Closed-form density-based framework for automatic detection of cellular morphology changes

Tarn Duong, Bruno Goud[1], and Kristine Schauer[1,2]

Unité Mixte de Recherche 144, Centre National de la Recherche Scientifique, Institut Curie, 26 rue d'Ulm, 75248 Paris Cedex 05, France

A primary method for studying cellular function is to examine cell morphology after a given manipulation. Fluorescent markers attached to proteins/intracellular structures of interest in conjunction with 3D fluorescent microscopy are frequently exploited for functional analysis. Despite the central role of morphology comparisons in cell biological approaches, few statistical tools are available that allow biological scientists without a high level of statistical training to quantify the similarity or difference of fluorescent images containing multifactorial information. We transform intracellular structures into kernels and develop a multivariate two-sample test that is nonparametric and asymptotically normal to directly and quantitatively compare cellular morphologies. The asymptotic normality bypasses the computationally intensive calculations used by the usual resampling techniques to compute the P-value. Because all parameters required for the statistical test are estimated directly from the data, it does not require any subjective decisions. Thus, we provide a black-box method for unbiased, automated comparison of cell morphology. We validate the performance of our test statistic for finite synthetic samples and experimental data. Employing our test for the comparison of the morphology of intracellular multivesicular bodies, we detect changes in their distribution after disruption of the cellular microtubule cytoskeleton with high statistical significance in fixed samples and live cell analysis. These results demonstrate that density-based comparison of multivariate image information is a powerful tool for automated detection of cell morphology changes. Moreover, the underlying mathematics of our test statistic is a general technique, which can be applied in situations where two data samples are compared.

hypothesis test | integrated density functional | optimal bandwidth selection | quantitative cell comparison

Fluorescent markers attached to proteins of interest in conjunction with modern fluorescent microscopy technologies are a useful proxy for studying subcellular compartments and their behavior after a given manipulation. Treatment with chemical compounds or specific gene silencing by RNA interference are commonly used at the scale of individual experiments to high-throughput studies. Visual inspection by expert biologists has been performed for several decades, ranging from early studies by microscopists like Ramon y Cajal to contemporary large scale, high-throughput screens (1–4). Although human observation may be very accurate, the three major drawbacks are that (*i*) it lacks quantitative measures, (*ii*) it may be biased, and (*iii*) it is time consuming.

The structural features of cells and the topological relationships between the numerous intracellular compartments give rise to multivariate data whose unbiased, automatic comparison is a major challenge. Importantly, alterations in cellular morphology also occur in many diseases, including cancer, requiring quantitative tools for their detection. Given that the majority of functional cell biology is based on image comparison, few tools are available that allow an unbiased, automatic comparison of the multivariate data encoded in fluorescent images. The cytometric tools developed so far are based on the extraction of a variety of numerical features from images in combination with classification

strategies (5–9). Features represent any measured property derived from the image, such as total/mean/standard deviation of fluorescence intensity, texture, Zernike shape descriptions, etc. (5, 10). Although feature-based approaches have proved to be very powerful in detection of morphological changes (11–14), they may suffer from lack of biologically meaningful, human interpretable measurements due to the acquisition of abstract numerical features and high-dimensional feature vector analysis. Furthermore, they require careful choice and calibration for each comparison (15). Current approaches also suffer from reduction of information as multidimensional information is transformed to one-dimensional metrics such as distances (16). There are few statistical approaches that directly assess intracellular organization, which makes automated image analysis of intracellular topology challenging. Thus, spatial comparisons could complement feature-based techniques for analyzing cell morphology alterations.

Recently, we showed that global spatial organization of defined subcellular structures (e.g., organelles, membrane domains) can be quantified by probabilistic density maps (17). We grew cells on adhesive micropatterns that enforce cells to take a certain shape, mimicking tissues' microenvironment (18). Image stacks of fluorescently marked proteins from several tens of cells were transformed into a cloud of coordinate points by segmentation analysis and were aligned using characteristic landmarks of micropatterns. To rigorously measure the topology of the fluorescently labeled subcellular structures, we centered Gaussian functions (kernels) with mean zero and an optimized variance at each of the data points and summed, revealing the underlying density throughout the cell. This analysis demonstrated that density estimation is a reliable statistical technique for the analysis of the morphology of subcellular structures whose point coordinates can be resolved, providing the basics for a comprehensive framework for statistical analysis. By transforming intracellular structures into three-dimensional kernels, alterations in cellular global organization, and thus cell morphology, can be translated into differences in density maps that are tractable by mathematical tools.

The problem of comparing two data samples has attracted much research to investigate its theoretical and practical aspects. Historically, the first methods involved small computational burdens. The well-known *t*-test developed in the Guinness brewery fits normal distributions with different means but with equal variances to each data sample, thus reducing the original problem to a simpler comparison for a difference in the means. However, this test is limited by the prespecification of the parametric form. Amongst the most widely known nonparametric tests for one-

dimensional continuous data are the Mann-Whitney, Kolmogor-ov-Smirnov, and Wald-Wolfowitz tests (19). The need for analogous tests for multivariate data has been addressed (20–22). However, these multivariate approaches have not met with the same wide acceptance as their univariate ancestors, because the former have not consistently yielded intuitive inferences when applied to experimental data. Given that the *t*-test is a density-based comparison, replacing parametric density estimates with their nonparametric counterparts should lead to a more flexible testing procedure. Kernel smoothing is a widely used computational technique for density estimation due to its intuitive construction and interpretation (23). Thus, it is an ideal basis for nonparametric density-based testing. Kernel-based tests have been developed with other discrepancy measures (24–27), but all rely on computationally intensive resampling methods to compute the critical quantiles of the null distribution. Although resampling methods provide a general framework for consistent tests, a second major trade-off is that they require sufficient familiarity, as resampling requires calibration for each data analysis situation at hand. These constraints prevent the wide adoption of bootstrap kernel density-based testing outside the computational statistical community. In particular, these tests are not easily available to biologists.

Here, we develop a test statistic that is asymptotically normal under the null hypothesis, allowing density-based, "black-box" comparisons of multivariate data. We use simulated and experimental data analysis to verify its performance for finite samples. Given that 3D organizations of cells can be expressed by probabilistic density maps, this test allows us to assess the statistical significance of the similarity or difference between two cellular topologies. Analyzing the data from fluorescent images of intracellular organelles, this test allows us to compare cellular morphology under different conditions in an automated and unbiased manner.

Results

Construction of the Test Statistic. We have used the usual squared discrepancy measure in order to construct a nonparametric and multivariate test statistic $\hat{T}$ that is asymptotically normal under the null hypothesis. (Algorithmic details are deferred to the *Methods* section).

Let $X_1, X_2, ... X_{n1}$ and $Y_1, Y_1, Y_2, ... Y_{n2}$ be *d*-variate random samples from their respective common densities f_1 and f_2. Concretely, $X_1, X_2, ... X_{n1}$ are the spatial coordinates of subcellular structures extracted from a first group of images, and likewise for $Y_1, Y_1, Y_2, ... Y_{n2}$ from a second group of images. So f_1 represents the steady-state spatial probability density function of the subcellular structures in the first images, and likewise for f_2. This is the same statistical framework used in Schauer et al. (17) to construct density maps from a single set of images. The kernel density estimates of f_1 and f_2 are

$$\hat{f}_1(\boldsymbol{x}; \mathbf{H}_1) = \frac{1}{n_1} \sum_{i=1}^{n_1} K_{\mathbf{H}_1}(\boldsymbol{x} - \boldsymbol{X}_i),$$
$$\hat{f}_2(\boldsymbol{x}; \mathbf{H}_2) = \frac{1}{n_2} \sum_{j=1}^{n_2} K_{\mathbf{H}_2}(\boldsymbol{x} - \boldsymbol{Y}_j) \qquad [1]$$

where K is the kernel function with $K_{\mathbf{H}_l}(\boldsymbol{x}) = |\mathbf{H}_l|^{-1/2} K(\mathbf{H}_l^{-1/2}\boldsymbol{x})$, and $\mathbf{H}_l$ is a bandwidth matrix, for $l = 1, 2$.

To test the null hypothesis $H_0 : f_1 = f_2$, we follow Anderson et al. (28), who proposed the following discrepancy measure $T = \int [f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})]^2 d\boldsymbol{x}$. As is the case in the rest of this manuscript whenever the limits of integration are omitted, integration is taken over the appropriate Euclidean space. We use the squared error measure, since it has the most extensive body of work in automatic optimal selection of the smoothing parameters in comparison to other discrepancy measures such as the absolute

error, Kullback-Leibler error, and Shannon-Jenson error. We rewrite the discrepancy as $T = \psi_1 + \psi_2 - (\psi_{1,2} + \psi_{2,1})$ where $\psi_l = \int f_l(\boldsymbol{x})^2 d\boldsymbol{x}$ and $\psi_{l_1,l_2} = \int f_{l_1}(\boldsymbol{x}) f_{l_2}(\boldsymbol{x}) d\boldsymbol{x}$. The test statistic is $\hat{T} = \hat{\psi}_1 + \hat{\psi}_2 - (\hat{\psi}_{1,2} + \hat{\psi}_{2,1})$ where

$$\hat{\psi}_1 = \frac{1}{n_1^2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_1} K_{\mathbf{H}_1}(\boldsymbol{X}_{i_1} - \boldsymbol{X}_{i_2}),$$

$$\hat{\psi}_2 = \frac{1}{n_2^2} \sum_{j_1=1}^{n_2} \sum_{j_2=1}^{n_2} K_{\mathbf{H}_2}(\boldsymbol{Y}_{j_1} - \boldsymbol{Y}_{j_2})$$

$$\hat{\psi}_{1,2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K_{\mathbf{H}_1}(\boldsymbol{X}_i - \boldsymbol{Y}_j),$$

$$\hat{\psi}_{2,1} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K_{\mathbf{H}_2}(\boldsymbol{X}_i - \boldsymbol{Y}_j).$$

We can interpret this test statistic as the comparing intrasample pairwise differences $\boldsymbol{X}_{i_1} - \boldsymbol{X}_{i_2}$ and $\boldsymbol{Y}_{j_1} - \boldsymbol{Y}_{j_2}$ to the intersample pairwise differences $\boldsymbol{X}_i - \boldsymbol{Y}_j$. So if the latter are larger than the former, then this indicates that the samples are different. The following theorem is our main result, which establishes the asymptotic normality under the null hypothesis of the test statistic $\hat{T}$.

Theorem 1. *Under the conditions in the Methods section, and assuming that the null hypothesis holds, $H_0 : f_1 = f_2 = f$. As $n_1, n_2 \to \infty$, then* $\frac{\hat{T} - \mu_T}{\sigma_T \sqrt{1/n_1 + 1/n_2}} \xrightarrow{d} N(0, 1)$, *where* $\mu_T = [n_1^{-1}|\mathbf{H}_1|^{-1/2} + n_2^{-1}|\mathbf{H}_2|^{-1/2}] K(0)$ *and* $\sigma_T^2 = 3[\int f(\boldsymbol{x})^3 d\boldsymbol{x} - (\int f(\boldsymbol{x})^2 d\boldsymbol{x})^2]$.

Null Distribution Parameter Estimation. To use the asymptotic null distribution, we need to estimate the mean parameter μ_T and the variance parameter σ_T^2. For μ_T, Chacón and Duong (29) showed an algorithm to obtain consistent estimators of the bandwidth matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ as minimizers of the asymptotic Mean Squared Error of $\hat{\psi}_1$ and $\hat{\psi}_2$ respectively. For σ_T^2, it is straightforward to show that an estimator is $\hat{\sigma}_T^2 = (n_1 \hat{\sigma}_1^2 + n_2 \hat{\sigma}_2^2)/(n_1 + n_2)$ where $\hat{\sigma}_1^2$ is an estimator of σ_1^2 the variance of $f_1(\boldsymbol{X})$ and $\hat{\sigma}_2^2$ an estimator of σ_2^2 the variance of $f_2(\boldsymbol{Y})$. Previous research has indicated that asymptotic normal approximations of a null distribution tend to reject the null hypothesis more often than is indicated by the nominal level of significance (25). One of the primary causes is the overestimation of the variance. In the context of kernel estimators, this usually arises from using a bandwidth which is optimal for density estimation, but which leads to an inflated variance estimate. Our proposed solution is to estimate the variance more directly using a larger bandwidth, since larger bandwidths reduce the variance by mitigating the effect that individual data points have on the value of the kernel estimator. Examining the first order Taylor's series expansion about the expected value: $f(\boldsymbol{X}) \sim f(E\boldsymbol{X}) + (\boldsymbol{X} - E\boldsymbol{X})^T Df(E\boldsymbol{X})$ where $Df = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_d}]^T$ is the column vector of first partial order derivatives, thus $\text{Var} f(\boldsymbol{X}) \sim [Df(E\boldsymbol{X})]^T [\text{Var} \boldsymbol{X}][Df(E\boldsymbol{X})]$. So plug-in estimators of σ_1^2 and σ_2^2 are $\hat{\sigma}_1^2 = [D\hat{f}_1(\bar{\boldsymbol{X}}; \boldsymbol{G}_1)]^T \boldsymbol{S}_1 [D\hat{f}_1(\bar{\boldsymbol{X}}; \boldsymbol{G}_1)]$ and $\hat{\sigma}_2^2 = [D\hat{f}_2(\bar{\boldsymbol{Y}}; \boldsymbol{G}_2)]^T \boldsymbol{S}_2 [D\hat{f}_2(\bar{\boldsymbol{Y}}; \boldsymbol{G}_2)]$ where $D\hat{f}_1(\boldsymbol{x}; \hat{\boldsymbol{G}}_1) = n_1^{-1} \sum_{i=1}^{n_1} DK_{\hat{\boldsymbol{G}}_1}(\boldsymbol{x} - \boldsymbol{X}_i)$, $D\hat{f}_2(\boldsymbol{x}; \hat{\boldsymbol{G}}_2) = n_2^{-1} \sum_{j=1}^{n_2} DK_{\hat{\boldsymbol{G}}_2}(\boldsymbol{x} - \boldsymbol{Y}_j)$, $\bar{\boldsymbol{X}}, \bar{\boldsymbol{Y}}$ are the sample means, $\boldsymbol{S}_l$ are the sample variances, and $\hat{\boldsymbol{G}}_l = [4/(d + 4)]^{2/(d+6)} \boldsymbol{S}_l n_l^{-2/(d+6)}$ are the normal scale selectors for a kernel estimator of the first density derivative (30).

Given these parameter estimates, the standard equation to obtain a *z*-score from $\hat{T}$ is $Z = \frac{\hat{T} - \hat{\mu}_T}{\hat{\sigma}_T \sqrt{1/n_1 + 1/n_2}}$. The *p*-value is then computed from this *z*-score using standard software or tables. The completely automatic testing procedure (including the para-

APPLIED
MATHEMATICS

CELL BIOLOGY

meter estimation, and the computation of the test statistic and its P-value) is programmed in the ks library (31) in the open-source R programming language.

Simulated Data Analysis. To verify the performance of our kernel density-based test for finite samples, we performed simulation studies using pairs of mixture normal densities, mostly taken from Chacón (32). The contour plots of these test densities as well as representative scatter plots for the two considered sample sizes ($n = 100$ and $n = 1000$) are displayed in Fig. 1. The first pair $N((-1/2, 0), \mathbf{I}_2)$ and $N((1/2, 0); \mathbf{I}_2)$ represent two single normal densities with identity variance, whose means are separated by distance of 1. This example was treated as base case. The second pair both are bimodal densities, $1/2N((1, -1), \Sigma) + 1/2N((-1, 1), \Sigma)$ and $1/2N((1, -1), \Sigma) + 1/2N((1, -1), \mathbf{I}_2)$ where $\Sigma = [4/9 \; 4/15; 4/15 \; 4/9]$. The lower right component of the pairs was exactly the same, but their upper right component was different, making it potentially a challenging case to distinguish between two finite samples. As a third example, we chose a pair 3, $N((0, 0), \mathbf{I}_2)$ and $1/2N((0, 0), \mathbf{I}_2) + 1/10N((0, 0), 1/16 \, \mathbf{I}_2) + 1/10N((-1, -1), 1/16 \, \mathbf{I}_2) + 1/10N((-1, 1), 1/16 \, \mathbf{I}_2) + 1/10N((1, -1), 1/16 \, \mathbf{I}_2) + 1/10N((1, 1), 1/16 \, \mathbf{I}_2)$, that have (approximately) zero mean and identity variance. Because this pair reveals different internal structure, it would most likely benefit from a density-based, rather than a moment-based, test.

First, we verified the asymptotic normality of $\hat{T}$ by comparing the density estimates of the z-scores with the standard normal (Fig. S1). The larger sample gave better estimates of the zero mean. On the other hand, performance in variance estimation was more uneven, since the $n = 1000$ samples did not lead to better variance estimates for pair 2. Related results have been observed previously (25), indicating that the variance estimation is the most difficult part in calibrating an asymptotically normal null distribution.

We performed simulations of the test statistic for two common nominal levels of significance $\alpha = 0.05, 0.01$ (Table 1), where α is the error rate of rejecting the null hypothesis H_0 when the null hypothesis is true (false positive). To estimate how close our statistical test in achieving this error rate, we computed the proportions of experiments ($\hat{\alpha}$) where two samples are simulated from the same distribution, which reject H_0. Given a level of significance, the other error that can be made is to accept the null hypothesis H_0 when it is false (false negative). We estimated this by computing the proportion of the experiments ($\hat{\beta}$) where two samples are simulated from different distributions, which H_0 is accepted. The empirical power is $1 - \hat{\beta}$. For the smaller sample size, we found that the empirical significance levels were close to

Table 1. Simulation results

		Nominal $\alpha = 0.05$		Nominal $\alpha = 0.01$		Z	
		$\hat{\alpha}$	$1 - \hat{\beta}$	$\hat{\alpha}$	$1 - \hat{\beta}$	mean	SD
$n = 100$	Pair 1	0.010	0.914	0.002	0.830	−0.226	0.781
	Pair 2	0.018	0.052	0.004	0.026	−0.285	0.919
	Pair 3	0.038	0.446	0.010	0.264	−0.160	1.019
$n = 1000$	Pair 1	0.018	1.000	0.006	1.000	−0.1379	0.760
	Pair 2	0.020	0.946	0.006	0.810	−0.0625	0.758
	Pair 3	0.074	1.000	0.026	1.000	−0.0001	1.191

Comparison of empirical level of significance ($\hat{\alpha}$) to the nominal level (α), empirical power ($1 - \hat{\beta}$) and central moments of the null distribution of the normalized $\hat{T}$ for $n = 100$; $n = 1000$.

the nominal values, but the power was low for pairs 2 and 3. This indicated that $n = 100$ data points were not sufficient to distinguish reliably between these more difficult comparisons. For the larger sample, the lack of power was resolved for all three pairs. The empirical levels of significance were more conservative for the larger sample size. This simulation evidence demonstrated that our proposed test does not identify more false positives than expected from the nominal level of significance and identifies almost all true negatives.

To evaluate the performance of our test, we compared it to a parametric alternative. The t-test is a well-known hypothesis test for univariate data, which has been generalized for multivariate data in Nel and Van der Merwe (33). The average P-values from 100 simulations of sample size 1000 were computed for each of the three pairs target densities (Table S1). As expected, the modified Nel and Van der Merwe (MNV) was more sensitive for the first pair, which only differed in mean (average P-value $= 0$). However, the kernel test was still highly significant (average P-value $= 1.142 * 10^{-29}$). For the next two pairs with similar mean values but clear differences in the internal organizations between the two densities, the MNV test gave nonsignificant average P-values of 0.5195 and 0.2158, whereas the kernel-based test gave highly significant average P-values of $1.353 * 10^{-8}$ and $3.386 * 10^{-23}$. This demonstrated that our density-based test outperformed the parametric MNV test in the detection of differences in internal organization.

Detection of Morphological Changes in Micropatterned Cells after Drug Treatment. To evaluate how our test is performing on experimental data, we compared the morphologies of intracellular structures under different experimental conditions (Fig. 2). As indicated in the flowchart (Fig. 2A), we analyzed the morphology changes of multivesicular bodies (MVB) induced by a treatment with the drug nocodazole (NZ) that depolymerizes microtubules, a major component of the cellular cytoskeleton. MVB are endosomes involved in several important cellular functions, including processing of nutrients, ligands and receptors during endocytosis, exosome secretion, and autophagy (34) that are transported along microtubules (35). Intracellular MVB were visualized by indirect immunofluorescence against CD63, a transmembrane protein enriched in MVB (Fig. 2B). Cells were cultured on micropatterns of extracellular matrix proteins that standardize cell shape and allow alignment of CD63-marked structures. Combining the signals of CD63-marked components from several tens of cells, we showed that the 3D organization of MVB is reproducible in these normalized conditions (17). Disruption of microtubules with NZ disconnects MVB from microtubules, leading to subtle changes in cell morphology (Fig. 2C). We transformed the fluorescent signal of normalized cells into coordinates by segmentation analysis as previously reported (17). All detected signals from a control group of cells and a group of cells exposed to treatment conditions were combined to the test populations f_1 and f_2, respectively. In previous analysis, we estimated that pooling signals from about 20–30 normalized cells (containing several hundreds of structures each)
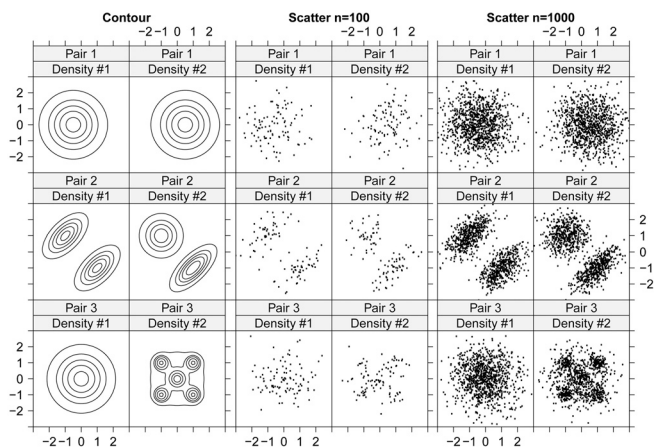

Fig. 1. Contour plots of simulated data. Contour plots for three pairs of simulated normal mixture densities (*Left*). Representative scatter plots for $n = 100$ (*Middle*). Representative scatter plots for $n = 1000$ (*Right*).
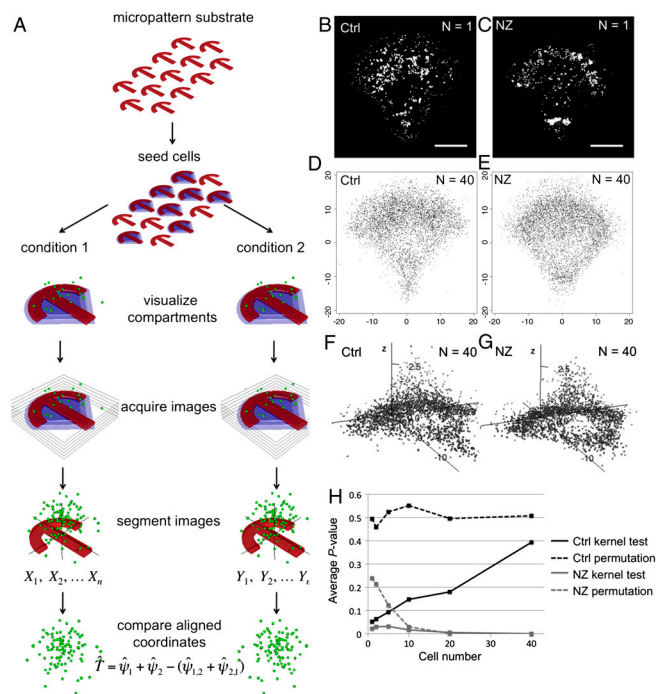
Fig. 2. Changes of CD63 morphology upon disruption of the microtubule cytoskeleton. (*A*) Flow chart of morphology comparison between two populations of cells seeded on micropatterned substrates. (*B*), (*C*) Maximum intensity projection of the deconvolved fluorescence of CD63-marked multivesicular bodies (MVB) without treatment (Ctrl) (*B*) and upon nocodazole (NZ) treatment (*C*). Scale bars are 10 μm. (*D*), (*E*) Representative 2D scatter plots for 40 cells with $n = 11786$ structures for Ctrl (*D*) and for 40 cells with $n = 13615$ structures for NZ (*E*). (*F*), (*G*) Representative 3D scatter plots for Ctrl (*F*) and NZ treatment (*G*) The z-axis is zoomed 500%. (*H*) Average *P*-values from kernel test (solid lines) and permutation analysis (dashed lines) as a function of the number of cells analyzed for 100 comparisons. Average *P*-values for test statistics for Ctrl (black) and NZ treatments (gray).

was required to produce reliable density maps (17). So we took these cell numbers as a starting point for our comparisons. Representative 2D and 3D scatter plots of 40 cells are shown in Fig. 2*D*–*G*; the 2D scatter plots of individual cells are represented in Fig. S2. The coordinates from 40 cells from each condition were compared by $\hat{T}$.

First, we compared a nontreated control group 1 of 40 cells with 11786 detected structures with a second control group 2 of 40 cells containing 12585 structures. The two control samples gave slightly different estimates of the CD63 steady-state distribution (Fig. S3). We estimated the normalized $\hat{T}_{2D} = 0.6492$ and $\hat{T}_{3D} = 1.2066$ corresponding to P_{2D}-value of 0.2581 and P_{3D}-value = 0.1138. So there is strong evidence that the minor differences between the two control samples are not significant.

We then compared the control group 1 with the NZ treated group of 40 cells with 13615 detected structures. For the control versus treatment condition, we estimated the normalized $\hat{T}_{2D} = 4.160$ and $\hat{T}_{3D} = 6.670$ giving rise to P_{2D}-value of $1.589 * 10^{-5}$ and P_{3D}-value = $1.280 * 10^{-11}$, indicating that there is strong evidence that the drug treatment significantly affects the distribution of MVB. These results agree with previous studies demonstrating microtubule-dependent movement of MVB (35). To further evaluate our approach, we performed additional analysis of diverse subcellular structures (SI Text, Fig. S4, and Table S2).

Second, we compared how our test is performing in comparison to a resampling strategy that was previously established for the comparison of fluorescent images (17). We calculated average *P*-values from either our test statistics or the permutation analysis as a function of the number of cells analyzed, taking 100 random samples of 1, 2, 10, 20, and 40 cells (Fig. 2*H*, Table S3). First, we

randomly picked two subsamples from the control conditions (Ctrl) to estimate the false positive rate of our test. Then we compared one control subsample with one subsample taken from treated condition (NZ). Dashed lines represent the permutation test; solid lines represent the kernel density test (Fig. 2*H*). According to the fundamental *P*-value calculations, *P*-values follow a uniform distribution on [0, 1] and thus mean 0.5, assuming the null hypothesis holds as expected for Ctrl. This is true for the permutation test, since it can mimic the sampling distribution of the test statistic. The kernel density test gives more small *P*-values (false positives) than predicted due to the asymptotic approximation (see also Fig. S5). However, an average *P*-value > 0.05 was obtained in each case, not rejecting the null hypothesis at the usual significance levels. Applied to the comparison between Ctrl and NZ treatment, the kernel density test gives lower *P*-values (more true positives) that the permutation test for <10 cells. Together, these two tests gave the same conclusions when testing a treatment for more than 10 cells (as typically analyzed), demonstrating that the normal approximation for the sampling distribution was as accurate as bootstrap resampling in this case. Thus, our test statistic is comparable to bootstrap resampling.

Next, we evaluated how sensitive our method is to errors in cell alignment on patterned substrates. We systematically estimated how strongly our test statistic degrades as a function of rotational and translational misalignment (see SI Text and Table S4). Overall, as expected, the *P*-values uniformly decrease as the magnitude of misalignment increases. As expected from the simulated data analysis, our test was highly sensitive when the entire cell sample to be compared was misaligned. Both rotations of as little as 10° and translations of 20–30 pixels were sufficient to give significant *P*-values ($P < 0.05$) (pattern size was 550 pixels). Our test was, however, less sensitive to random misalignment in individual cells within one sample. Significant *P*-values for rotations appeared at 30–40° and at 30–50 pixels for translations. This analysis highlights the importance of cell alignment to reduce false positive results. Together, our analysis on CD63 demonstrated that the density-based comparison is well suited to detect changes in steady-state morphology of cells cultured under controlled conditions of adhesion.

Detection of Morphological Changes in Live-Cell Assays in Unconstrained Cells. To demonstrate that our density-based framework is also valid for the detection of morphological changes in unconstrained cells that are classically studied, we applied our test to live cell analysis. Because cells maintained a consistent orientation during a given time period, fast changes in intracellular organization as observed after drug treatments could be analyzed by our density-based method. We comprehensively benchmarked the statistical method on the dynamics of MVB in unconstrained cells before and after treatment with NZ (Fig. 3). We acquired 3D stacks over 24 min with acquisition at each 60 s, extracted 3D positional information of labeled compartments by segmentation, and compared morphology changes.

We split the images of Movie S1 into four groups (1–4) containing six images each (Fig. 3*A*). Groups 1 and 2 represented the nontreated control groups, with 1080 and 1002 detected CD63-positive structures that were acquired before addition of the drug. Groups 3 and 4 were the treatment test groups, containing 1019 and 801 structures that were recorded after the addition of the drug. The corresponding 2D and 3D scatter plots of the four groups are shown in Fig. 3*B* and *C*. We applied our test statistic on each of the possible combination of pairs. The corresponding *P*-values for the 2D and 3D comparison are listed in Table S5. The results indicated that whereas no significant changes in CD63 morphology was detected before drug treatment (P_{2D}-value = 0.4136; P_{3D}-value = 0.3565), the treatment with NZ significantly affected CD63-morphology (P_{2D}-value = $3.998 * 10^{-6}$; P_{3D}-value = $4.844 * 10^{-6}$). The effect of the drug

was more significant for later time points in agreement with visual inspection of the images, demonstrating that the statistical significance can quantify compound influence. Thus, our approach allowed unbiased automated detection of morphological changes in live-cell assays in unconstrained cells.

Discussion

We have developed a test statistic which inherits the advantages of kernel density estimates to facilitate generally applicable two-sample comparisons of multivariate data. By drawing on recent distributional results for kernel estimators, we were able to express its null distribution in a closed asymptotic form, thus circumventing the requirement for resampling to determine the critical quantiles of the null distribution.

This test allowed us to compare complex data from fluorescent microscopy without reducing the provided information into simple summary statistics. This allowed quantitative comparison of cellular morphology by directly measuring the three-dimensional organization of intracellular structures visualized by fluorescent microscopy. Note that our proposed image comparison focuses on the spatial localization of structures whose point coordinates can be resolved. It is also applicable to the comparison of continuous structures, such as microtubules, but performance is not optimal. As our test statistic requires independent data points (like most kernel-based estimators), the representation of continuous structures by several connected coordinates leads to smaller *P*-values. This is a disadvantage compared to commonly used feature-based techniques that collate any type of measured quantity from the microscopy images and therefore can also be applied even to diffuse fluorescent patterns. However, feature-based comparisons require the critical stage of feature selection to be calibrated carefully for each comparison (15), especially in order to compute *P*-values. So the challenge to automatize optimal feature selection into a black-box method remains an open problem. Our simpler, more direct approach of comparing spatial distributions does not face equivalent problems allowing full

automatization. As image acquisition facilities are developed at an accelerated speed, there is a rising need for image analysis tools that estimate the required test parameters directly from the data and do not require computationally intensive analytical techniques. As feature-based and density-based approaches use different numerical information, they are complementary to each other and the combination of both of them should lead to an improvement of image analysis.

Such computational imaging methods are indispensable tools for the high-content and high-throughput image acquisition capability of advanced microscopes that daily acquire thousands of high-resolution images in time-lapse experiments. We have shown that our density-based mathematical framework is powerful for phenotype profiling and can be easily adjusted to high-throughput analysis. Attempts are underway to incorporate this computational imaging comparison into a high-throughput workflow to screen for cell morphological changes due to chemical compounds treatment and siRNA-based gene silencing.

A second disadvantage of our approach is that it requires the cells to have a constant shape in order to construct spatial density maps and the test statistic. Fortunately, the micropatterning technique allows us to grow cells reproducibly into standardized shapes in culture. One important advantage of growing cells in controlled conditions of adhesion is that cells are much closer to their physiological state in tissues, where cells are restrained, than in classical (unconstrained) culture conditions on Petri dishes. Another advantage is that standardizing cells by micropatterning technology represents an important step towards quantitative approaches in cell biology. We have also shown that our testing procedure can be applied to live cell comparisons. By orienting an unconstrained cell through time, we validate that unconstrained cells are in principle analyzable. However, alignment of unconstrained cells with the help of computational approaches (36, 37) will be a requisite in order to apply density-based comparison as a general approach for unconstrained cells. An important future application is to compare cell spatial morphology in tissues, in which many cell types show reproducible shapes and inherent polarization. In particular, this application would be important to detect alterations in the cellular architecture during pathological processes such as cancer. Since imaging and alignment of tissue cells is more challenging, it is not yet suitable to apply our testing procedure.

A promising extension of our test is to elaborate the regions of the sample space, which are the largest contributors to the overall statistical difference. There have been some attempts to tackle this problem (38) using a heuristic density differences approach based on data mining approaches, but which is unable to make rigorous statistical inferences. Developing a rigorous analogue would be an important advance for the analysis of the multivariate samples comparisons. Another future challenge is analyzing structures with a diffuse fluorescence. An alternative is to consider diffuse fluorescence patterns as functional data; i.e., not to be finite dimensional vectors (as in multivariate analysis) but infinite dimensional functions. The state of the art in formal hypothesis testing for functional data analysis is less advanced than in multivariate data analysis, leaving a testing procedure for functional data analogous to our proposed statistic an open problem.
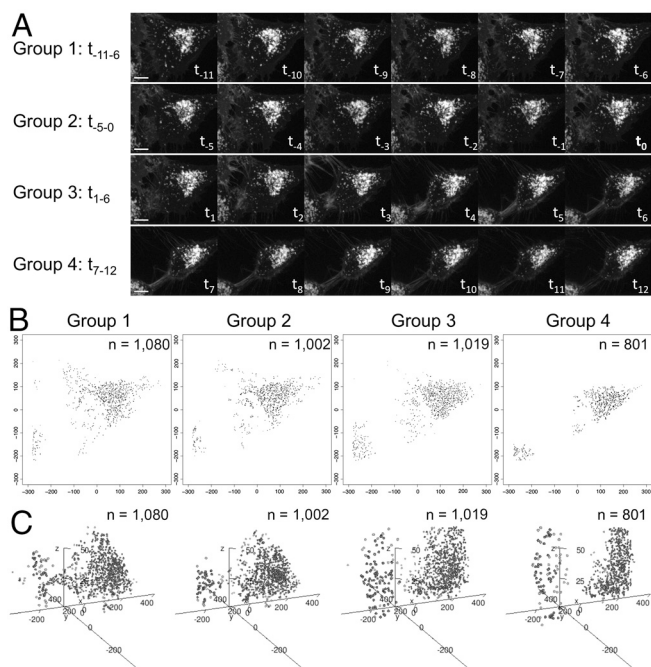

Fig. 3. Time lapse of CD63 morphology change upon microtubule disruption. (*A*) Maximum intensity projections of EGFP-CD63-marked multivesicular bodies (MVB) upon nocodazole (NZ) treatment. Each group represents six time frames with 1-min intervals: Group 1 and Group 2 show CD63 before addition of NZ; Group 3 and Group 4 show CD63 after the addition of NZ. Scale bars are 10 μm. (*B*), (*C*) Representative 2D (*B*) and 3D (*C*) scatter plots for each group with *n* structures analyzed.

Methods

Proof of Theorem 1: To establish the asymptotic sampling distribution of $\hat{T}$, we follow the approach of Chacón and Duong (29). Suppose that the conditions hold. For $l = 1, 2$,

(F) The target densities f_l have two derivatives, which are bounded, continuous, and square integrable.

(H) The bandwidths $\mathbf{H}_l = \mathbf{H}_l(n_l)$ are a sequence of symmetric, positive definite matrices such that all elements of $\mathbf{H}_l \to 0$ and $n_l^{-1}|\mathbf{H}_l|^{-1/2} \to 0$ as $n_l \to \infty$.

(K) The kernel K is a symmetric probability density function such that $m_0(K^2) = \int K(\mathbf{x})^2 d\mathbf{x}$ is finite, and that $\int \mathbf{x}\mathbf{x}^T K(\mathbf{x})d\mathbf{x} = m_2(K)\mathbf{I}_d$ for some real number $m_2(K)$ and $\mathbf{I}_d$ is the $d \times d$ identity matrix.

(N) The sample sizes n_1, n_2 are such that n_1/n_2 and n_2/n_1 are bounded away from zero and infinity as n_1, $n_2 \to \infty$. The proof is deferred to the *SI Text*.

Cells and Sample Preparation. Cell culture and sample preparation for fixed cells was as in ref. 17. Antibodies used were primary α-CD63 (Invitrogen), Sec13 (17), α-tubulin (BD Biosciences) and Alexa-Fluor 488, Cy-3, or Cy-5-coupled secondary antibodies. EGFP-CD63-expressing stable cells (generated by transfection of the plasmid pEGFP-CD63 (34) into RPE-1 cells) were seeded on iwaki glass base dishes (Asahi Glass) for live cell observation. To depolymerize microtubules, NZ was added to a final concentration of 20 μM. Cells were imaged before and after addition of NZ.

Immunofluorescence Image Acquisition and Processing. Image acquisition of fixed cells was as in ref. 17. Live cell imaging was performed on a Yokogawa spinning disc mounted on an Eclipse TE2000 Inverted Microscope using 60x Plan Apo VC 1.4 Oil objective, Laser 491 nm and CCD camera (Roper CoolSnap HQ2). Z-series of images were taken every 0.2 μm every 60 s.

Images were segmented with MetaMorph (Universal Imaging Corporation) as described in ref. 17. Briefly, the centroids of fluorescent objects were detected as fluctuations that are 15-fold larger than noise. The watershed function was routinely applied to precisely detect individual structures in dense regions. The coordinates of the segmented structures were aligned using the center of the micropatterns as in ref. 17 and used in a completely automatic testing procedure programmed in the ks library (31) in the open-source R programming language.

1. Farhan H, et al. (2010) MAPK signaling to the early secretory pathway revealed by kinase/phosphatase functional screening. *J Cell Biol* 189:997–1011.
2. Kiger AA, et al. (2003) A functional genomic analysis of cell morphology using RNA interference. *J Biol* 2:27.
3. Kim JK, et al. (2005) Functional genomic analysis of RNA interference in *C. elegans*. *Science* 308:1164–1167.
4. Ramon y Cajal S (1960) *Studies on Vertebrate Neurogenesis* (CC Thomas, Springfield, IL).
5. Boland MV, Murphy RF (2001) A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics* 17:1213–1223.
6. Carpenter AE, et al. (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol* 7:R100.
7. Jones TR, et al. (2009) Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. *Proc Natl Acad Sci USA* 106:1826–1831.
8. Lamprecht MR, Sabatini DM, Carpenter AE (2007) CellProfiler: free, versatile software for automated biological image analysis. *Biotechniques* 42:71–75.
9. Loo LH, et al. (2009) An approach for extensibly profiling the molecular states of cellular subpopulations. *Nat Methods* 6:759–765.
10. Ljosa V, Carpenter AE (2009) Introduction to the quantitative analysis of two-dimensional fluorescence microscopy images for cell-based screening. *PLoS Comput Biol* 5: e1000603.
11. Chen SC, Zhao T, Gordon GJ, Murphy RF (2007) Automated image analysis of protein localization in budding yeast. *Bioinformatics* 23:i66–i71.
12. Moffat J, et al. (2006) A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* 124:1283–1298.
13. Perlman ZE, et al. (2004) Multidimensional drug profiling by automated microscopy. *Science* 306:1194–1198.
14. Singh DK, et al. (2010) Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. *Mol Syst Biol* 6:369–379.
15. Logan DJ, Carpenter AE (2010) Screening cellular feature measurements for image-based assay development. *J Biomol Screen* 15:840–846.
16. Andrey P, et al. (2010) Statistical analysis of 3D images detects regular spatial distributions of centromeres and chromocenters in animal and plant nuclei. *PLoS Comput Biol* 6:e1000853.
17. Schauer K, et al. (2010) Probabilistic density maps to study global endomembrane organization. *Nat Methods* 7:560–566.
18. Thery M, et al. (2006) Anisotropy of cell adhesive microenvironment governs cell internal organization and orientation of polarity. *Proc Natl Acad Sci USA* 103:19771–19776.
19. Gibbons JD, Chakraborti S (2003) Nonparametric statistical inference. *Nonparametric Statistical Inference* (Marcel Dekker, New York).
20. Bickel PJ (1969) A distribution free version of the Smirnov two-sample test in the p-variate case. *Ann Math Stat* 40:1–23.
21. Friedman JH, Rafsky LC (1979) Multivariate generalizations of the Wald-Wolfowitz and Smirnov 2-sample tests. *Ann Stat* 7:697–717.
22. Liu RY, Singh K (1993) A quality index based on data depth and multivariate rank-tests. *J Am Stat Assoc* 88:252–260.
23. Simonoff JS (1996) *Smoothing Methods in Statistics* (Springer-Verlag, New York).
24. Alba Fernandez V, Jimenez Gamero MD, Munoz Garcia J (2008) A test for the two-sample problem based on empirical characteristic functions. *Comput Stat Data Anal* 52:3730–3748.
25. Cao R, Van Keilegom I (2006) Empirical likelihood ratio tests for two-sample problems via non-parametric density estimation. *Can J Stat* 34:61–77.
26. Martínez-Camblor P, De Una-Alvarez J, Corral N (2008) k-sample test based on the common area of kernel density estimators. *J Stat Plan Inference* 138:4006–4020.
27. Molanes-Lopez EM, Cao R (2008) Plug-in bandwidth selector for the kernel relative density estimator. *Ann Inst Stat Math* 60:273–300.
28. Anderson NH, Hall P, Titterington DM (1994) Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *J Multivar Anal* 50:41–54.
29. Chacon JE, Duong T (2010) Multivariate plug-in bandwidth selection with unconstrained bandwidth matrices. *Test* 19:375–398.
30. Chacon JE, Duong T, Wand MP (2011) Asymptotics for general multivariate kernel density derivative estimators. *Stat Sinica* 21:807–840.
31. Duong T (2007) ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *J Stat Softw* 21:1–16.
32. Chacon JE (2009) Data-driven choice of the smoothing parametrization for kernel density estimators. *Can J Stat* 37:249–255.
33. Nel DG, Van der Merve CA (1986) A solution to the multivariate Behrens-Fisher problem. *Commun Stat Theory Methods* 15:3719–3735.
34. Ostrowski M, et al. (2010) Rab27a and Rab27b control different steps of the exosome secretion pathway. *Nat Cell Biol* 12:19–30.
35. Driskell OJ, Mironov A, Allan VJ, Woodman PG (2007) Dynein is required for receptor sorting and the morphogenesis of early endosomes. *Nat Cell Biol* 9:113–120.
36. Pincus Z, Theriot JA (2007) Comparison of quantitative methods for cell-shape analysis. *J Microsc* 227:140–156.
37. Zhao T, Murphy RF (2007) Automated learning of generative models for subcellular location: building blocks for systems biology. *Cytometry A* 71:978–90.
38. Duong T, Koch I, Wand MP (2009) Highest density difference region estimation with application to flow cytometric data. *Biom J* 51:504–521.

APPLIED MATHEMATICS

CELL BIOLOGY

Supporting Information

Duong et al. 10.1073/pnas.1117796109

SI Text

Analysis of Continuous Intracellular Structures in Micropatterned Cells. To further evaluate our approach, we performed additional analysis of diverse subcellular structures. We analyzed a sample of each 32 cells that we stained for endoplasmic reticulum exit sites marked by Sec13 (a membrane domain) and microtubules marked by α-tubulin, in the presence and absence of the drug nocodazole (Fig. S5). Additionally, we analyzed Rab8-marked membrane domains that were visualized by a stably overexpressed green fluorescent protein (GFP)-Rab8 fusion. Rab8 is a small GTPase that regulates trafficking into the primary cilium (Fig. S5). Thus, we asked whether our test statistic could distinguish between the presence or absence of a primary cilium in a given cell population. To represent continuous structures (microtubules and Rab8-marked membranes domains) by a cloud of coordinates, we "cut" them in several small structures by increasing the watershed during segmentation. Because these coordinates from continuous structures were not independent, our test statistic became suboptimal. Indeed, the formula for the estimated variance is divided by the number of independent structures; thus, the variance is overestimated for dependent structures, leading to smaller P-values. Whereas the P_{2D}-values for control half samples for microtubules and Rab8-marked membranes were not significant, the corresponding P_{3D}-values became significant (Table S3). Contrary, for Sec13 that are independent structures (same cells as for microtubules), the 2D and 3D P-values are all nonsignificant for control half samples. Thus, for dependent structures, the estimated variance needs to be divided by a number smaller than the number of structures. The adjustment factor required is not straightforward to calculate, as it requires modeling of the correlation between the dependent structures. Nonetheless, even though the computed P_{3D}-values for microtubules and Rab8-marked membranes were smaller than their true values, we could still clearly distinguish between polarized and depolarized microtubules and the absence or presence of a primary cilium marked by Rab8. The P-values between control half samples were orders of magnitude bigger than those for the comparison of different conditions (Table S3).

Sensitivity of the Statistical Test to Cell Alignment. To evaluate how sensitive our method is to errors in cell alignment on patterned substrates, we performed rotational or translational operations to represent different types of experimentally misaligned images. We focused on the untreated samples control 1 and control 2, artificially misaligned control 2, and compared each of these misaligned control 2 samples to original control 1 sample (Table S4). First, we misaligned the entire cell sample control 2 by (a) rotation of all structures about the Z-axis by a fixed angle, ranging from 10° to 50° in steps of 10°; (b) translation of all structures in the X-axis by a fixed distance, ranging from 10 to 50 pixels, in steps of 10 pixels. As expected from the simulated data, our test was very sensitive to this batch misalignment, since even 10° was sufficient to give a significant P-value (a). For the translations in the X-axis, the 2D P-value became significant for 30 pixels, and 3D P-value for 20 pixels. Next, we randomly misaligned individual cells within one sample by (c) rotation of each structure by a random angle about the Z-axis, uniformly randomly selected in the interval $[-d°, d°]$, with d ranging from 10 to 100, in

steps of 10; (d) translation of each structure by a random distance in the X-Y plane, uniformly randomly selected $[-d, d]$ pixels, with d ranging from 10 to 100, in steps of 10. In this case, our test was less sensitive. Significant 2D P-values appeared at $[-40°, 40°]$, 3D P-values at $[-30°, 30°]$; thus, small misrotations did not affect the statistical significance. For the translations, $[-50, 50]$ pixels were required to perturb the control 2 sample sufficiently to produce a significant 2D P-value, and $[-30, 30]$ pixels for the 3D test. Together, this analysis demonstrated that misaligned samples lead to false positives.

Proof: For general r, let $\psi_l^{(r)} = \int D^{\otimes r} f_l(\boldsymbol{x}) f_l(\boldsymbol{x}) d\boldsymbol{x}$. Here, D is the differential operator and the r-fold Kronecker product is denoted by the $\otimes^r$ superscript; e.g., $D^{\otimes^3} = D \otimes D \otimes D$. Thus, $D^{\otimes^r}$ is the r-th order differential operator arranged into a vector of length d^r (1). From Chacón and Duong (2), the expected value of $\hat{\psi}_l$ is $E\hat{\psi}_l = \psi_l + \{1/2(\text{vec}^T \mathbf{H}_l)\psi_l^{(2)} + n^{-1} K_{\mathbf{H}_l}(0)\}\{1 + o(1)\}$ For the cross-product integrated density functional,

$$E\hat{\psi}_{l_1,l_2} = E K_{\mathbf{H}_{l_1}}(\boldsymbol{X} - \boldsymbol{Y}) = \iint K_{\mathbf{H}_{l_1}}(\boldsymbol{x} - \boldsymbol{y}) f_1(\boldsymbol{x}) f_2(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y}$$
$$= \psi_{l_1,l_2} + \frac{1}{2}(\text{vec } ^T \mathbf{H}_{l_1}) \int D^{\otimes^2} f_{l_1}(\boldsymbol{x}) f_{l_2}(\boldsymbol{x}) d\boldsymbol{x} \{1 + o(1)\}.$$

where $(l_1, l_2) = (1, 2)$ or $(2, 1)$. Thus, the expected value of $\hat{T}$ is

$$E\hat{T} = E\hat{\psi}_1 + E\hat{\psi}_2 - (E\hat{\psi}_{1,2} + E\hat{\psi}_{2,1})$$
$$= \psi_1 + \psi_2 - (\psi_{1,2} + \psi_{2,1})$$
$$+ \left[\frac{1}{2}(\text{vec } ^T \mathbf{H}_1)\psi_1^{(2)} + \frac{1}{2}(\text{vec } ^T \mathbf{H}_2)\psi_2^{(2)} + n_1^{-1}|\mathbf{H}_1|^{-1/2} K(0)\right.$$
$$+ n_2^{-1}|\mathbf{H}_2|^{-1/2} K(0) - \frac{1}{2}(\text{vec}^T \mathbf{H}_1) \int D^{\otimes^2} f_1(\boldsymbol{x}) f_2(\boldsymbol{x}) d\boldsymbol{x}$$
$$\left. - \frac{1}{2}(\text{vec } ^T \mathbf{H}_2) \int D^{\otimes^2} f_2(\boldsymbol{x}) f_1(\boldsymbol{x}) d\boldsymbol{x}\right]\{1 + o(1)\}.$$

The variance is

$$\text{Var } \hat{T} = \text{Var}[\hat{\psi}_1 + \hat{\psi}_2 - (\hat{\psi}_{1,2} + \hat{\psi}_{2,1})]$$
$$= \text{Var } \hat{\psi}_1 + \text{Var } \hat{\psi}_2 + \text{Var } \hat{\psi}_{1,2} + \text{Var } \hat{\psi}_{2,1}$$
$$- 2\text{Cov}(\hat{\psi}_1, \hat{\psi}_{1,2}) - 2\text{Cov}(\hat{\psi}_1, \hat{\psi}_{2,1}) - 2\text{Cov}(\hat{\psi}_2, \hat{\psi}_{1,2})$$
$$- 2\text{Cov}(\hat{\psi}_2, \hat{\psi}_{2,1}) + 2\text{Cov}(\hat{\psi}_{1,2}, \hat{\psi}_{2,1})$$

since $\text{Cov}(\hat{\psi}_1, \hat{\psi}_2) = 0$. From Chacón and Duong (2),

$$\hat{\psi}_1 = \{4n_l^{-1}(\xi_1 - \psi_l^2) + 2n_l^{-2}\psi_l|\mathbf{H}_l|^{-1/2} m_0(K^2)\}\{1 + o(1)\},$$

where $\xi_l = \int f_l(\boldsymbol{x})^3 d\boldsymbol{x}$. From Lee and Dehling (3, Theorem 1), $\text{Var } \hat{\psi}_{1,2} = n_1^{-1}\text{Var}[(K_{\mathbf{H}_1} * f_2)(\boldsymbol{X})]\{1 + o(1)\}$ a n d $\text{Var } \hat{\psi}_{2,1} =$

$n_2^{-1}\mathrm{Var}[(K_{\mathbf{H}_2}*f_1)(\boldsymbol{Y})]\{1+o(1)\}$ where $*$ is the convolution operator. So we have $\mathrm{Var}\,\hat{\psi}_{l_1,l_2} = n_{l_1}^{-1}[\int f_{l_1}(\boldsymbol{x})^2 f_{l_2}(\boldsymbol{x})d\boldsymbol{x} - \psi_{l_1,l_2}^2]\{1+o(1)\}$ since

$$\mathrm{E}[(K_{\mathbf{H}_1}*f_2)(\boldsymbol{X})] = \iint K_{\mathbf{H}_1}(\boldsymbol{x}-\boldsymbol{y})f_1(\boldsymbol{x})f_2(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} = \psi_{1,2}\{1+o(1)\}$$

$$\mathrm{E}[(K_{\mathbf{H}_2}*f_1)(\boldsymbol{Y})] = \iint K_{\mathbf{H}_2}(\boldsymbol{x}-\boldsymbol{y})f_2(\boldsymbol{x})f_1(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} = \psi_{2,1}\{1+o(1)\}$$

$$\mathrm{E}[(K_{\mathbf{H}_1}*f_2)(\boldsymbol{X})^2] = \int (K_{\mathbf{H}_1}*f_2)(\boldsymbol{x})^2 f_1(\boldsymbol{x})d\boldsymbol{x}$$
$$= \int \{f_2(\boldsymbol{x})+o(1)\}^2 f_1(\boldsymbol{x})d\boldsymbol{x}$$
$$= \int f_1(\boldsymbol{x})f_2(\boldsymbol{x})^2\{1+o(1)\}d\boldsymbol{x},$$

and likewise $\mathrm{E}[(K_{\mathbf{H}_2}*f_1)(\boldsymbol{X})^2] = \int f_1(\boldsymbol{x})^2 f_2(\boldsymbol{x})\{1+o(1)\}d\boldsymbol{x}$. For the remaining covariance terms,

$$\mathrm{Cov}(\hat{\psi}_1,\hat{\psi}_{1,2}) = n_1^{-1}\mathrm{E}[K_{\mathbf{H}_1}(\boldsymbol{X}_1-\boldsymbol{X}_2)K_{\mathbf{H}_1}(\boldsymbol{X}_1-\boldsymbol{Y})]$$
$$- n_1^{-1}\mathrm{E}[K_{\mathbf{H}_1}(\boldsymbol{X}_1-\boldsymbol{X}_2)]\mathrm{E}[K_{\mathbf{H}_1}(\boldsymbol{X}_1-\boldsymbol{Y})]$$
$$= n_1^{-1}[\iiint K_{\mathbf{H}_1}(\boldsymbol{x}-\boldsymbol{y})K_{\mathbf{H}_1}(\boldsymbol{x}-\boldsymbol{z})f_1(\boldsymbol{x})f_1(\boldsymbol{y})f_2(\boldsymbol{z})d\boldsymbol{x}d\boldsymbol{y}d\boldsymbol{z}$$
$$- \psi_1\psi_{1,2}]\{1+o(1)\}$$
$$= n_1^{-1}[\int f_1(\boldsymbol{x})^2 f_2(\boldsymbol{x})d\boldsymbol{x} - \psi_1\psi_{1,2}]\{1+o(1)\}.$$

So more generally, by symmetry, $\mathrm{Cov}(\hat{\psi}_{l_0}, \hat{\psi}_{l_1,l_2}) = n_{l_0}^{-1}[\int f_{l_0}(\boldsymbol{x})f_{l_1}(\boldsymbol{x})f_{l_2}(\boldsymbol{x})d\boldsymbol{x} - \psi_{l_0}\psi_{l_1,l_2}]\{1+o(1)\}$; and

$$\mathrm{Cov}(\hat{\psi}_{1,2}, \hat{\psi}_{2,1}) = n_1^{-1}n_2^{-1}\mathrm{Cov}[K_{\mathbf{H}_1}(\boldsymbol{X}-\boldsymbol{Y})K_{\mathbf{H}_2}(\boldsymbol{X}-\boldsymbol{Y})]$$
$$+ (n_1-1)^{-1}n_1^{-1}n_2^{-1}\mathrm{Cov}[K_{\mathbf{H}_1}(\boldsymbol{X}_1-\boldsymbol{Y})K_{\mathbf{H}_2}(\boldsymbol{X}_2-\boldsymbol{Y})]$$
$$+ n_1^{-1}(n_2-1)^{-1}n_2^{-1}\mathrm{Cov}[K_{\mathbf{H}_1}(\boldsymbol{X}-\boldsymbol{Y}_1)K_{\mathbf{H}_2}(\boldsymbol{X}-\boldsymbol{Y}_2)]$$
$$= \left\{\frac{1}{2}n_1^{-1}n_2^{-1}[\psi_{1,2}|\mathbf{H}_1|^{-1/2}\int K(\boldsymbol{x})K_{\mathbf{H}_2\mathbf{H}_1^{-1}}(\boldsymbol{x})d\boldsymbol{x} - \psi_{1,2}^2\right.$$
$$+ \psi_{2,1}|\mathbf{H}_2|^{-1/2}\int K(\boldsymbol{x})K_{\mathbf{H}_1\mathbf{H}_2^{-1}}(\boldsymbol{x})d\boldsymbol{x} - \psi_{2,1}^2]$$
$$+ n_1^{-1}[\int f_1(\boldsymbol{x})^2 f_2(\boldsymbol{x})d\boldsymbol{x} - \psi_{1,2}^2]$$
$$\left. + n_2^{-1}[\int f_1(\boldsymbol{x})f_2(\boldsymbol{x})^2 d\boldsymbol{x} - \psi_{2,1}^2]\right\}\{1+o(1)\}.$$

where assumption (N) ensures that the integrals $\int K(\boldsymbol{x})K_{\mathbf{H}_2\mathbf{H}_1^{-1}}(\boldsymbol{x})d\boldsymbol{x}$ and $\int K(\boldsymbol{x})K_{\mathbf{H}_1\mathbf{H}_2^{-1}}(\boldsymbol{x})d\boldsymbol{x}$ are $O(1)$. Combining these expressions, the variance is

$$\mathrm{Var}\,\hat{T} = \{4n_1^{-1}(\xi_1 - \psi_1^2) + 2n_1^{-2}\psi_1|\mathbf{H}_1|^{-1/2}m_0(K^2)$$
$$+ 4n_1^{-1}(\xi_2 - \psi_2^2) + 2n_2^{-2}\psi_2|\mathbf{H}_2|^{-1/2}m_0(K^2)$$
$$+ 3n_1^{-1}[\int f_1(\boldsymbol{x})^2 f_2(\boldsymbol{x})d\boldsymbol{x} - \psi_{1,2}^2]$$
$$+ 3n_2^{-1}[\int f_1(\boldsymbol{x})f_2(\boldsymbol{x})^2 d\boldsymbol{x} - \psi_{2,1}^2]$$
$$- 4n_1^{-1}[\int f_1(\boldsymbol{x})^2 f_2(\boldsymbol{x})d\boldsymbol{x} - \psi_1\psi_{1,2}]$$
$$- 4n_2^{-1}[\int f_1(\boldsymbol{x})f_2(\boldsymbol{x})^2 d\boldsymbol{x} - \psi_2\psi_{1,2}]$$
$$+ n_1^{-1}n_2^{-1}[\psi_{1,2}|\mathbf{H}_1|^{-1/2}\int K(\boldsymbol{x})K_{\mathbf{H}_2\mathbf{H}_1^{-1}}(\boldsymbol{x})d\boldsymbol{x} - \psi_{1,2}^2$$
$$+ \psi_{2,1}|\mathbf{H}_2|^{-1/2}\int K(\boldsymbol{x})K_{\mathbf{H}_1\mathbf{H}_2^{-1}}(\boldsymbol{x})d\boldsymbol{x}$$
$$- \psi_{2,1}^2]\}\{1+o(1)\}.$$

From Chacón and Duong (2), an optimal bandwidth for $\hat{\psi}_l$ is $\mathbf{H}_l = O(n_l^{-2/(d+2)})$ then $\mathrm{Var}\,\hat{T} = O(n_1^{-1} + n_2^{-1} + n_1^{-1}n_1^{-2/(d+2)} + n_2^{-1}n_2^{-2/(d+2)} + n_1^{-1}n_2^{-2/(d+2)} + n_1^{-1}n_2^{-2/(d+2)})$; i.e., the $O(n_1^{-1} + n_2^{-1})$ terms dominate, so

$$\mathrm{Var}\,\hat{T} = \{4n_1^{-1}(\xi_1 - \psi_1^2) + 4n_1^{-1}(\xi_2 - \psi_2^2) + n_1^{-1}[4\psi_1\psi_{1,2}$$
$$- \int f_1(\boldsymbol{x})^2 f_2(\boldsymbol{x})d\boldsymbol{x} - 3\psi_{1,2}^2] + n_2^{-1}[4\psi_2\psi_{1,2}$$
$$- \int f_1(\boldsymbol{x})f_2(\boldsymbol{x})^2 d\boldsymbol{x} - 3\psi_{2,1}^2]\}\{1+o(1)\}.$$

Under the null hypothesis, $f_1 \equiv f_2 \equiv f$, so we have $\psi_1^{(r)} = \psi_2^{(r)} = \psi^{(r)}$ for all r, and that $\psi_{1,2} = \psi_{2,1} = \psi$. Thus, the $O(\mathrm{tr}\mathbf{H}_1 + \mathrm{tr}\mathbf{H}_2)$ terms in the bias vanish. Expanding the Taylor series further yields that the next leading terms are $O(\mathrm{tr}^2\mathbf{H}_1 + \mathrm{tr}^2\mathbf{H}_2)$. Thus the null expected value simplifies to

$$\mathrm{E}\hat{T} = \mu_T = [n_1^{-1}|\mathbf{H}_1|^{-1/2}K(\mathbf{0}) + n_2^{-1}|\mathbf{H}_1|^{-1/2}K(\mathbf{0})]\{1+o(1)\}$$

since the $O(\mathrm{tr}^2\mathbf{H}_1 + \mathrm{tr}^2\mathbf{H}_2) = O(n_1^{-4/(d+2)} + n_2^{-4/(d+2)})$ terms, which are dominated by the $O(n_1^{-1}|\mathbf{H}_1|^{-1/2} + n_2^{-1}|\mathbf{H}_2|^{-1/2}) = O(n_1^{-2/(d+2)} + n_2^{-2/(d+2)})$ terms. So $\mathrm{E}\hat{T}$ is asymptotically zero by assumption (H). The null variance simplifies to

$$\mathrm{Var}\,\hat{T} = 3(n_1^{-1} + n_2^{-1})\lfloor\int f(\boldsymbol{x})^3 d\boldsymbol{x} - (\int f(\boldsymbol{x})^2 d\boldsymbol{x})^2\rfloor\{1+o(1)\}.$$

Standard U-statistics theory, e.g., Theorem 7.1 (4), asserts the asymptotic normality of $\hat{\psi}_l$. More recently, Theorem 2 (3) establishes the asymptotic normality of $\hat{\psi}_{l_1,l_2}$, so the asymptotic normality of $\hat{T}$ follows immediately.

1. Holmquist, B (1996) The d-variate vector Hermite polynomial of order k. *Linear Algebra Appl* 237–238:155–190.
2. Chacón JE, Duong T (2010) Multivariate plug-in bandwidth selection with unconstrained bandwidth matrices. *Test* 19:375–398.
3. Lee M-LT, Dehling HG (2005) Generalized two-sample U-statistics for clustered data. *Stat Neerl* 59:313–323.
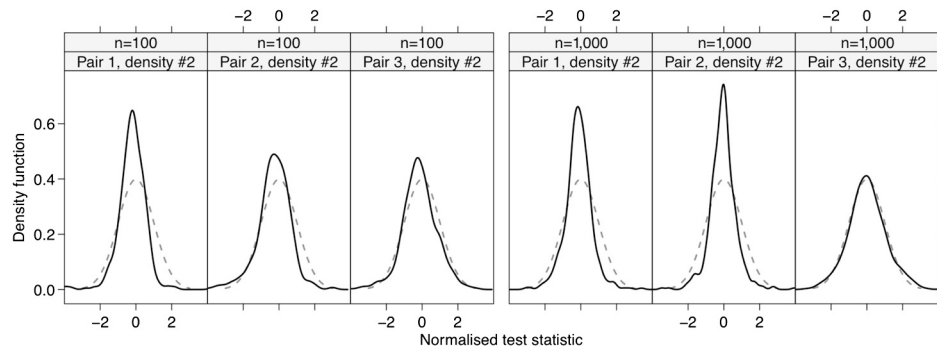4. Hoeffding W (1948) A class of statistics with asymptotically normal distribution. *Ann Math Stat* 19:239:325.

Fig. S1. Simulation results. Comparison of empirical null distribution of the normalized $\hat{T}$ (solid black line) to the standard normal (gray dashed line) for the second density in each target pair, for sample sizes $n = 100$; $n = 1000$.
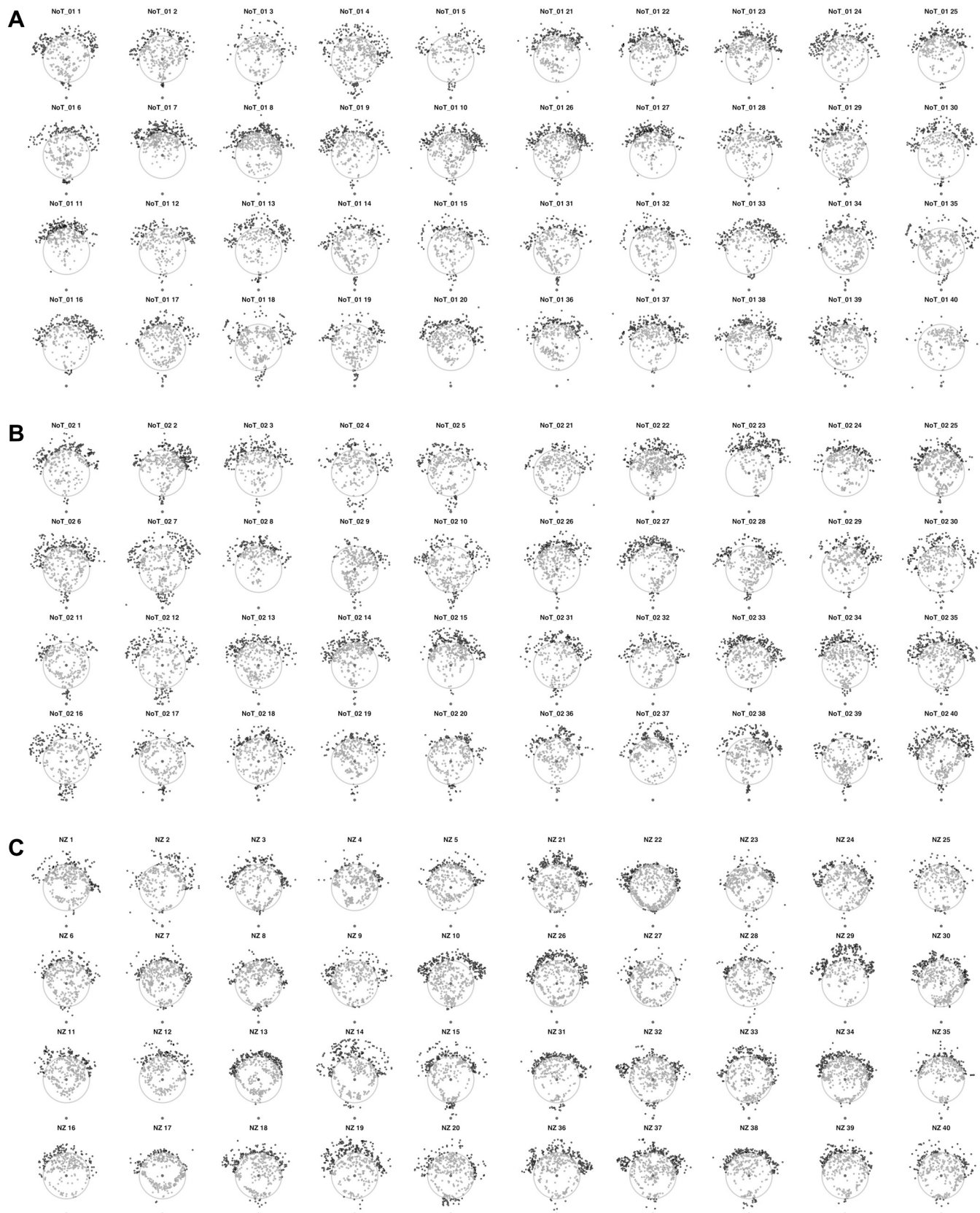
Fig. S2. 2D scatter plots for individual cells after alignment. (*A*), (*B*) Control experiments without treatments (Ctrl) with 40 cells each. (*C*) Nocodazole-treatment experiment (NZ) with 40 cells.
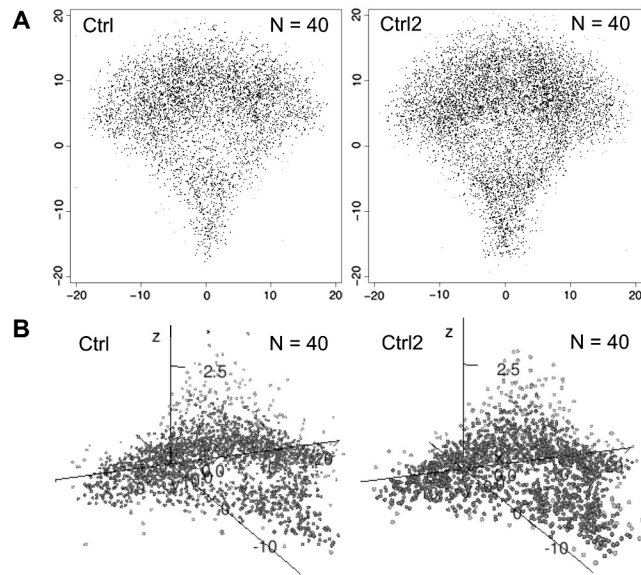
Fig. S3. Comparison of the morphology of multivesicular bodies (MVB) in independent control (Ctrl) experiments. (*A*) Representative 2D scatter plots for two control groups of 40 cells each, with $n = 11786$ structures for Ctrl and $n = 12585$ structures for Ctrl2. (*B*) Representative 3D scatter plots for Ctrl and Ctrl2. The *z*-axis is zoomed 500%.
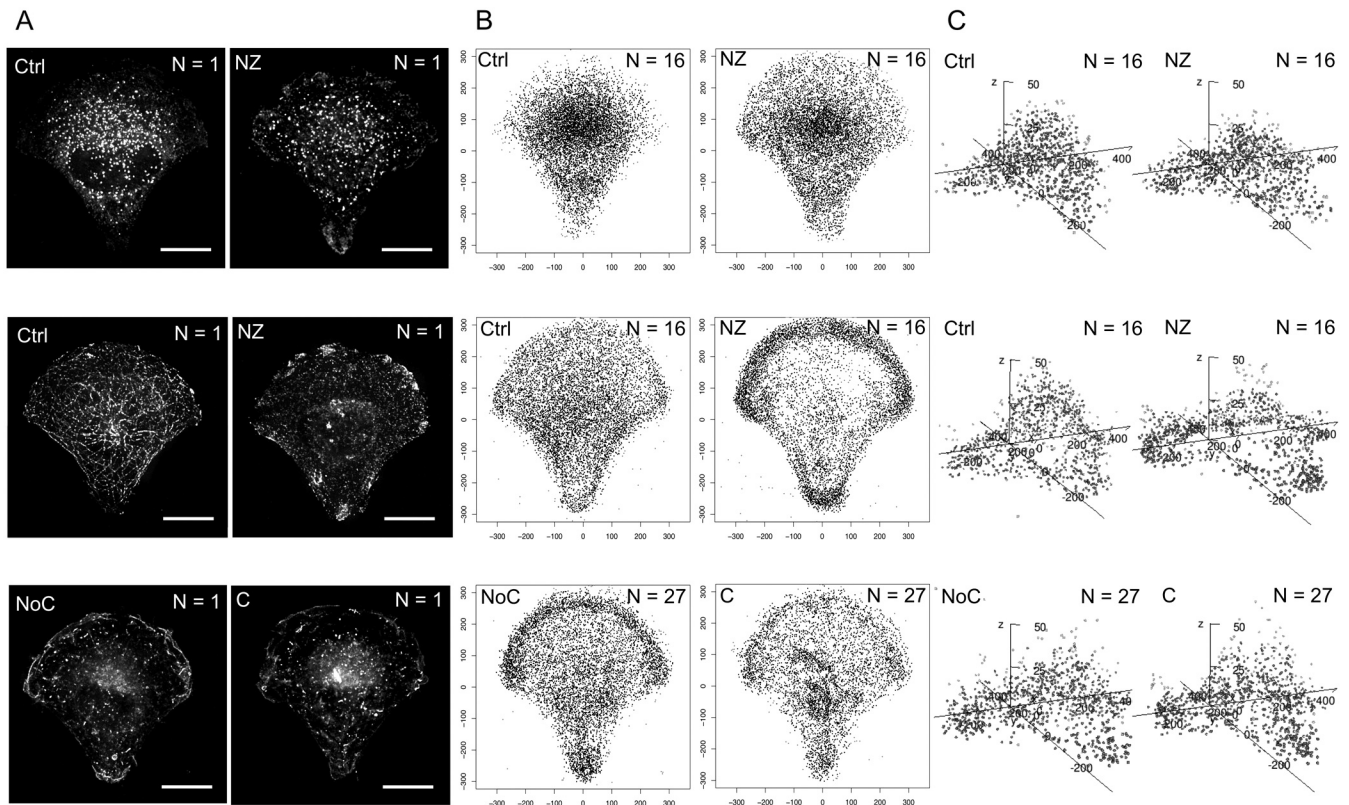

Fig. S4. Comparison of the morphology of ER exit sites, microtubules, and Rab8-marked membranes in different conditions. (*A*) Maximum intensity projection of the deconvolved fluorescence of Sec13-marked ER exit sites (upper panel), α-tubulin-marked microtubules (middle panel) without treatment (Ctrl), and upon nocodazole (NZ) treatment, as well as Rab8-marked membranes in the absence (NoC) and presence (C) of a primary cilium (lower panel). Scale bars are 10 μm. (*B*) Representative 2D scatter plots for N cells. (*C*) Representative 3D scatter plots for N cells. The *z*-axis is zoomed 500%.
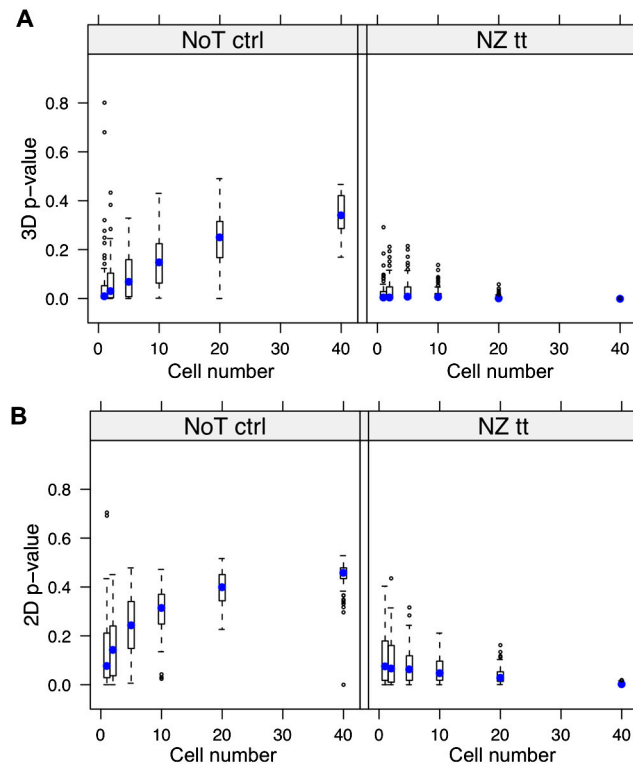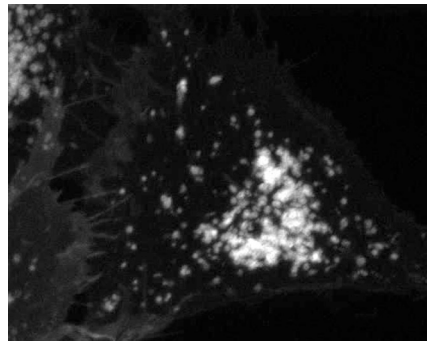
Fig. S5. *P*-value distributions from density-based test. Box plot of 3D (*A*) and 2D (*B*) *P*-values for CD63 control and nocodazole-treatment conditions as a function of the number of cells analyzed for 100 comparisons to indicate the spread of the *P*-values from kernel test. See also Fig. 2*H*.

Movie S1. Time lapse of the morphology of multivesicular bodies (MVB) upon disruption of the microtubule cytoskeleton. RPE-1 cells stably expressing EGFP-CD63-marked MVBs were images using Z-series of 0.2 μm at time intervals of 60 s. Nocodazole was added at frame 12.

Movie S1 (AVI)

Table S1. *P*-value comparison between density-based test and parametric *t*-test

Pair	Sample size	2D KDE test	2D *t*-test (MNV test)
1	1000	$1.142 * 10^{-29}$	0
2	1000	$1.353 * 10^{-8}$	0.5195
3	1000	$3.386 * 10^{-23}$	0.2158

Average *P*-values of 2D density-based and parametric *t*-test (MNV) test statistics from 100 comparisons of simulated data drawn from the 3 pairs of normal mixture densities in Fig. 1.

Table S2. Comparison of the morphology of ER exit sites, microtubules and Rab8-marked membranes in different conditions

| | 2D P-values | | | 3D P-values | | |
	Microtubules	ER exit sites	Rab8	Microtubules	ER exit sites	Rab8
Condition1 vs Condition1	1,1460E−01	3,5994E−01	1,3073E−01	**6,1724E−04**	3,1606E−01	**8,2577E−03**
Condition1 vs Condition 2	**6,5862E−41**	**1,2151E−02**	**4,3428E−04**	**2,7758E−228**	**3,0281E−04**	**2,6783E−05**

Average P-values of 2D and 3D KDE-based test statistic from 100 comparisons for the morphology of microtubules (N = 16 cells per condition), ER exit sites (N = 16 cells per condition), and Rab8 membranes (N = 27 cells per condition). Significant P-values are bold. See also Fig. S4.

Table S3. P-value comparison between density-based test and permutation analysis

| | CD63 NoT control | |
Sample size	Permutation	3D KDE test
1	4,9460E−01	5,1911E−02
2	4,5910E−01	6,3279E−02
5	5,2430E−01	9,3159E−02
10	5,5170E−01	1,4794E−01
20	4,9590E−01	1,8011E−01
40	5,0770E−01	3,9350E−01
	CD63 NZ	
Sample size	Permutation	3D KDE test
1	2,3900E−01	2,2195E−02
2	2,1410E−01	3,0114E−02
5	1,2260E−01	3,1677E−02
10	2,9600E−02	1,6527E−02
20	1,1000E−03	5,4816E−03
40	0	1,6996E−04

Average P-values of permutation and 3D KDE-based test statistics from 100 comparisons, in which corresponding number of cells were picked randomly from 100 Ctrl or 66 NZ-treated cells. Average P-values for the permutation test were previously calculated (1).

1 Schauer K, et al. (2010) Probabilistic density maps to study global endomembrane organization. *Nat Methods* 7:560–566.

Table S4. Sensitivity of the statistical test to cell alignment

Comparison of control 1 to	2D *P*-value	3D *P*-value
control 2	2.5811E−01	1.1379E−01
(a) control 2: Fixed Z-rotation (°)		
10	**1.2026E−02**	**3.1615E−04**
20	**1.2107E−08**	**4.4409E−16**
30	**4.3601E−22**	**9.7947E−42**
40	**1.4477E−43**	**1.6058E−79**
50	**1.0115E−71**	**1.6590E−124**
(b) control 2: Fixed X-translation (pix)		
10	1.7983E−01	5.9271E−02
20	5.1206E−02	**3.7231E−03**
30	**2.7942E−03**	**3.3882E−06**
40	**9.3938E−06**	**1.8697E−12**
50	**5.4986E−10**	**2.0579E−23**
(c) control 2: Random Z-rotation (°)		
[−10, 10]	2.3342E−01	1.0459E−01
[−20, 20]	1.5036E−01	5.0567E−02
[−30, 30]	5.9005E−02	**9.3238E−03**
[−40, 40]	**1.1310E−02**	**4.4162E−04**
[−50, 50]	**7.3779E−04**	**3.1049E−06**
[−60, 60]	**9.4306E−06**	**1.9971E−09**
[−70, 70]	**2.2935E−08**	**1.9107E−13**
[80, −80]	**2.0369E−11**	**1.4874E−17**
[90, −90]	**4.6629E−15**	**7.6345E−22**
[−100, 100]	**2.0921E−18**	**2.8145E−25**
(d) control 2: Random X-Y-translation (pix)		
[−10, 10]	2.3646E−01	1.0205E−01
[−20, 20]	1.9582E−01	7.7122E−02
[−30, 30]	1.4037E−01	**4.3900E−02**
[−40, 40]	8.1624E−02	**1.6126E−02**
[−50, 50]	**3.3639E−02**	**2.9152E−03**
[−60, 60]	**8.5085E−03**	**1.8488E−04**
[−70, 70]	**1.1634E−03**	**3.1264E−06**
[80, −80]	**6.4743E−05**	**7.7590E−09**
[90, −90]	**1.2998E−06**	**2.4933E−12**
[−100, 100]	**8.9372E−09**	**1.1102E−16**

Average *P*-values of 2D and 3D KDE-based test statistic from 100 comparisons, in which control 1 was compared to a misaligned control 2. Significant *P*-values are bold.

Table S5. *P*-values for morphology changes of unconstrained cells in live cell experiment

Group	1 ($t_{−11−6}$)	2 ($t_{−5−0}$)	3 ($t_{1−6}$)	4 ($t_{7−12}$)
1 ($t_{−11−6}$)	2D	**3.57E-01**	**4.29E-02**	**4.84E-06**
2 ($t_{−5−0}$)	4,14E−01		**4.08E-02**	**2.48E-05**
3 ($t_{1−6}$)	4,96E−02	3,76E−02		**8.48E-02**
4 ($t_{7−12}$)	4,00E−06	7,69E−07	1,15E−01	**3D**

P-values for the 2D and 3D (bold) KDE test statistic for the comparison between the groups of images containing each six time frames of EFGP-CD63-marked structures. Group 1 and 2 are nontreated controls acquired before addition of the drug. Group 3 and 4 are nocodazole-treatment conditions acquired after the addition of the drug. See also Fig. 3*B*.