Biophysical Journal, Volume *113*

Supplemental Information

Investigating Focal Adhesion Substructures by Localization Microscopy

Hendrik Deschout, Ilia Platzman, Daniel Sage, Lely Feletti, Joachim P. Spatz, and Aleksandra Radenovic

Investigating focal adhesion substructures by localization microscopy and expectation maximization of a Gaussian mixture

H. Deschout, I. Platzman, D. Sage, L. Feletti, J. P. Spatz and A. Radenovic

Supporting Material

Supporting Text

1. Expectation maximization of a Gaussian mixture (EMGM)

1.1 Classic algorithm

We apply expectation maximization of a Gaussian mixture (EMGM) [1] on single-molecule localization (SMLM) data to investigate the substructure of focal adhesions (FAs). The main assumption is that the FA subunits can be described as bivariate Gaussians. The spatial probability distribution of an FA subunit is thus given by:

$$G(\boldsymbol{r}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\boldsymbol{r}-\boldsymbol{\mu})^{\mathrm{T}}\cdot\boldsymbol{\Sigma}^{-1}\cdot(\boldsymbol{r}-\boldsymbol{\mu})\right) \quad (1)$$

where $\boldsymbol{r}$ is the position in which the Gaussian is being evaluated, $\boldsymbol{\mu}$ the center position of the Gaussian, and $\boldsymbol{\Sigma}$ the covariance matrix of the Gaussian. Assume one or more FAs consisting out of N positions $\boldsymbol{r}_n$. According to our assumption, these FAs can be modeled by a mixture of bivariate Gaussians. Assume that this mixture consists of K components with the weight of component k described by the mixing coefficient π_k. These mixing coefficients fulfil the condition:

$$\sum_{k=1}^{K}\pi_k = 1 \quad (2)$$

Expectation maximization is a popular algorithm to identify the properties $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and π_k of each component the Gaussian mixture. After choosing initial values, the expectation step consists of evaluating the posterior probability that localization $\boldsymbol{r}_n$ was generated from component k:

$$\gamma_{nk} = \frac{\pi_k G(\boldsymbol{r}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j G\left(\boldsymbol{r}_n\middle|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)} \quad (3)$$

In the maximization step, the parameters are re-estimated using the posterior probabilities:

$$\begin{aligned}\boldsymbol{\mu}_k^{\mathrm{new}} &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}\boldsymbol{r}_n \\ \boldsymbol{\Sigma}_k^{\mathrm{new}} &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}(\boldsymbol{r}_n-\boldsymbol{\mu}_k^{\mathrm{new}})\cdot(\boldsymbol{r}_n-\boldsymbol{\mu}_k^{\mathrm{new}})^{\mathrm{T}} \\ \pi_k^{\mathrm{new}} &= \frac{N_k}{N}\end{aligned} \quad (4)$$

Where N_k is defined as the number of localizations that belong to component k:

$$N_k = \sum_{n=1}^{N}\gamma_{nk} \quad (5)$$

Finally, the likelihood of the updated Gaussian mixture is calculated and checked for convergence:

$$\mathcal{L} = \prod_{n=1}^{N}\sum_{j=1}^{K}\pi_k^{\mathrm{new}}G\left(\boldsymbol{r}_n\middle|\boldsymbol{\mu}_j^{\mathrm{new}},\boldsymbol{\Sigma}_j^{\mathrm{new}}\right) \quad (6)$$

If the convergence criterion is not satisfied, the expectation and maximization steps described in Eqs. (3) and (4) are repeated.

1.2 Initialization by greedy learning

EMGM is known to be sensitive to local maxima. To avoid finding such a solution, initial values of the parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and π_k (see Supporting Text, Section 1.1) need to be chosen sufficiently close to the real values. In the context of SMLM, these values are not known. Although several approaches have been reported in order to initialize the model parameters for EMGM, there is no widely accepted method. Popular approaches are randomly generating the initial parameter values, or estimating them using the *k*-means clustering algorithm [1].

An interesting alternative to these initialization methods is the so-called "greedy learning" approach [2], based on repeating the EMGM by starting from a trivial Gaussian mixture consisting of one component, and each time adding an extra component. The EMGM solution obtained for a P-1 component mixture is used as initialization for the P component mixture, by deleting one component and inserting two random components, based on the deleted one. This can be done P-1 times, for each component of the old mixture, and the solution with the highest likelihood is retained. By doing so, one proceeds until a desired number of components K is attained. Additionally, each step consisting of P-1 initializations can be repeated Q times to increase the accuracy of the result. The total number of EMGM repeats to obtain the correct solution of K components is thus given by $Q\left(1+\sum_{i=1}^{K} i\right)$.

This shows that the initialization procedure becomes computationally more expensive for datasets containing more components. The computation time on a mid-range personal computer for the simulations shown in Fig. 2 ranged from ~3 s (for K = 1 and Q = 3) to ~1000 s (for K = 20 and Q = 3). Note that we actually used $Q\left(1+\sum_{i=1}^{K}[i+1]\right)$ initializations due to an extra background "component" (see Supporting Text, Section 1.4).

1.3 Model selection by hypothesis testing

When applying EMGM, the number of components K for the Gaussian mixture needs to be chosen. In the context of SMLM, this number is unknown. In order to select the most appropriate number of components, one can repeat the EMGM procedure for a range of K values. The likelihood value is not a good selection criterion, as increasing the number of components increases the likelihood monotonously. A solution provided by information theory is the Akaike or Bayes information criterion [1], which penalizes an increasing number of components and therefore leads to a maximum value for a certain K value. However, this value has been reported to typically overestimate the real number of components [3].

Hypothesis testing can provide a more conservative approach towards selecting to right mixture model [4]. Assume two mixtures calculated by EMGM, one containing K-1 components and the other containing K components. The K component model will have a larger likelihood than the K-1 component model. Consider the null hypothesis that the K-1 component model is the correct one, which will correspond to a specific distribution of likelihood increments. If the real model consists of more than K-1 components, the likelihood increment can be expected to be larger than the values described by the null hypothesis distribution. This distribution, however, is unknown, but can be simulated from the identified K-1 component model, i.e. a number of bootstrapped data sets are generated assuming the null hypothesis and the increments in likelihood are obtained by applying EMGM for both K-1 and K components. Comparing the real likelihood increment with the bootstrap null hypothesis distribution allows to determine the *p*-value, in turn allowing to accept or reject the null hypothesis. Choosing the maximum allowed *p*-value sufficiently small, e.g. equal to 0.01, means

that there is only a 1% chance to select a mixture model that contains too many components, preventing overestimation of the number of components.

1.4 Localization background

While initialization and model selection issues are inherent to EMGM, other problems arise because of the nature of SMLM data. One important problem is that not necessarily all localizations are part of FAs, but instead can belong to a background. Consider a SMLM dataset consisting of N positions that belong to a mixture of multivariate Gaussians, and an extra N_b positions that belong to a background, within an area A. In case of a simple uniform background, the probability distribution of the background localizations is given by:

$$B = \frac{1}{A} \tag{7}$$

The algorithm can readily be adjusted to incorporate the background described by B. First of all, the posterior probability that localization $\boldsymbol{r}_n$ was generated from component k (see Eq. (3)) is now given by:

$$\gamma_{nk} = \frac{\pi_k G(\boldsymbol{r}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j G(\boldsymbol{r}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) + B} \tag{8}$$

And an equivalent posterior probability for the background can be defined as:

$$\delta_n = \frac{B}{\sum_{j=1}^{K} \pi_j G(\boldsymbol{r}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) + B} \tag{9}$$

The re-estimation of the parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ can be done as before, while the re-estimation of the mixing coefficients (see Eq. (4)) has to be adjusted as follows:

$$\pi_k^{\mathrm{new}} = \frac{N_k}{N + N_\mathrm{b}} \tag{10}$$

where N_b can be calculated using the background posterior probabilities:

$$N_\mathrm{b} = \sum_{n=1}^{N} \delta_n \tag{11}$$

Finally, the calculation of the likelihood of the updated Gaussian mixture (see Eq. (6)) is adjusted as follows:

$$\mathcal{L} = \prod_{n=1}^{N+N_\mathrm{b}} \left\{ \sum_{j=1}^{K} \pi_k^{\mathrm{new}} G(\boldsymbol{r}_n|\boldsymbol{\mu}_j^{\mathrm{new}}, \boldsymbol{\Sigma}_j^{\mathrm{new}}) + B \right\} \tag{12}$$

The background can effectively be considered as an extra component of the Gaussian mixture, requiring an adaptation of the initialization procedure (see Supporting Text, Section 1.2). Initialization of a P component Gaussian mixture is done P times instead of $P - 1$ times (i.e. $P - 1$ initializations corresponding to each component of the previous solution, and 1 initialization corresponding to the background of the previous solution).

1.5 Localization uncertainty

The localizations in SMLM data contain measurement uncertainties [5]. The localization uncertainty can be described as an extra contribution $\boldsymbol{\varepsilon}$ to the real position of the molecule. This contribution is described by a spatial probability distribution that is usually modeled as a Gaussian:

$$E(\boldsymbol{\varepsilon}|s) = \frac{1}{2\pi s} \exp\left(-\frac{|\boldsymbol{\varepsilon}|^2}{2s^2}\right) \tag{13}$$

The standard deviation s is often termed as the localization uncertainty or precision. An observed localization $\boldsymbol{r}$ belonging to component k is described by the sum of $\boldsymbol{\varepsilon}$ and the real emitter position. Since both variables are independent, the spatial probability distribution of their sum is given by the convolution of their corresponding spatial probability distributions (see Eqs. (1) and (13)):

$$N(\boldsymbol{r}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, s) = \int_{-\infty}^{+\infty} E(\boldsymbol{r} - \boldsymbol{r}'|s) G(\boldsymbol{r}'|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\, d\boldsymbol{r}' \tag{14}$$

This is the convolution of two bivariate Gaussians, which can be solved as [6]:

$$G(\boldsymbol{r}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, s) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}_k + s^2\boldsymbol{I}|}} \exp\left(-\frac{1}{2}(\boldsymbol{r} - \boldsymbol{\mu}_k)^{\mathrm{T}} \cdot (\boldsymbol{\Sigma}_k + s^2\boldsymbol{I})^{-1} \cdot (\boldsymbol{r} - \boldsymbol{\mu}_k)\right) \tag{15}$$

where $\boldsymbol{I}$ is the identity matrix. This expression describes the observed spatial probability distribution of component k. In order to incorporate the effect of the localization uncertainty in EMGM, we need to adjust the algorithm in two ways. First of all, the expectation step needs to be adjusted, since the expression for the posterior probability γ_{nk} of position $\boldsymbol{r}_n$ of component k contains the spatial probability distribution of that component (see Eq. (3)). Substitution of Eq. (15) in Eq. (3) yields the adjusted posterior probability:

$$\gamma_{nk} = \frac{\pi_k G(\boldsymbol{r}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, s_n)}{\sum_{j=1}^{K} \pi_j G(\boldsymbol{r}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, s_n)} \tag{16}$$

where s_n is the localization uncertainty corresponding to localization $\boldsymbol{r}_n$. Secondly, the maximization step needs to be adjusted, because the apparent spatial probability distribution is a bivariate Gaussian with a covariance matrix equal to $\boldsymbol{\Sigma}_k + s^2\boldsymbol{I}$ (see Eq. (15)). This means that the presence of localization uncertainties affects both the shape and size of the observed component k. The re-estimation of the covariance matrix (see Eq. (4)) should be adjusted as follows:

$$\boldsymbol{\Sigma}_k^{\mathrm{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \{(\boldsymbol{r}_n - \boldsymbol{\mu}_k^{\mathrm{new}}) \cdot (\boldsymbol{r}_n - \boldsymbol{\mu}_k^{\mathrm{new}})^{\mathrm{T}} - s_n^2\boldsymbol{I}\} \tag{17}$$

The contribution coming from the localization uncertainty is included within the sum, since the value of the localization uncertainty can change for different localizations. Note that Eq. (17) suggests that the covariance matrix values of certain mixture components can possibly become negative during the EMGM procedure. If this occurs during EMGM, the covariance matrix is not updated, and the value of the previous iteration is retained.

2. Simulations

2.1 Simulation details

The simulations shown in Fig. 2 were performed in Matlab (The Mathworks). Briefly, Gaussian mixtures consisting of K components were simulated. The localizations in each component were obtained from a Gaussian probability distribution, using the Matlab function *mvnrnd*. The Gaussian standard deviation was σ_x = σ_y = 20 nm (except for Fig. 2E), and the number of localizations for each component was N_k = 100. The number of mixture components K was varied between 1 and 20 in Fig. 2B, and fixed at 4 in Fig. 2C-F. The centers of the mixture components were placed in a square grid with a spacing $d_{x,y}$ equal to five times $\sigma_{x,y}$ (except for Fig. 2F).

A uniform localization background was added in Fig. 2C by randomly generating a number of localizations from a uniform distribution, using the Matlab function *rand*. The number of background localizations was determined from the localization background density bg, which was varied between 0 and 50,000 #/µm^2, in steps of 1000 #/µm^2. The effect of the localization uncertainty shown in Fig. 2D was simulated by adding to each localization coordinate a value randomly generated from a Gaussian distribution with standard deviation s, using the Matlab function *randn*. The value of the localization uncertainty s was varied from 0 to a 40 nm, in steps of 1 nm. To account for the apparent increase in component size, the spacing between the component centers was adjusted to five times $\sqrt{\sigma_{x,y}^2 + s^2}$. The changing component eccentricity shown in Fig. 2E was simulated by increasing the component standard deviation σ_x from 2.8 to 20 nm, and simultaneously decreasing the standard deviation σ_y from 140 to 20 nm, resulting in eccentricities σ_x/σ_y increasing from 0.02 to 1. In Fig. 2F, the spacing $d_{x,y}$ between the component centers was increased from 0 to 200 nm, in steps of 5 nm. For each case, 100 simulations were performed.

2.2 Number of mixture components

The simulation results in Fig. 2B show that EMGM increasingly underestimates the number of mixture components for an increasing value of K. Additionally, the number of non-existing components (i.e. false positives) identified by EMGM also increases with K, as illustrated in Fig. S3B. We define K_{id} as the number of mixture components correctly identified by EMGM, and K_{fp} as the number of false positive components found by EMGM. Using the simulated data from Fig. 2B, we calculated the probability of obtaining a completely correct EMGM result (i.e. K_{id} = K and K_{fp} = 0) as a function of K. The results are shown in Fig. S3C. For mixtures with K < 10, this probability is on average equal to 94%. For larger numbers, the method starts to underestimate K, most likely because the contribution of correctly fitting individual components to the total likelihood becomes smaller with an increasing number. Fig. S3D shows the average values of K_{id} and K_{fp} as a function of K. The average number of false positives is smaller than 1 for mixtures with K < 10.

While mixtures of identical Gaussian components with equidistantly spaced centers allow an unambiguous interpretation of the effect of changing one of the mixture characteristics, they are not representative of the reality. We therefore performed additional simulations showing a complexity closer to the experimental situation. We simulated mixtures with a number of components K varying between 1 and 10 (i.e. the range in which the EMGM approach was found to perform well), while the component centers, orientation, and eccentricities were randomly generated. More specifically, the

standard deviation σ_x and σ_y were each randomly generated between 4 and 40 nm, while the center positions were randomly generated within a square region with an area of $K\pi$(20 nm)^2. Resulting components with an eccentricity σ_x/σ_y lower than 0.1 were rejected. The components were allowed to approach each other closely, the only restriction being that their 2σ ellipses did not overlap (resulting in a relative spacing that does not go below 4, cfr. Fig. S10). The results are shown in Fig. S4. Interestingly, the performance of our EMGM approach for these realistic datasets is not much worse than for the idealized case (Fig. 2B and Fig. S3). The probability of identifying all components correctly is slightly lower (Fig. S4C), and there is a larger spread on the average number of correctly identified components K_{id} (Fig. S4D).

2.3 Number of initializations

The initialization procedure (see Supporting Text, Section 1.2) consists of P-1 separate initializations for a P component Gaussian mixture. If the localization background is considered as an extra component, the procedure actually consists of P separate initializations for a P component mixture (see Supporting Text, Section 1.4). This procedure can be repeated several times Q to improve the accuracy of the EMGM result, resulting in a total of QP initializations for a P component Gaussian mixture. In order to investigate the effect of the value of Q on the EMGM performance, we performed simulations similar to the ones shown in Fig. 2B, for different values of Q. Fig. S5A shows that an increasing Q results in less underestimation of K, although the improvement is small for Q > 3. The number of false positive components K_{fp} does not seem to be affected by the value of Q (Fig. S5B). We therefore used Q = 3 (see Materials and Method).

2.4 Localization background

The adapted EMGM performs excellently in the presence of a uniform localization background (see Fig. 2C and Fig. S6, A and B). Only for values of the localization background density that are not representative for our experimental conditions (e.g. bg = 50,000 #/µm^2 in Fig S6C), the algorithm starts to underestimate the true amount of mixture components and finds false positive components. Using the simulated data from Fig. 2C, we calculated the probability of obtaining a completely correct EMGM result (i.e. $K_{\mathrm{id}} = K$ and K_{fp} = 0) as a function of bg (see Fig. S6C). For mixtures with bg < 25,000 #/µm^2, this probability is on average equal to 93%. Fig. S6D shows the average values of K_{id} and K_{fp} as a function of bg, confirming that the EMGM performance deteriorates for values larger than 25,000 #/µm^2. This is not a surprise, since the characteristic localization density of the component mixtures themselves is lower (each component counts 100 localization and has a standard deviation of $\sigma_{x,y}$ = 20 nm, resulting in a 2σ ellipse area of 0.016 µm^2, which yields a characteristic localization density around 20,000 #/µm^2).

The results shown in Fig. 2C and Fig. S6 were obtained from simulated Gaussian mixtures with a fixed number of components K = 4. We therefore also investigated the simultaneous effect of the localization background and the number of components on the EMGM performance. We simulated mixtures similar to Fig. 2B, varying K between 1 and 10 (i.e. the range in which the EMGM approach was found to perform well, see Supporting Text, Section 2.2) for different values of bg in the same range as in Fig. 2C. The results shown in Fig. S7 indicate that our EMGM approach generally performs well for values of bg up to 10,000 #/µm^2. For larger values, the method increasingly underestimates K, while the number of false positive components increases.

2.5 Localization uncertainty

The simulation results in Fig. 2D show that the estimated standard deviation $\sigma_{x,y}$ of the mixture components is slightly affected by an increasing localization uncertainty s. However, as illustrated in Fig. S8C, a high value of s can have an important impact on the values of K_{id} and K_{fp}. We assessed the probability of obtaining a completely correct EMGM result (i.e. $K_{\mathrm{id}} = K$ and $K_{\mathrm{fp}} = 0$) as a function of s, using the simulated data shown in Fig. 2D. The results are shown in Fig. S8D, indicating that the probability decreases strongly when s becomes larger than 30 nm. This is to be expected, since the localization uncertainty is larger than the standard deviation $\sigma_{x,y}$ = 20 nm of the mixture components itself. Fig. S8E shows K_{id} and K_{fp} as a function of s. For localization uncertainties larger than 30 nm, the average number of correctly identified components slightly decreases, while the average number of false positives increases more strongly.

The results shown in Fig. 2D and Fig. S8 were obtained from simulated Gaussian mixtures with a fixed number of components K = 4. We therefore also investigated the simultaneous effect of the localization uncertainty and the number of components on the EMGM performance. We simulated mixtures similar to Fig. 2B, varying K between 1 and 10 (i.e. the range in which the EMGM approach was found to perform well, see Supporting Text, Section 2.2) for different values of s in the same range as in Fig. 2D. The results shown in Fig. S9 indicate that our EMGM approach performs well for values of s up to 30 nm. For larger localization uncertainties, the EMGM algorithm breaks down. Interestingly, the effect of the localization uncertainty does not seem to depend on the number of mixture components, unlike for the localization background (Fig. S7).

2.6 Component eccentricity

The results in Fig. 2E suggest that the component eccentricity σ_x/σ_y does not have an effect on the performance of our EMGM approach. To verify this, we performed simulations similar to the ones shown in Fig. 2F, repeated for different values of σ_x/σ_y. The spacing d_x in the x-direction between the component centers was increased from 0 to 100 nm, while the spacing in the y-direction was taken equal to d_x divided by σ_x/σ_y (to ensure the same relative overlap between the components in both directions). Surprisingly, the results shown in Fig. S10A seem to suggest that the performance of the EMGM algorithm improves with an increasing eccentricity (i.e. a smaller value of σ_x/σ_y). This can be explained by the decreasing overlap between the components for the same spacing. Indeed, plotting the result as a function of the ratio d_x/σ_x shows almost no difference between the eccentricities (see Fig. S10B).

2.7 Number of localizations

The simulations presented in Fig. 2 describe Gaussian mixtures with components that each consist of N_k = 100 localizations. However, as illustrated in Fig. S11A, the performance of the EMGM algorithm can depend on the value of N_k. We assessed the probability of obtaining a completely correct EMGM result (i.e. $K_{\mathrm{id}} = K$ and $K_{\mathrm{fp}} = 0$) as a function of N_k, using simulations similar to Fig. 2A. The results are shown in Fig. S11D, indicating that the probability decreases strongly when N_k becomes smaller than 50. Fig. S11E shows K_{id} and K_{fp} as a function of N_k, indicating that this low probability is mainly due to EMGM not detecting all mixture components for low numbers of localizations.

3. Applying EMGM on experimental data

3.1 Scanning procedure

The number of FA substructures present in a typical SMLM dataset is not known, and can be assumed to be larger than 10. However, the simulation results in Fig. 2B indicate that the EMGM analysis is optimal when the Gaussian mixture consists of a smaller number of components. We therefore split the SMLM dataset into smaller subsets and perform the EMGM analysis on each subset separately. This can be done simply by scanning the original region of interest along non-overlapping square subregions with side length L, as illustrated in Fig. S2, A and B. However, this scanning procedure clips Gaussian mixture components that are not completely contained in a single subregion. A solution is repeating the scan with subregions that are shifted over a distance equal to $L/2$. If this shift is done in three different directions (as shown in Fig. S2, B-E), each component with dimensions below $L/2$ is completely included in at least one subregion of at least one scan. Considering that the FA substructures of interest have sizes below the diffraction limit, we choose L = 2 µm.

3.2 Combining procedure

Combining the EMGM results obtained from the scanning procedure (see Supporting Text, Section 3.1) consists of two steps: (1) the EMGM results of the subregions within each separate scan need to be combined, resulting in four different EMGM descriptions of the same original dataset, and (2) combining these four results yields the final EMGM result.

For the first step, we make the approximation that all components identified in a subregion are completely described by the localizations within that subregion. The posterior probability (see Eq. (3)) of a localization within a certain subregion belonging to a component identified in another subregion will therefore be zero. This means that the posterior probabilities of all M subregions of a single scan can be assembled into a sparse matrix $\boldsymbol{\gamma}_{\text{scan}}$ to describe the posterior probabilities of the full dataset:

$$\boldsymbol{\gamma}_{\text{scan}} = \begin{bmatrix} \boldsymbol{\gamma}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\gamma}_M \end{bmatrix} \tag{18}$$

where the matrices $\boldsymbol{\gamma}_i$ describe the posterior probabilities of the localizations inside subregion i, with i = 1, …, M. The posterior probabilities corresponding to the full dataset for a localization to belong to the background (see Eq. (9)) are similarly given by:

$$\boldsymbol{\delta}_{\text{scan}} = \begin{bmatrix} \boldsymbol{\delta}_1 \\ \vdots \\ \boldsymbol{\delta}_M \end{bmatrix} \tag{19}$$

This approximation is not optimal for the case of Gaussian mixture components being clipped (see Supporting Text, Section 3.1). The column in $\boldsymbol{\gamma}_{\text{scan}}$ corresponding to such a clipped component is therefore deleted, and its values are added to $\boldsymbol{\delta}_{\text{scan}}$. The criterion for determining whether a component is clipped is chosen as whether its 2σ error ellipse (containing around ~95% of localizations) is completely inside the subregion or not.

The resulting $\boldsymbol{\gamma}_{\text{scan}}$ does not provide a complete description of the Gaussian mixture in the full dataset due to the deletion of components that are clipped during the scanning procedure. However, each clipped component that is deleted from a certain scan is, in theory, identified in at least one of the three other scans (see Fig. S2). The second step therefore consists of merging the $\boldsymbol{\gamma}_{\text{scan}}$ matrices of the four different scans. For this purpose, the Pearson correlation between the posterior probabilities of each pair of components i and j belonging to different scans is calculated:

$$\rho_{ij} = \frac{\sum_n (\gamma_{in} - \overline{\gamma_{in}})\,(\gamma_{jn} - \overline{\gamma_{jn}})}{\sqrt{\sum_n (\gamma_{in} - \overline{\gamma_{in}})^2 \sum_k (\gamma_{jn} - \overline{\gamma_{jn}})^2}} \tag{20}$$

The sum runs over all localizations n that have a non-zero posterior probability (i.e. excluding all localizations outside subregion i and j). The correlation is tested against the null hypothesis that the posterior probabilities of components i and j are not correlated (i.e. it is verified that the correlation is larger than the values described by a simulated null hypothesis distribution). Two components identified in two different scans are considered to be identical if their correlation is significant according to the null hypothesis and if the correlation is larger than any other significant correlation involving either i or j. After identifying all identical components, their posterior probabilities are combined by averaging, while the posterior probabilities of components identified in only one scan are retained. This results in a final $\boldsymbol{\gamma}$ that describes the full dataset without clipped components. The background posterior probabilities are combined similarly into a final $\boldsymbol{\delta}$.

4. Merging procedure

The merging procedure illustrated in Fig. 5A is performed by splitting the mixture components obtained by EMGM into two categories: the ones whose 1σ error ellipse intersects with at least one other error ellipse, called the "overlapping" components, and the ones whose 1σ error ellipse does not intersect with another one, called the "isolated" components. The 1σ error ellipse is chosen because it corresponds to the probability of containing ~40% of all localizations. This means that localizations on the intersection between two such error ellipses have approximately an equal probability to belong to both corresponding components, therefore suggesting that they can be viewed as a single merged object. Once a set of K_{overlap} overlapping components have been verified, a new merged object can be calculated by summing their posterior probabilities γ_{nk} (see Eq. (3)):

$$\gamma_{n,\text{merged}} = \sum_{k=1}^{K_{\text{overlap}}} \gamma_{nk} \tag{21}$$

The properties of the merged object can then be calculated using Eq. (4). This gives rise to a third category, called the "merged" components.

5. PAINT imaging of integrin β3

5.1 Sample preparation

We used a commercial kit (Ultivue-2, Ultivue) for our points accumulation in nanoscale topography (PAINT) [7] experiments. The sample was prepared according to the manufacturer's recommendations. Briefly, we seeded around 10^5 REF cells on a fibronectin-coated 25 mm diameter cover slip, incubated them at 37° C in cell culture medium, washed them with PBS after 24h, and fixed them with 2.5% paraformaldehyde at 37° C for 10 minutes (see Materials and Methods). After removing the fixative, the cells were washed three times with PBS, the cover slip was placed into a custom made holder, and they were incubated in PBS for 10 minutes at 37° C.

The cells were subsequently reduced by incubating them for 10 minutes in a freshly prepared 0.1% sodium borohydride solution at room temperature. Afterwards, the cells were washed three times with PBS, and incubated in PBS for 10 minutes at room temperature. Next, the cells were incubated for 1.5h at room temperature in a blocking and permeabilization buffer consisting of PBS with 3% bovine serum albumin and 0.2% Triton X-100.

The primary antibody staining was carried out by incubating the cells overnight at 4 °C with integrin β3 mouse monoclonal antibodies (sc-7311, Santa Cruz Biotechnology) diluted 100 times in staining buffer composed of PBS with 1% bovine serum albumin and 0.2% Triton X-100. Next, the cells were washed four times with PBS, and incubated in PBS for 10 minutes at room temperature. The secondary antibody staining was carried out by first incubating the cells in Antibody Dilution Buffer (Ultivue-2, Ultivue) for 10 minutes at room temperature, and then for 2h with Goat-anti-Mouse-D1 antibodies (Ultivue-2, Ultivue) diluted 100 times in Antibody Dilution Buffer. Next, the cells were washed four times with PBS, and incubated in PBS for 10 minutes at room temperature.

5.2 Imaging procedure

Prior to imaging, 100 nm gold nanospheres (C-AU-0.100, Corpuscular) were added to the sample for lateral drift correction (see Materials and Methods). Imaging was performed using image strand I1-560 (Ultivue-2, Ultivue) diluted in Image Buffer (Ultivue-2, Ultivue) at a concentration of 1 nM. The imaging procedure was similar as for the PALM measurements (see Materials and Methods).

5.3 Discussion

We used PAINT to image fixed rat embryonic fibroblast (REF) cells where integrin β3 was antibody stained. The resulting PAINT images show FAs as patchy structures (Fig. S14). We hypothesize that this is caused by difficulties in labeling integrin with antibodies, for instance due to cell membrane areas that are curved inwards, resulting in an integrin epitope that is more difficult to access. We also noticed that mostly the cell periphery was labelled, again suggesting that not all integrins are accessible for the antibodies.

6. Production of nano-patterned substrates

Nano-patterned substrates were prepared by means of block-copolymer micelle nanolithography (BCML) as previously described [8-10]. Briefly, quasi-hexagonally ordered gold nanoparticle arrays on cleaned 25 mm diameter microscope cover slips (#1.5 Micro Coverglass, Electron Microscopy Sciences) were fabricated using a toluene solution of poly(styrene)-block-poly(2-vinyl pyridine) (PS-b-P2VP, Polymer Source Inc.) [9, 10]. The PS-b-P2VP toluene solution was treated with $HAuCl_4$ (Sigma Aldrich) at a stoichiometric loading of ($P2VP/HAuCl_4$) = 0.5 and stirred for at least 24h in order to obtain gold nanoparticles (AuNPs) with a diameter between 6-8 nm. The lateral distance between the individual AuNPs was adjusted by varying the micellar coating process (spinning speed). Details concerning the applied block polymers and the spin casting processes are included in Table S1.
The area between the AuNPs was passivated with PLL-*g*-PEG (PLL(20kDa)-g[3.5]-PEG(2kDa), Susos AG) to prevent non-specific adhesion. The substrates were first activated in an oxygen plasma at 0.4 mbar and 150 W for 10 minutes. The PLL-*g*-PEG was diluted to a concentration of 0.25 mg/ml in a 10 mM HEPES buffer at pH 7.4. The freshly activated substrates were incubated upside down for 45 minutes at room temperature on a 60 µl drop of the PLL-*g*-PEG solution on parafilm in a moist chamber. Afterwards the substrates are washed once with milli-Q water. Following passivation, each surface was functionalized with cRGD pentapeptide (Peptide Specialty Laboratories GmbH) at a concentration of 25 µM in MilliQ water for 2h at room temperature. The cRGD pentapeptide was conjugated with a PEG spacer (6 units) that serves as a breach between the peptide and the cysteine. The physisorbed material was removed by thorough rinsing with MilliQ water and PBS.

References

1. Bishop, C.M., *Pattern recognition and machine learning*. 2006: Springer.
2. Verbeek, J.J., N. Vlassis, and B. Krose, *Efficient greedy learning of Gaussian mixture models.* Neural Computation, 2003. **15**(2): p. 469-485.
3. Busemeyer, J.R. and Y.M. Wang, *Model comparisons and model selections based on generalization criterion methodology.* Journal of Mathematical Psychology, 2000. **44**(1): p. 171-189.
4. Punzo, A., R.P. Browne, and P.D. McNicholas, *Hypothesis testing for parsimonious Gaussian mixture models.* arXiv, 2014. 1405.0377.
5. Deschout, H., et al., *Precisely and accurately localizing single emitters in fluorescence microscopy.* Nature Methods, 2014. **11**(3): p. 253-266.
6. Vinga, S. and J.S. Almeida, *Renyi continuous entropy of DNA sequences.* Journal of Theoretical Biology, 2004. **231**(3): p. 377-388.
7. Sharonov, A. and R.M. Hochstrasser, *Wide-field subdiffraction imaging by accumulated binding of diffusing probes.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(50): p. 18911-18916.
8. Arnold, M., et al., *Activation of integrin function by nanopatterned adhesive interfaces.* Chemphyschem, 2004. **5**(3): p. 383-388.
9. Platzman, I., et al., *Surface properties of nanostructured bio-active interfaces: impacts of surface stiffness and topography on cell-surface interactions.* Rsc Advances, 2013. **3**(32): p. 13293-13303.
10. Pallarola, D., et al., *Focal adhesion stabilization by enhanced integrin-cRGD binding affinity.* BioNanoMaterials, 2017. https://doi.org/10.1515/bnm-2016-0014.

Supporting Figures

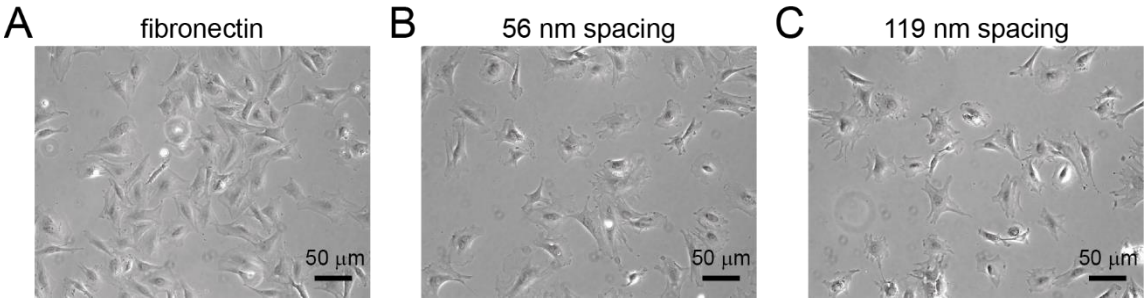

Figure S1. Phase-contrast microscopy imaging of REF cells. (A-C) The REF cells were growing on (A) a fibronectin-coated substrate, (b) a nano-patterned substrate with 56 nm spacing between AuNPs, or (C) a nano-patterned substrate with 119 nm spacing between AuNPs. The images were recorded 24h after transection with the integrin β3 vector.

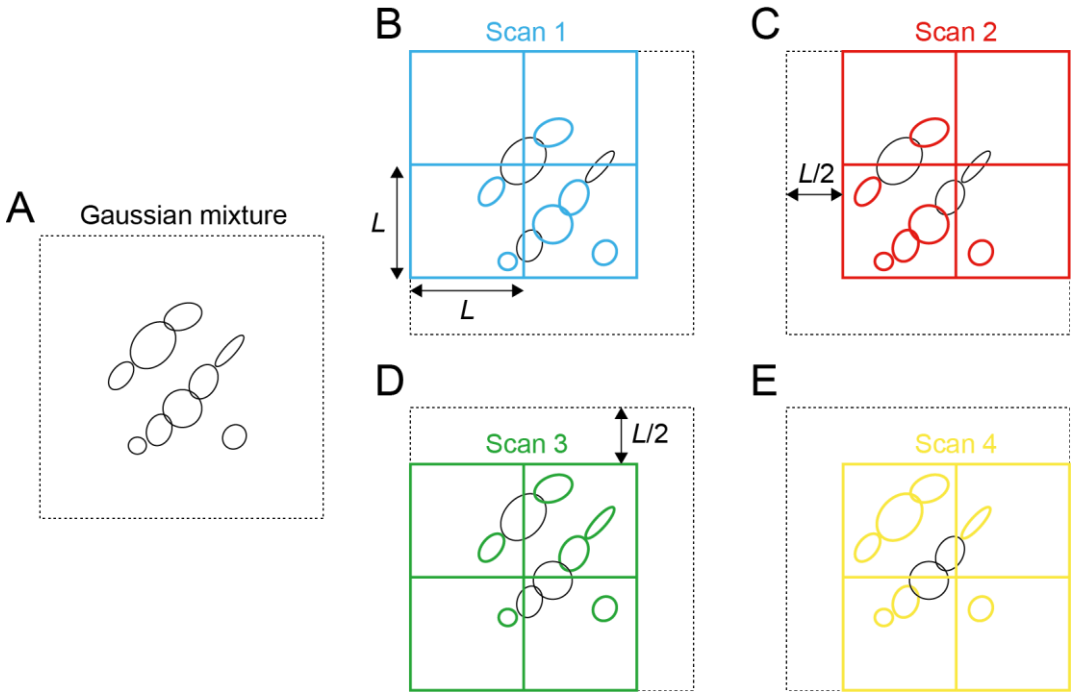

Figure S2. Scanning procedure for EMGM analysis of SMLM data. (A) Illustration of a Gaussian mixture with components represented by black ellipses. (B-E) Scanning procedure consisting of 4 different scans. During each scan, the EMGM analysis is performed on separate square subregions with a side length L, indicated by the colored squares. The Gaussian mixture components that can be correctly identified in a certain scan are indicated by the ellipses that have the same color as the squares. In between scans, the subregions are shifted over a distance $L/2$ in one of the following directions: left, right, up, or down.

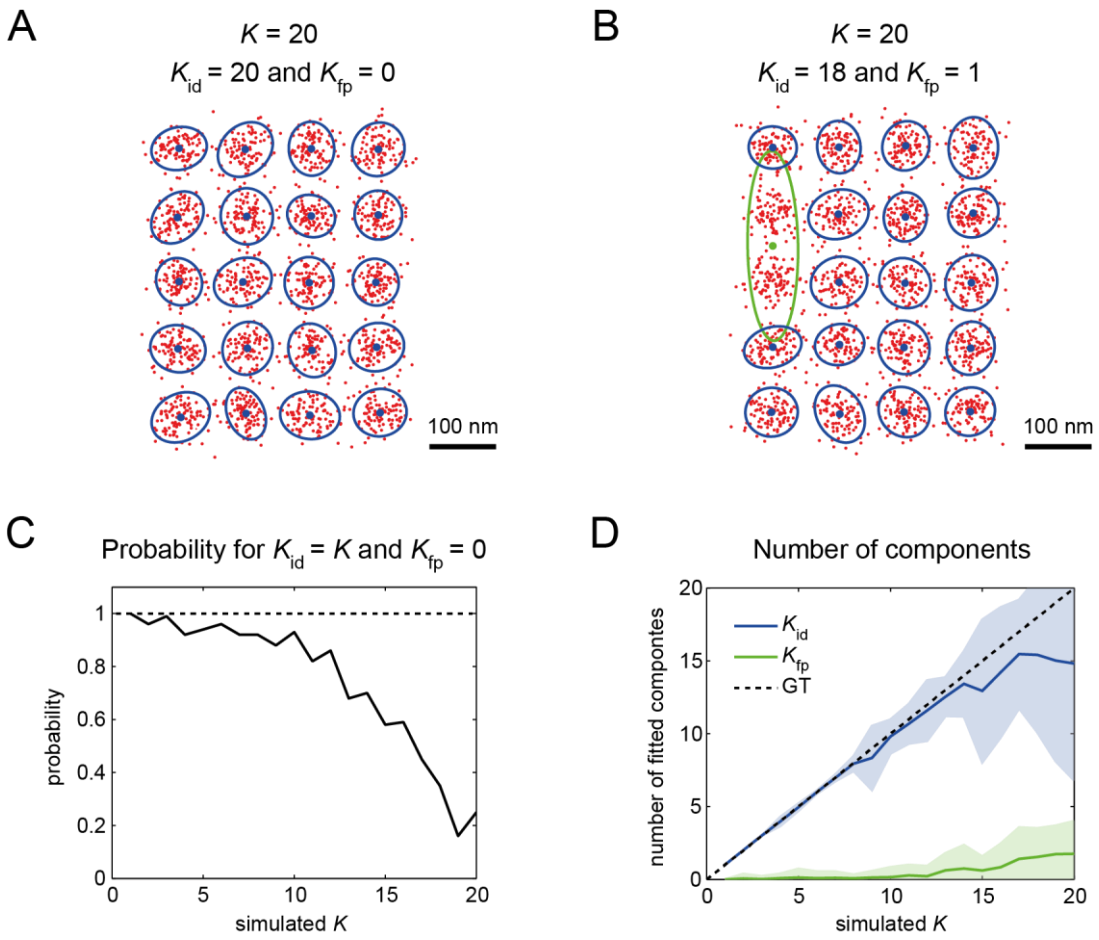

Figure S3. Influence of the number of mixture components K on the EMGM performance. (A-B) Example EMGM results for simulated Gaussian mixtures with K = 20 components. EMGM correctly identified K_{id} = 20 components and found K_{fp} = 0 false positive components for (A). EMGM correctly identified K_{id} = 18 components and found K_{fp} = 1 false positive component for (B). The red dots symbolize the simulated localizations. The blue/green dots symbolize the center positions of the correct/false positive components, the blue/green ellipses symbolize the 2σ error ellipses of the correct/false positive components. (D) The simulated probability of obtaining a completely correct EMGM result (i.e. $K_{\mathrm{id}} = K$ and K_{fp} = 0) as a function of K. (E) The simulated average values of K_{id} and K_{fp} as a function of K. The dashed line represents the ground truth (GT) and the shaded areas the standard deviation (n = 100).

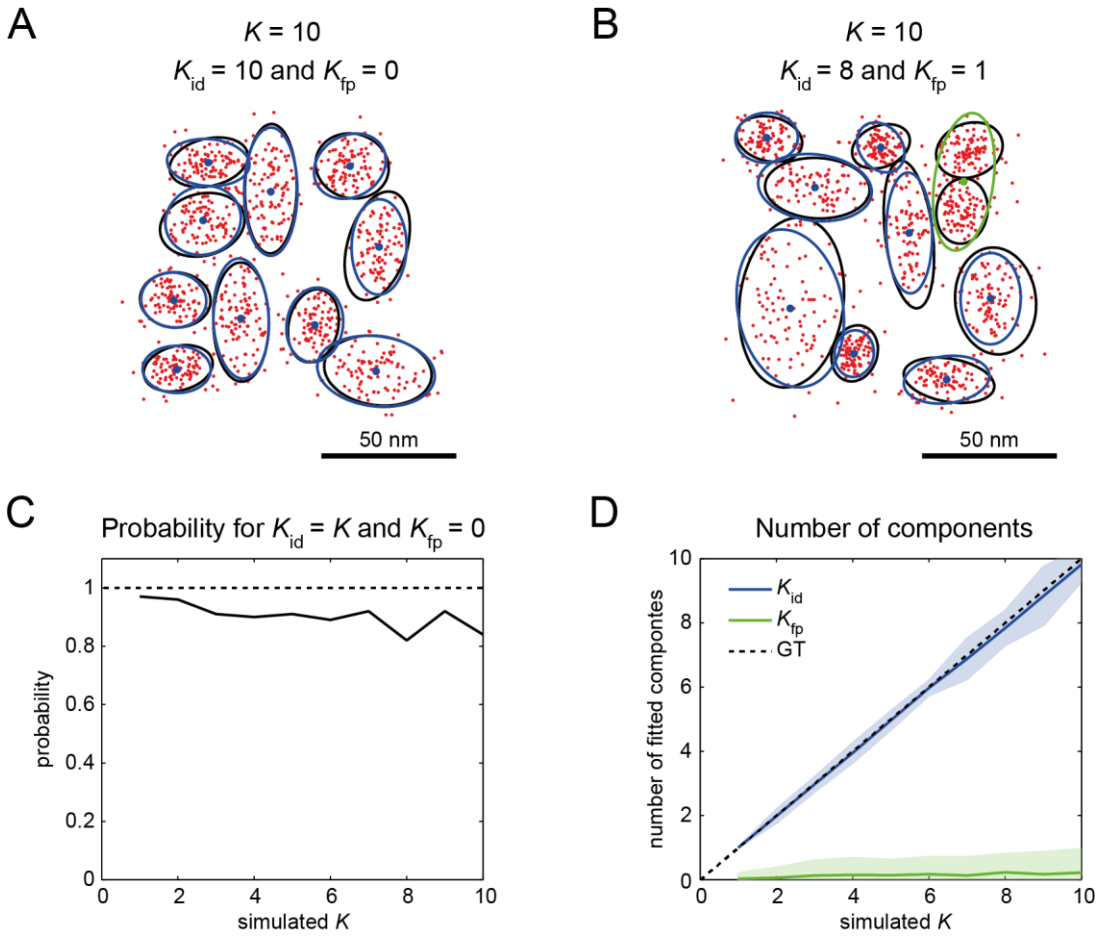

Figure S4. EMGM analysis on simulated random Gaussian mixtures. (A-B) Example EMGM results for simulated Gaussian mixtures with K = 10 components. EMGM correctly identified K_{id} = 10 components and found K_{fp} = 0 false positive components for (A). EMGM correctly identified K_{id} = 8 components and found K_{fp} = 1 false positive component for (B). The red dots symbolize the simulated localizations. The blue/green dots symbolize the center positions of the correct/false positive components, the blue/green ellipses symbolize the 2σ error ellipses of the correct/false positive components. The black ellipses symbolize the simulated components. (D) The simulated probability of obtaining a completely correct EMGM result (i.e. K_{id} = K and K_{fp} = 0) as a function of K. (E) The simulated average values of K_{id} and K_{fp} as a function of K. The dashed line represents the ground truth (GT) and the shaded areas the standard deviation (n = 100).

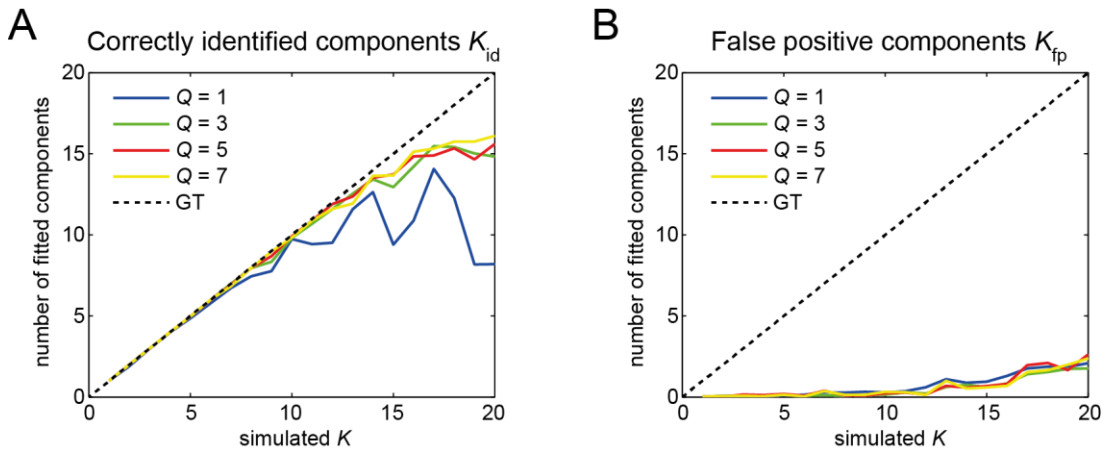

Figure S5. Influence of the number of initialization procedures Q on the EMGM performance. Gaussian mixtures with different values of K were simulated and analyzed by EMGM. (A) The average value of the number of correctly identified compenents K_{id} as a function of K, for different values of Q. (B) The average value of number of false positive components K_{fp} as a function of K, for different values of Q. The dashed line represents the ground truth (GT) and n = 100 simulations were performed.

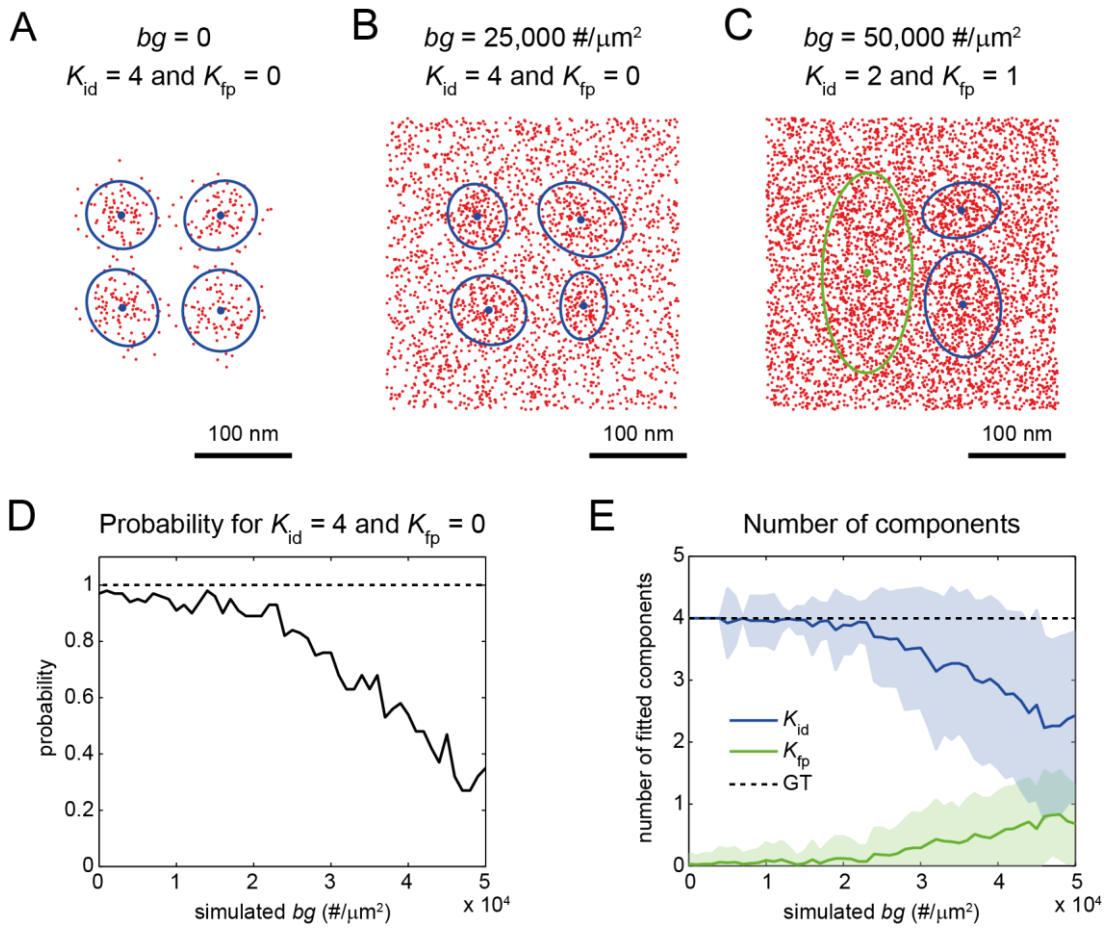

Figure S6. Influence of the localization background on the EMGM performance. (A-C) Example EMGM results for simulated Gaussian mixtures. Each mixture consists of K = 4 components with localization background density (A) bg = 0, (B) bg = 25,000 #/µm², or (C) bg = 50,000 #/µm². EMGM correctly identified K_{id} = 4 components and found K_{fp} = 0 false positive components for (A) and (B). EMGM correctly identified K_{id} = 2 components and found K_{fp} = 1 false positive component for (C). The red dots symbolize the simulated localizations. The blue/green dots symbolize the center positions of the correct/false positive components, the blue/green ellipses symbolize the 2σ error ellipses of the correct/false positive components. (D) The simulated probability of obtaining a completely correct EMGM result (i.e. K_{id} = 4 and K_{fp} = 0) as a function of bg. (E) The simulated average values of K_{id} and K_{fp} as a function of bg. The dashed line represents the ground truth (GT) and the shaded areas the standard deviation (n = 100).

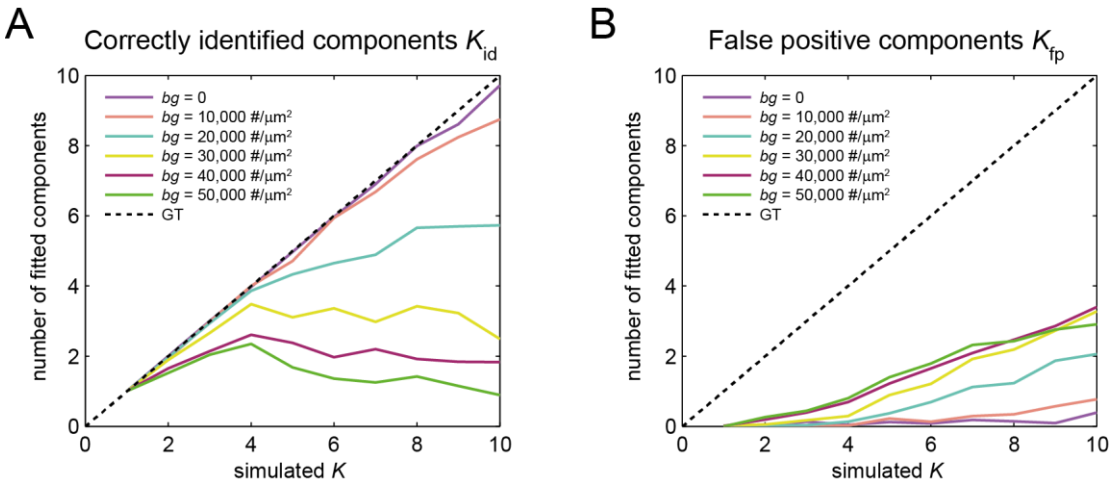

Figure S7. Influence of the localization background and the number of mixture components K on the EMGM performance. (A) Simulated average number of correctly identified components K_{id} as a function of K, for different values of the localization background density bg. (B) Simulated average number of false positive components K_{fp} as a function of K, for different values of the localization background density bg. The dashed line represents the ground truth (GT) and n = 100 simulations were performed.

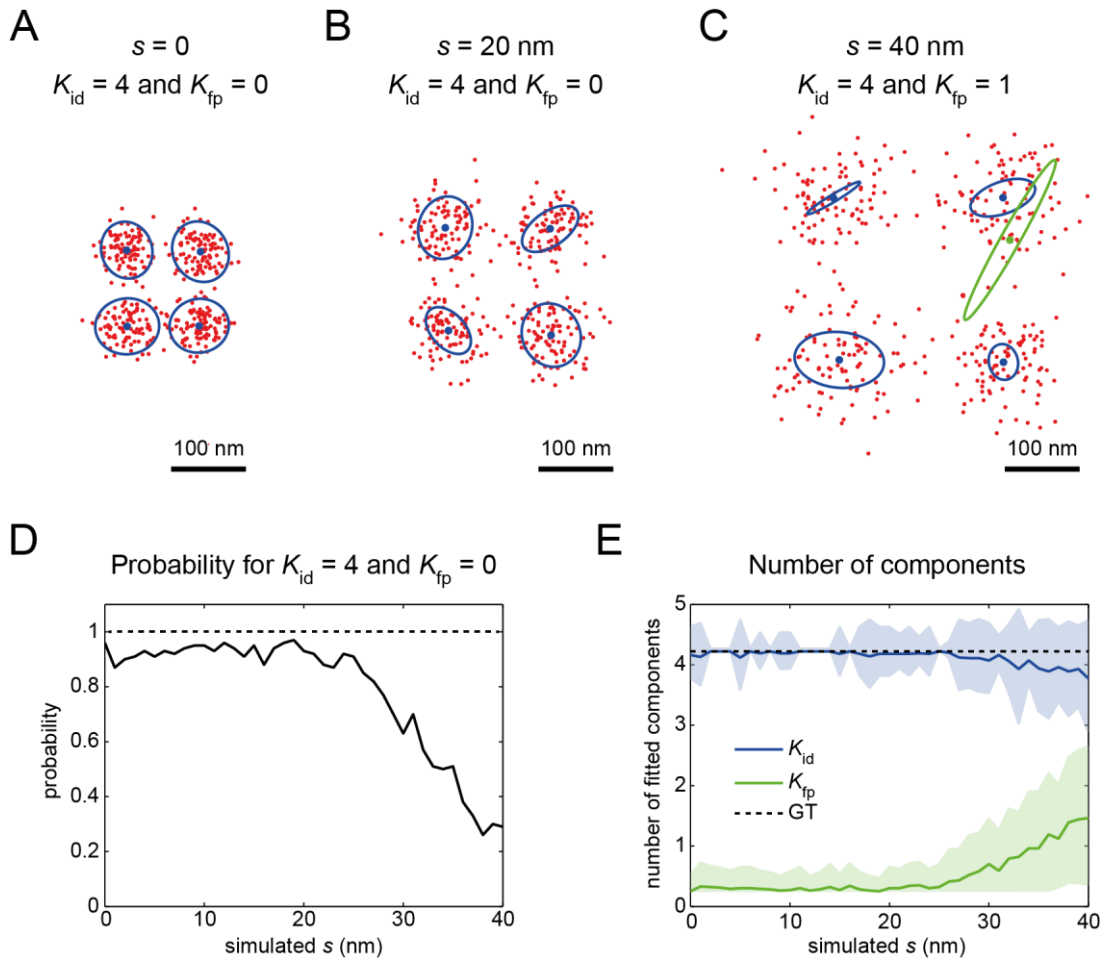

Figure S8. Influence of the localization uncertainty on the EMGM performance. (A-C) Example EMGM results for simulated Gaussian mixtures. Each mixture consists of K = 4 components with localization uncertainty (A) s = 0, (B) s = 20 nm, or (C) s = 40 nm. EMGM correctly identified K_{id} = 4 components and found K_{fp} = 0 false positive components for (A) and (B). EMGM correctly identified K_{id} = 4 components and found K_{fp} = 1 false positive component for (C). The red dots symbolize the simulated localizations. The blue/green dots symbolize the center positions of the correct/false positive components, the blue/green ellipses symbolize the 2σ error ellipses of the correct/false positive components. (D) The simulated probability of obtaining a completely correct EMGM result (i.e. K_{id} = 4 and K_{fp} = 0) as a function of s. (E) The simulated average values of K_{id} and K_{fp} as a function of s. The dashed line represents the ground truth (GT) and the shaded areas the standard deviation (n = 100).

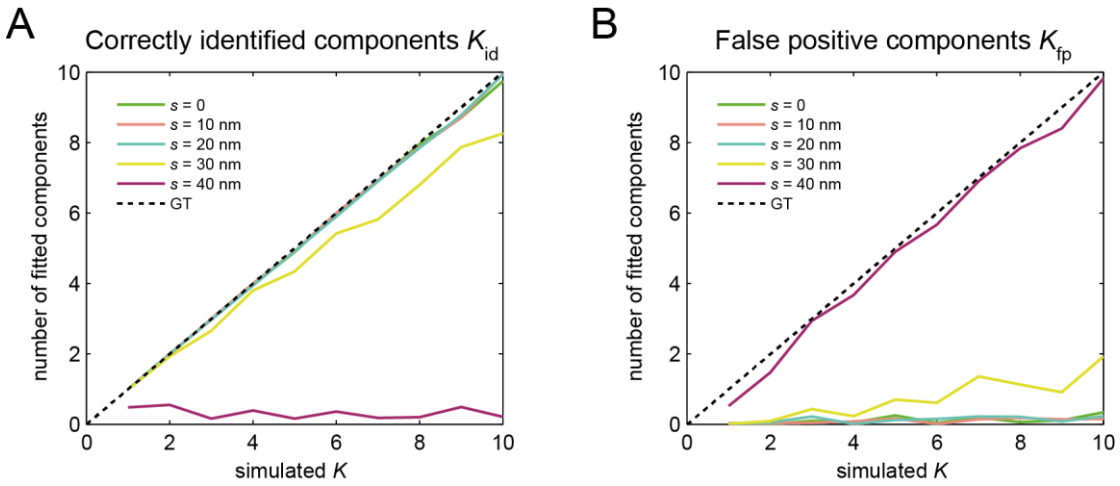

Figure S9. Influence of the localization uncertainty s and the number of mixture components K on the EMGM performance. (A) Simulated average number of correctly identified components K_{id} as a function of K, for different values of s. (B) Simulated average number of false positive components K_{fp} as a function of K, for different values of s. The dashed line represents the ground truth (GT) and n = 100 simulations were performed.

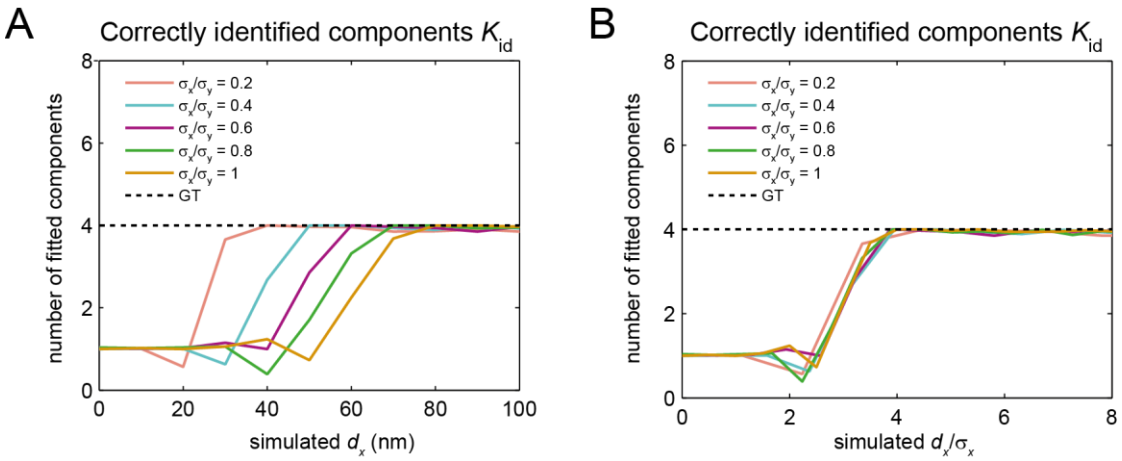

Figure S10. Influence of the eccentricity σ_x/σ_y and the spacing d_x on the EMGM performance. (A) Simulated average number of mixture components correctly identified by EMGM as a function of d_x for different values of σ_x/σ_y. (B) Simulated average number of mixture components correctly identified by EMGM as a function of d_x/σ_x. The dashed line represents the ground truth (GT) and the average values were obtained from n = 100 simulations.

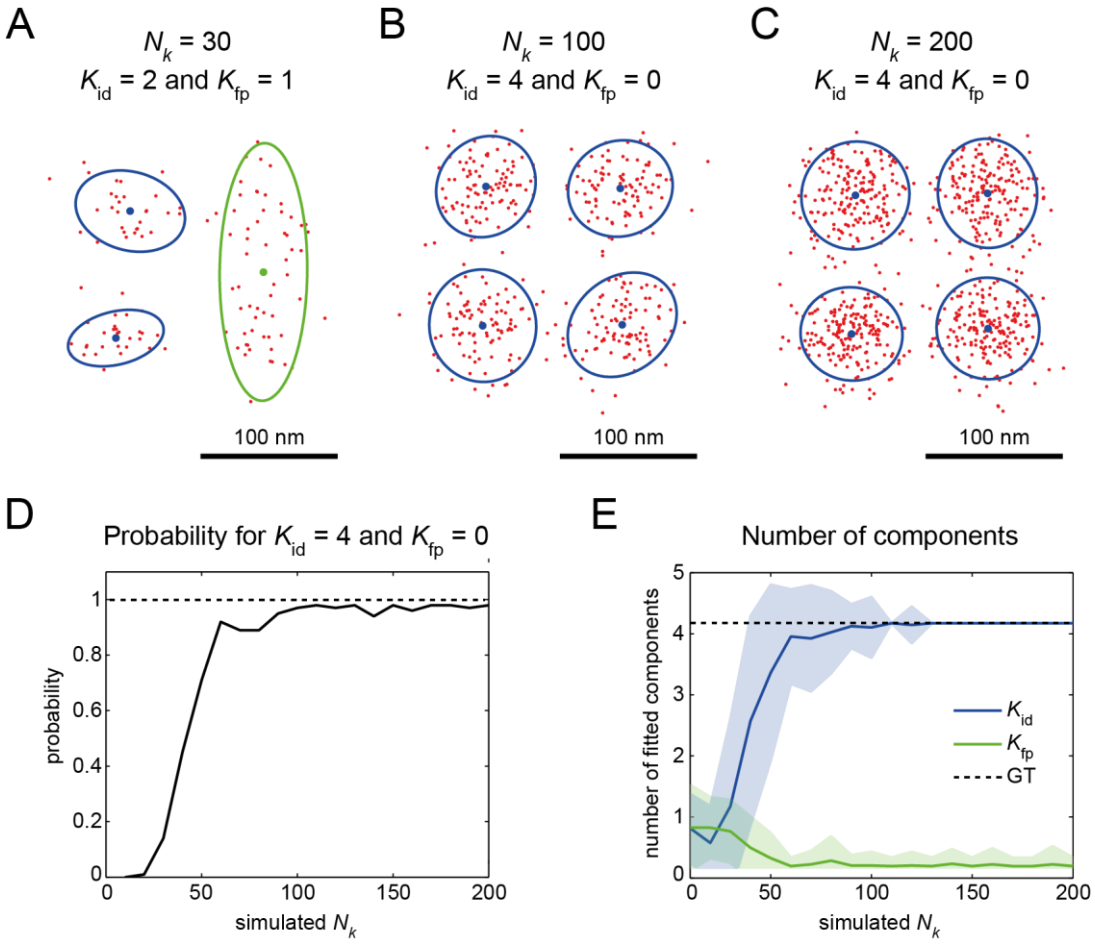

Figure S11. Influence of the number of localizations on the EMGM performance. (A-C) Example EMGM results for simulated Gaussian mixtures. Each mixture consists of K = 4 components with localization number (A) N_k = 30, (B) N_k = 100, or (C) N_k = 200. EMGM correctly identified K_{id} = 2 components and found K_{fp} = 1 false positive components for (A). EMGM correctly identified K_{id} = 4 components and found K_{fp} = 0 false positive component for (B) and (C). The red dots symbolize the simulated localizations. The blue/green dots symbolize the center positions of the correct/false positive components, the blue/green ellipses symbolize the 2σ error ellipses of the correct/false positive components. (D) The simulated probability of obtaining a completely correct EMGM result (i.e. K_{id} = 4 and K_{fp} = 0) as a function of N_k. (E) The simulated average values of K_{id} and K_{fp} as a function of N_k. The dashed line represents the ground truth (GT) and the shaded areas the standard deviation (n = 100).

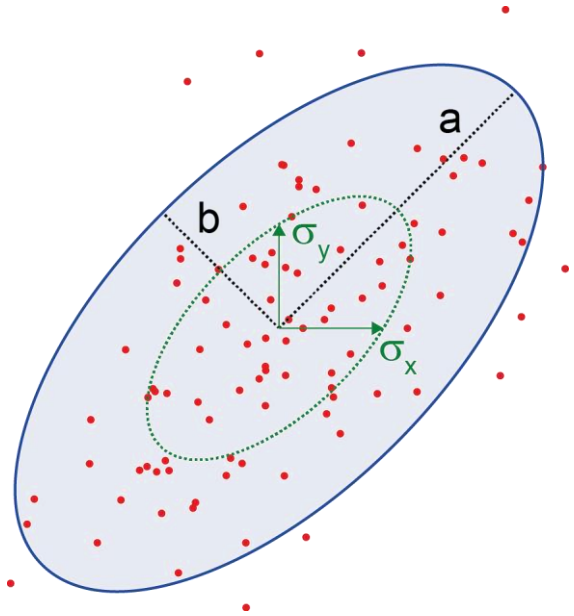

Figure S12. Illustration of a Gaussian component with standard deviation σ_x and σ_y, together with the corresponding 2σ error ellipse with major axis a and minor axis b.

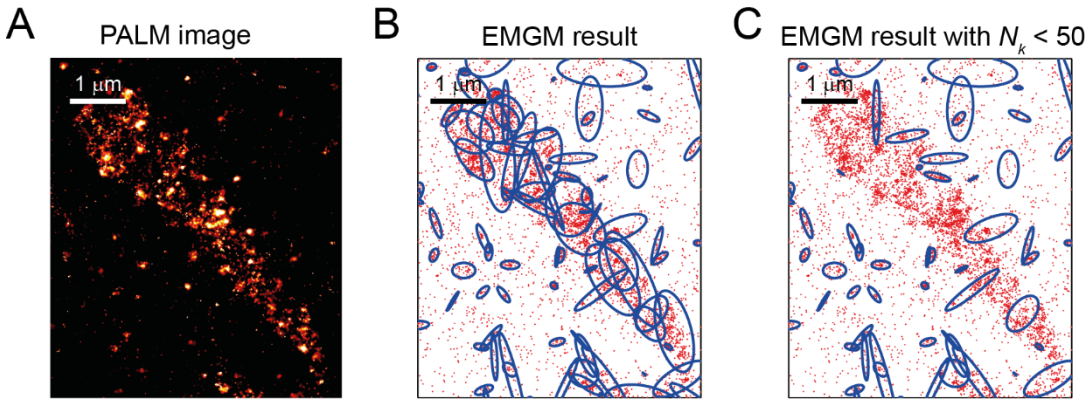

Figure S13. Focal adhesion substructures with small localization numbers N_k identified by EMGM. (A) PALM image of a small area in a fixed REF cell expressing integrin β3 labelled with mEos2, growing on a fibronectin-coated substrate (see Fig. 3B). (B) Result of the EMGM analysis of the PALM data shown in (A). The red dots symbolize the localizations, and the blue ellipses the 2σ error ellipses of the mixture components. (C) Same as (B) showing only the components with N_k < 50.

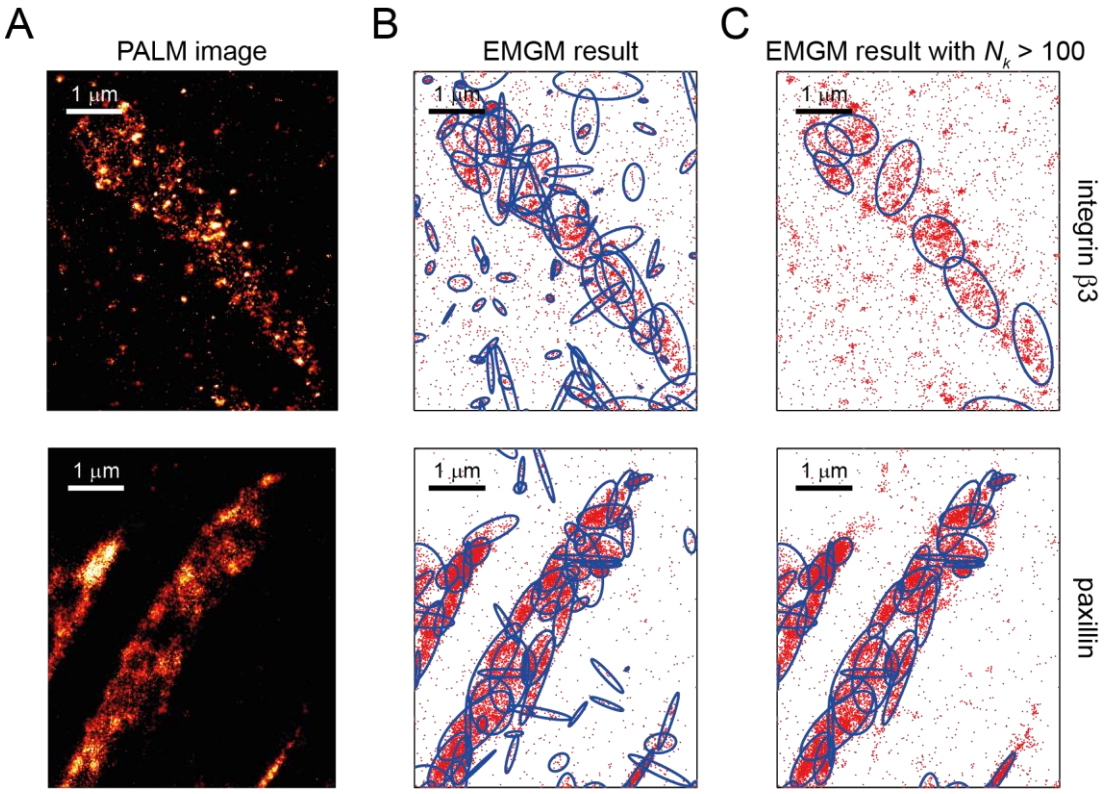

Figure S14. Focal adhesion substructures with large localization numbers N_k identified by EMGM. (A) PALM images of a small area in a fixed REF cell expressing paxillin or integrin β3 labelled with mEos2, growing on a fibronectin-coated substrate (see Fig. 3B). (B) Result of the EMGM analysis of the PALM data shown in (A). The red dots symbolize the localizations, and the blue ellipses the 2σ error ellipses of the mixture components. (C) Same as (B) showing only the components with N_k > 100.

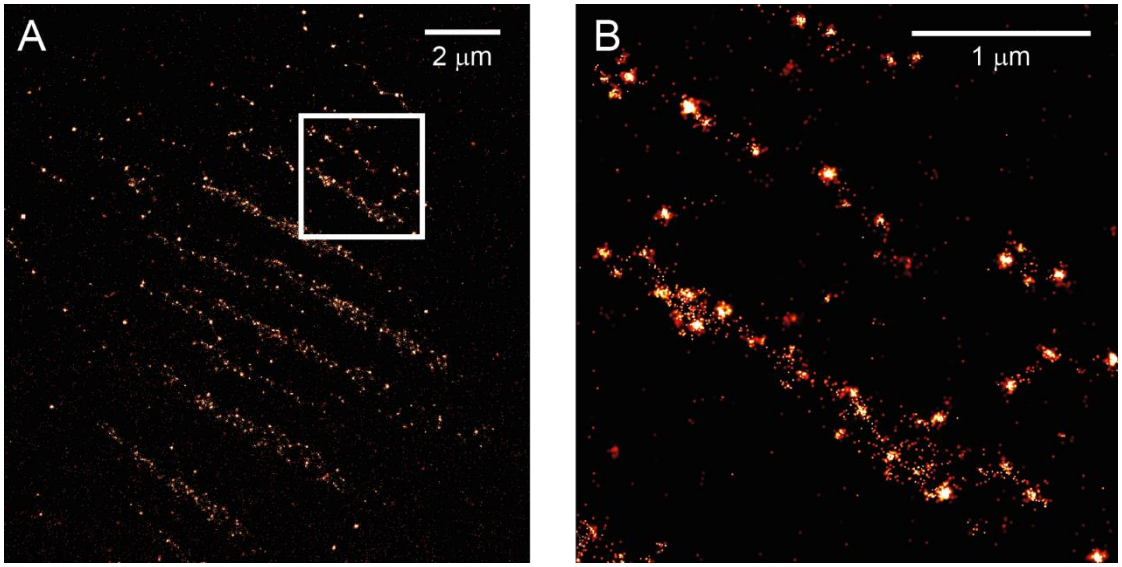

Figure S15. PAINT imaging of focal adhesions. (A) PAINT image of a fixed REF cell where integrin β3 was antibody stained. (B) Zoom-in of the region in (A) indicated by the white rectangle.

Supporting Tables

Polymer $PS_{(units)}$-*b*-$P2VP_{(units)}$	PDI	Polymer concentration [mg/ml]	Spinning speed [rpm]	Distance on glass [nm]
PS_{1056}-*b*-$P2VP_{671}$	1.09	5	2000	56 ± 9
		2.5	6000	119 ± 11

Table S1 Details concerning the block polymers and the spin casting processes used for the fabrication of the nano-patterned substrates.