localization events-based sample drift correction for localization microscopy with redundant cross-correlation algorithm. *Optics express* **22**, 15982–91 (2014).

Acknowledgments

We thank Hazel Soto-Montoya, Dr. William Navarre and Dr. Joshua N. Milstein for providing a bacteria strain (K12ΔHNS/pHNS-mEos3.2C1) and equipment for single-molecule localization imaging. This work was supported by the Human Frontier Science Program (LT000752/2014-C to Y.W.) and by the Arkansas Biosciences Institute, the major research component of the Arkansas Tobacco Settlement Proceeds Act of 2000. The plasmid mEos3.2-C1 was a gift from Michael Davidson (Addgene plasmid # 54550).

Author contributions statement

Y.W. conceived the project, S.J. and Y.W. conducted the simulations, Y.W. performed the experiments, S.J., S.D.C. and Y.W. analyzed the results. S.J., S.D.C. and Y.W. wrote and reviewed the manuscript.

Additional Information

Competing interests

The authors declare no competing financial interests.

Corresponding author

Correspondence to Yong Wang.

Figure Legends

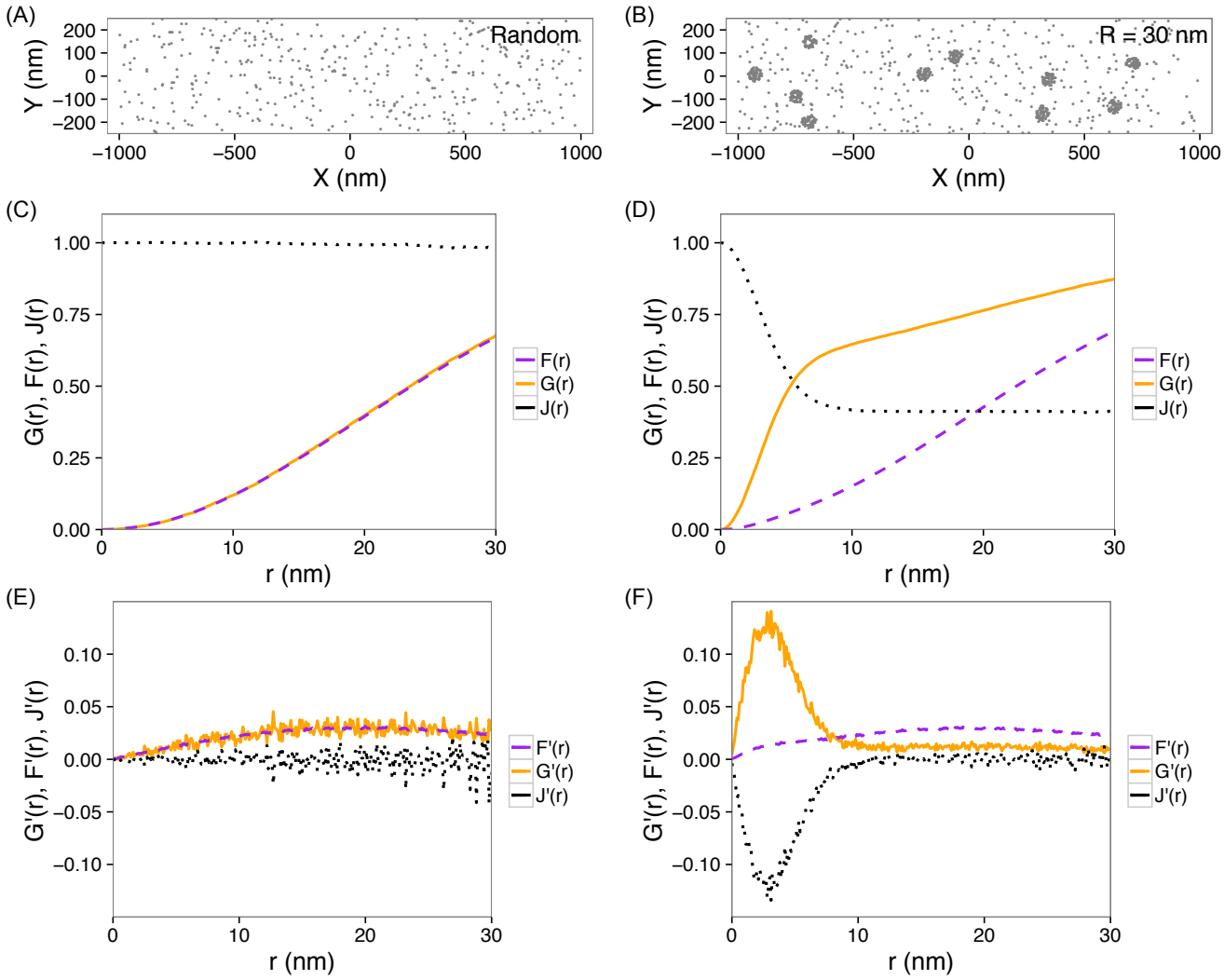

Figure 1. $G(r)$, $F(r)$ and $J(r)$ functions, and their derivatives. (A) Simulated noise points. (B) Simulated points forming clusters with a radius of $R = 30$ nm, in the presence of noise points. (C, D) $G(r)$, $F(r)$ and $J(r)$ functions calculated from the points in (A) and (B), respectively. (E, F) Derivatives, $G'(r)$, $F'(r)$ and $J'(r)$, calculated from the points in (A) and (B), respectively.

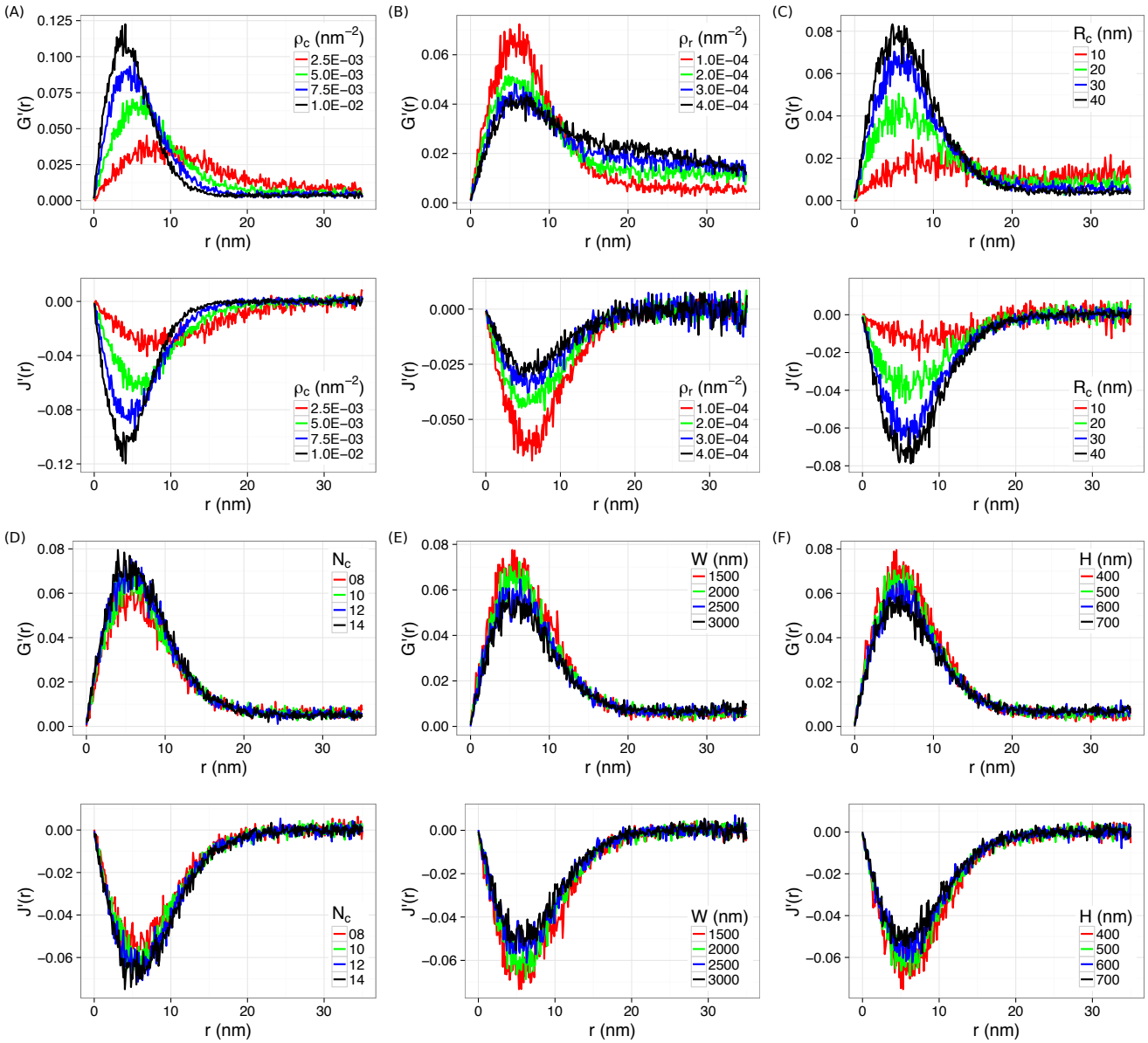

Figure 2. Changes in $G'(r)$ and $J'(r)$ by varying a cluster feature at a time. (A) ρ_c, (B) ρ_r, (C) R_c, (D) N_c, (E) W, and (F) H.

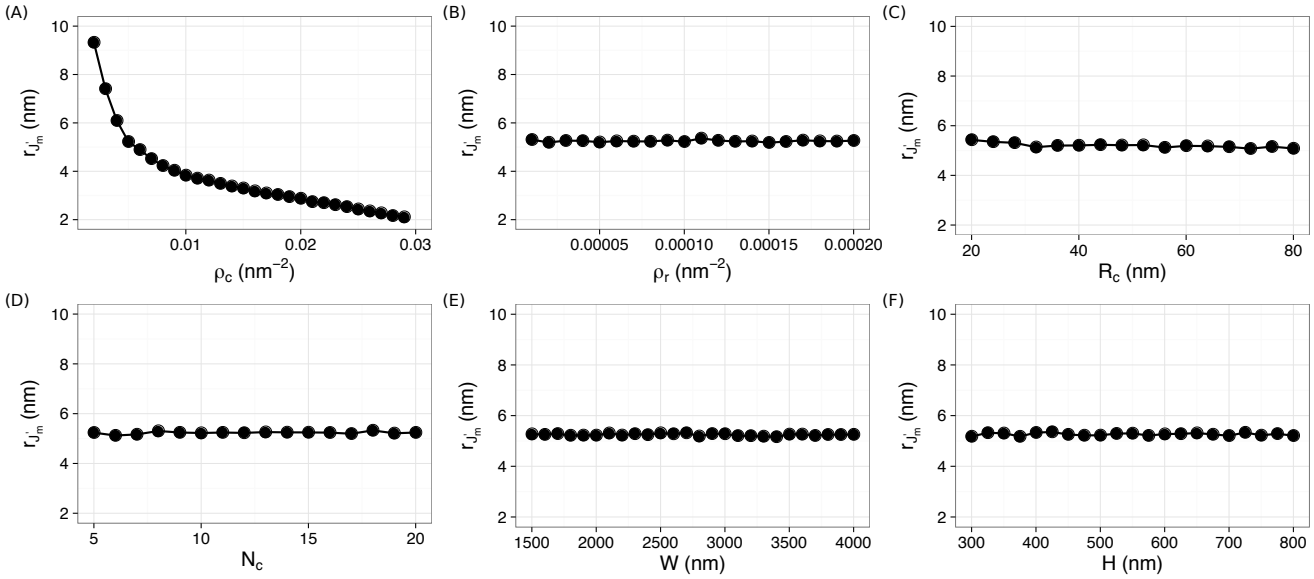

Figure 3. Dependence of $r_{J'_m}$ on the clustering features: (A) ρ_c, (B) ρ_r, (C) R_c, (D) N_c, (E) W, and (F) H.

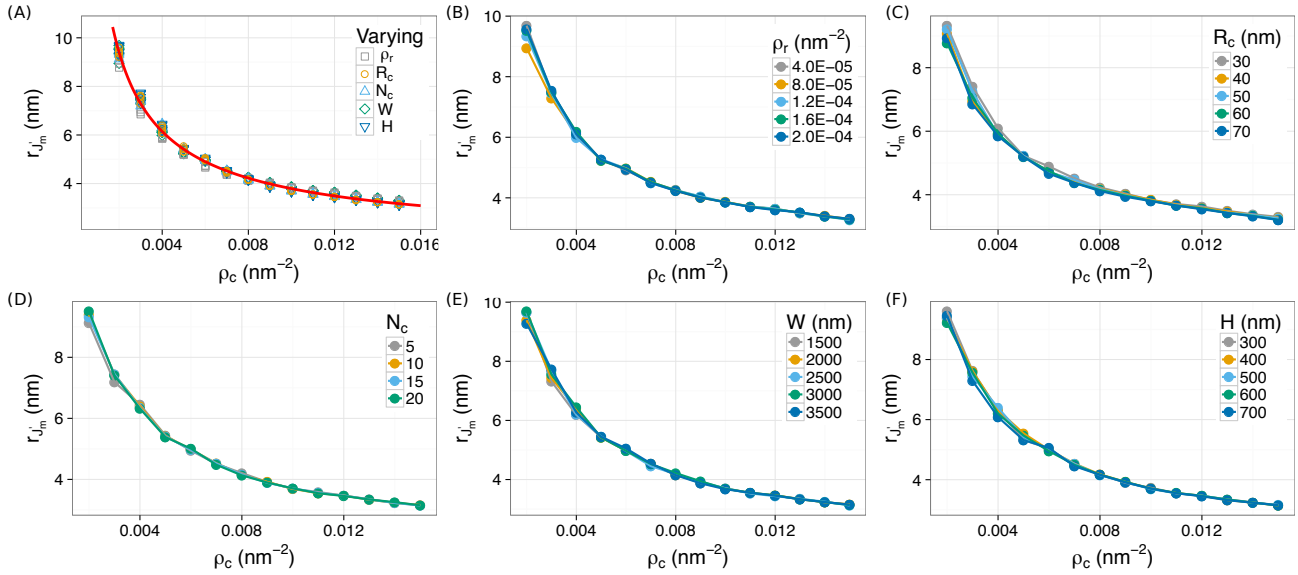

Figure 4. The $r_{J'_m} - \rho_c$ relation is independent on all the other cluster features, R_c, ρ_r, N_c, W, and H. All data points collapse onto a single power-law curve, $r_{J'_m} = A \cdot {\rho_c}^{-\alpha} + b$. Least-square fitting gives $\alpha = 0.76 \pm 0.03$.

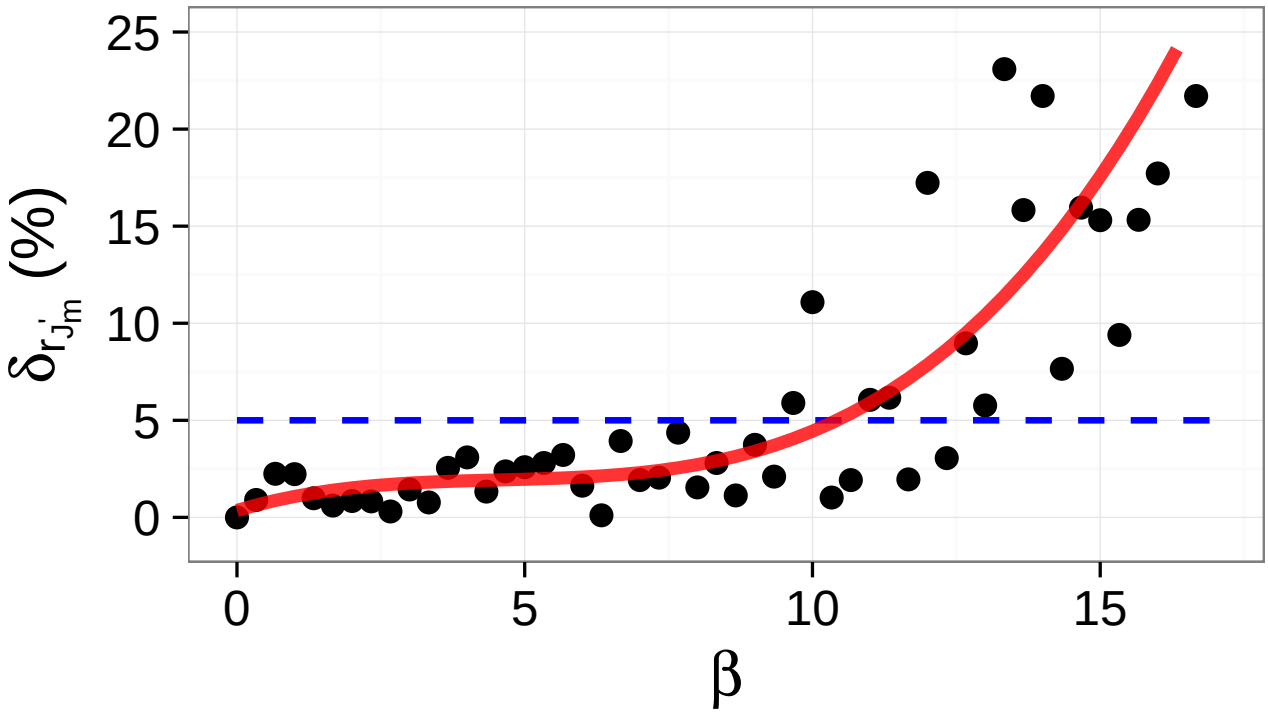

Figure 5. Robustness of the $r_{J'_m} - \rho_c$ relation. Given a certain density of clustering points (ρ_c), $r_{J'_m}$ remain (almost) constant (i.e., the relative error $\delta_{r_{J'_m}}$ is $< 5\%$) until there are 10 times more noise points than clustering points ($\beta \leq 10$).

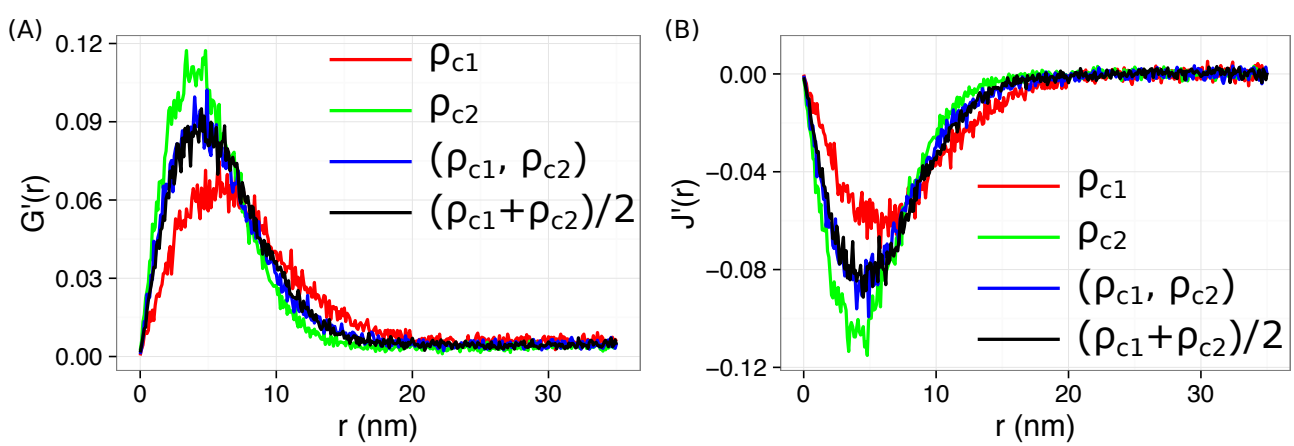

Figure 6. $G'(r)$ and $J'(r)$ for data with heterogeneous clusters with two different clustering densities.

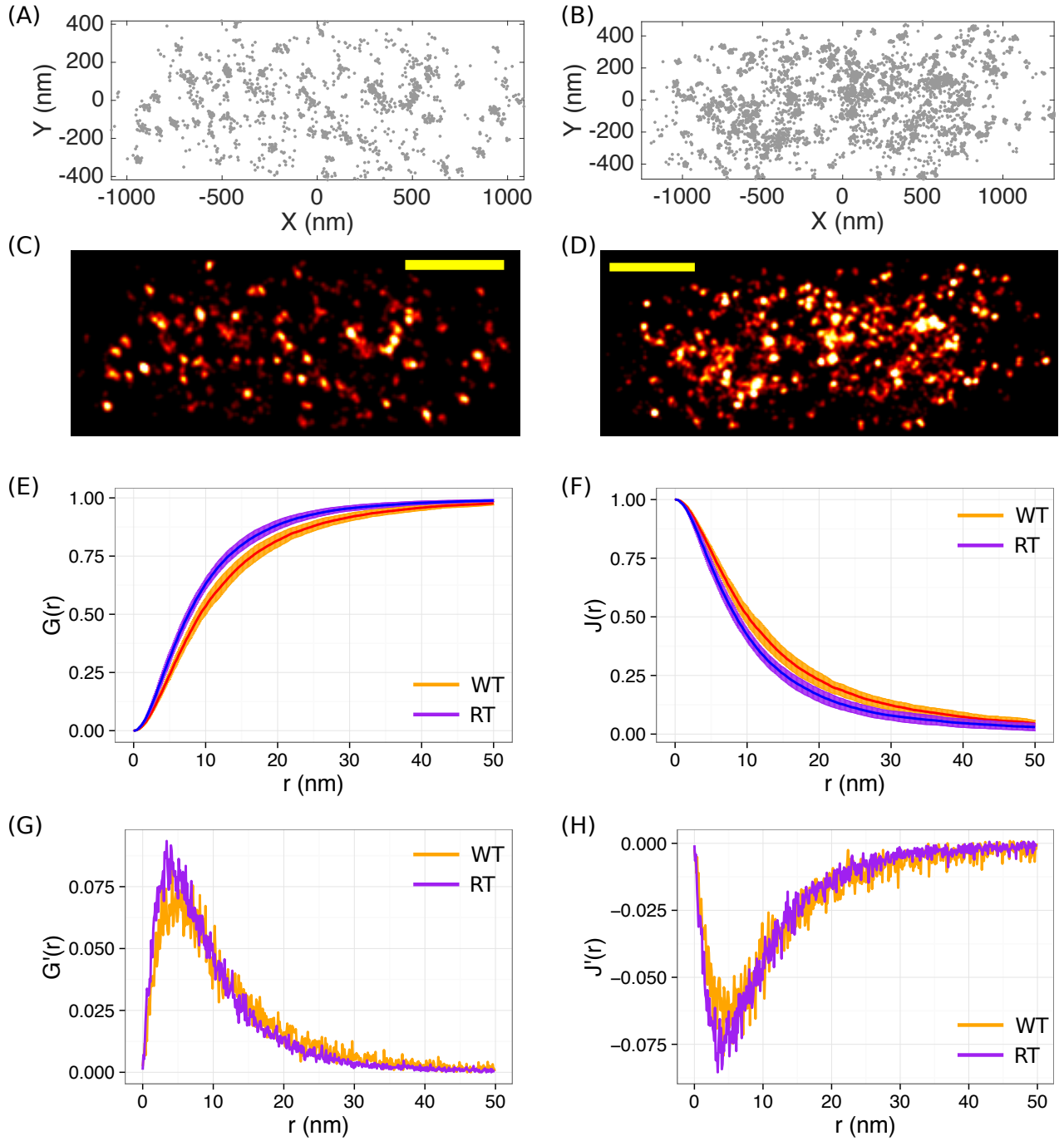

Figure 7. Clustering of H-NS proteins upon cold shock. (A, B) Localizations of H-NS proteins in *E. coli* (A) before and (B) after cold shock. (C, D) Super-resolved images of the bacteria in (A) and (B), respectively. (E, F) $G(r)$ and $J(r)$ functions computed for bacteria before (WT) and after (RT) cold shock. The error bars are standard deviation computed from 20 (WT) or 31 (RT) cells. (G, H) $G'(r)$ and $J'(r)$ computed from (E) and (F), respectively.

Supplementary Information

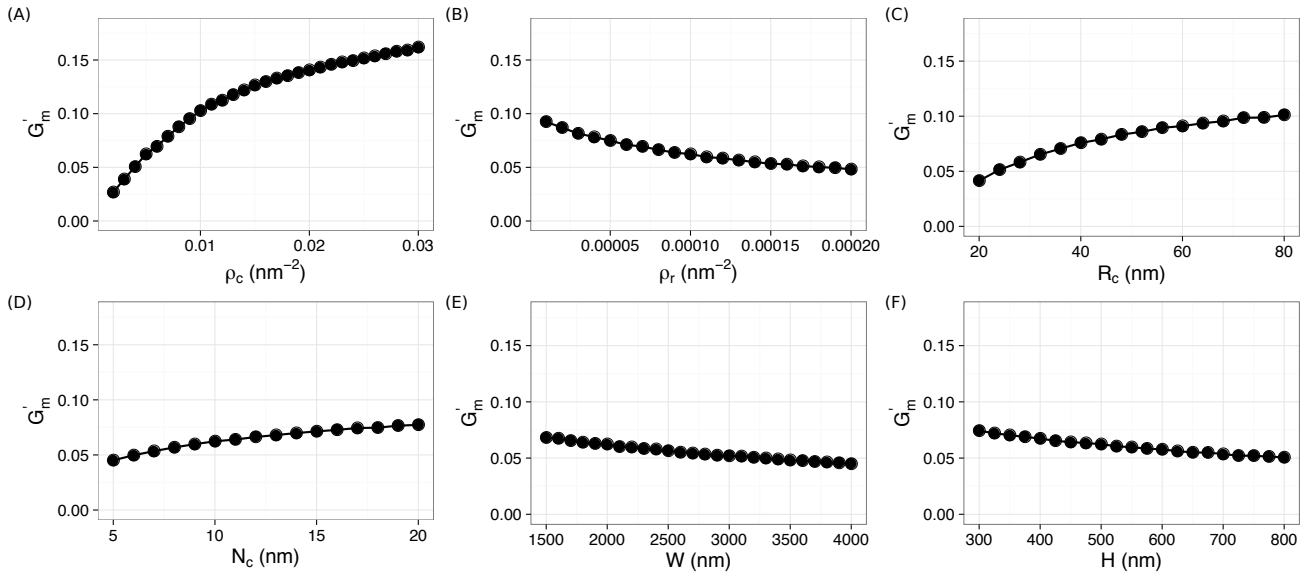

Figure S1. Dependence of G'_m on the clustering features: (A) ρ_c, (B) ρ_r, (C) R_c, (D) N_c, (E) W, and (F) H.

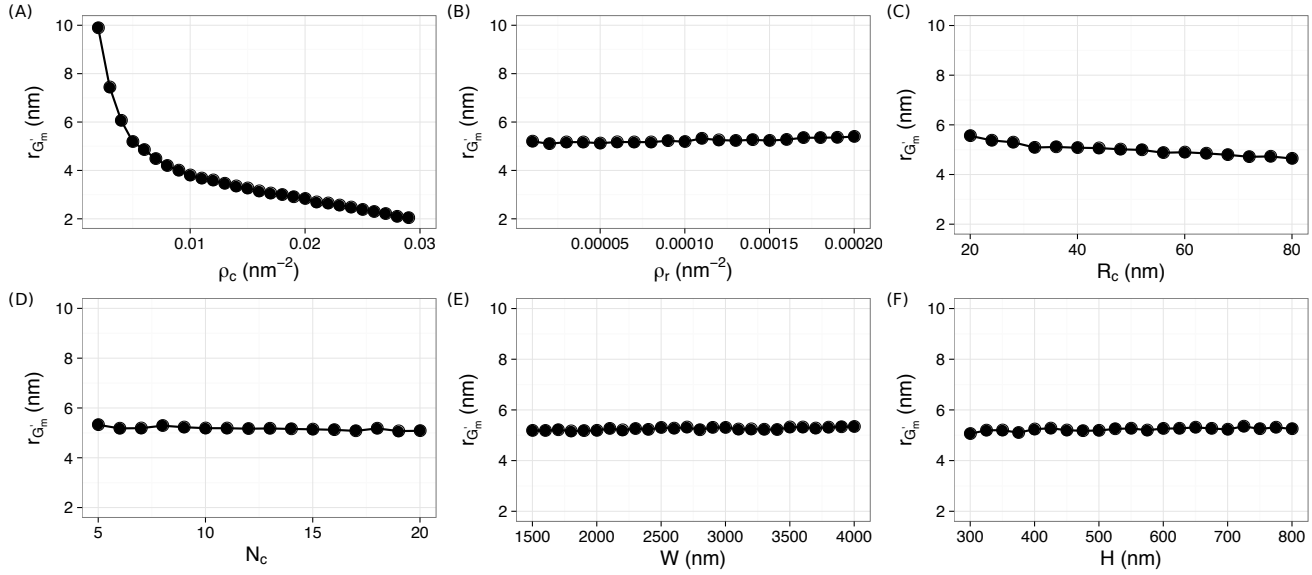

Figure S2. Dependence of $r_{G'_m}$ on the clustering features: (A) ρ_c, (B) ρ_r, (C) R_c, (D) N_c, (E) W, and (F) H.

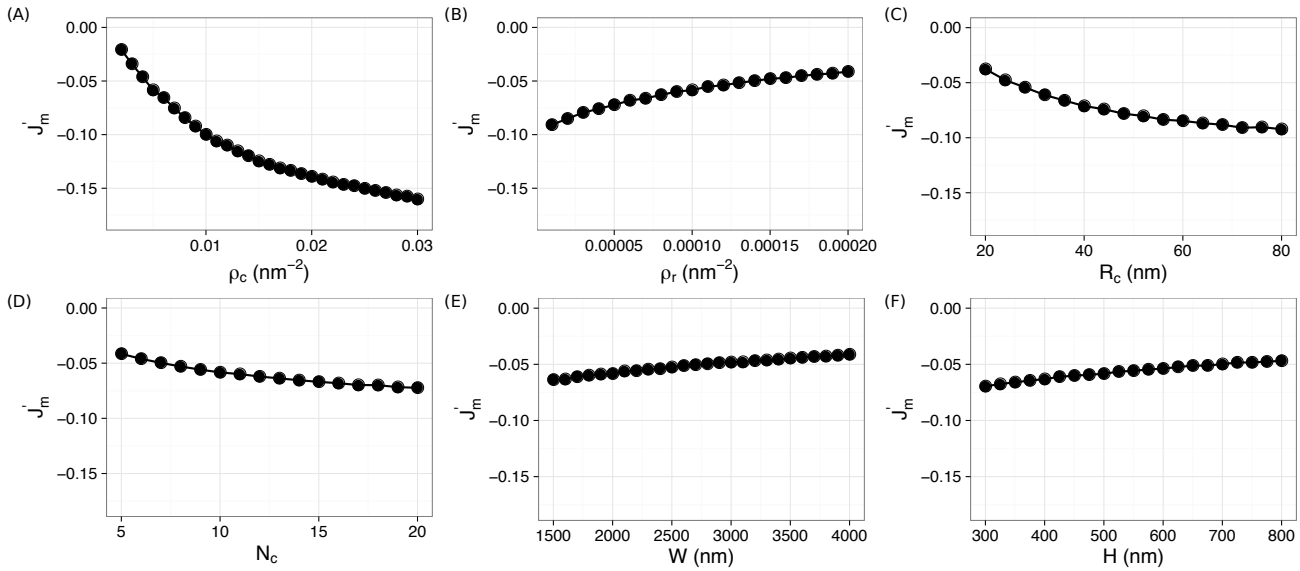

Figure S3. Dependence of J'_m on the clustering features: (A) ρ_c, (B) ρ_r, (C) R_c, (D) N_c, (E) W, and (F) H.

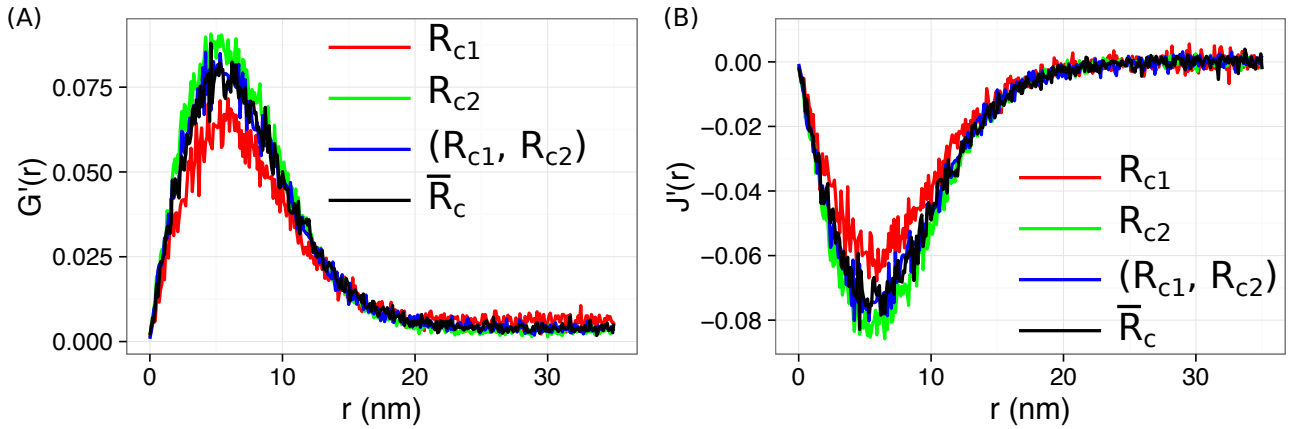

Figure S4. $G'(r)$ and $J'(r)$ from heterogeneous clusters with different radii, $\bar{R}_c = \sqrt{\left(R_{c1}^2 + R_{c2}^2\right)/2}$.

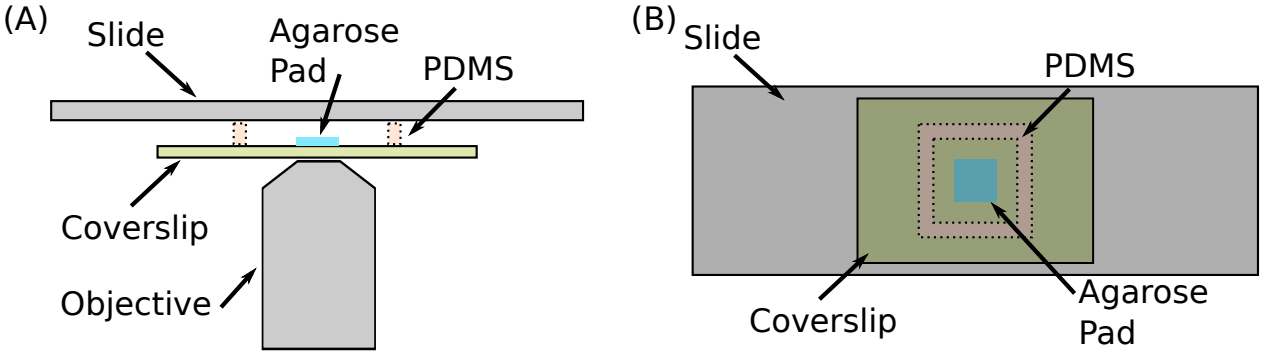

Figure S5. Mounting of *E. coli* bacteria expression fluorescent H-NS protein on agarose pads for SMLM imaging. (A) Side view. (B) Top view.