Decoupling global and local relations in biased biological systems

Assaf Zaritsky[1,2], Uri Obolski[3&], Zuzana Kadlecova[1&], Zhuo Gan[1,2], Yi Du[1], Sandra Schmid[1], Gaudenz Danuser[1,2*]

[1]Department of Cell Biology, UT Southwestern Medical Center, Dallas, TX 75390, USA.

[2]Department of Bioinformatics, UT Southwestern Medical Center, Dallas, TX 75390, USA.

[3]Department of Molecular Biology and Ecology of Plants, Tel-Aviv University, Tel Aviv, Israel.

[&]Equal contribution

[*] To whom correspondence should be sent: Gaudenz Danuser, 5323 Harry Hines Blvd., Dallas, Texas 75390-9039. Phone: 214-648-3835. Fax: 214-648-7491. Email: Gaudenz.Danuser@UTSouthwestern.edu.

Condensed title: DeBias: analysis of coupled biological variables

Keywords: Global bias, Local interaction, Colocalization, Cytoskeleton alignment, Collective cell migration, Endocytosis.

Number of characters: 34,100

Abstract

Analysis of coupled variables is a core concept of cell biological inference, with colocalization of two molecules as a proxy for protein interaction being an ubiquitous example. However, external effectors may influence the observed colocalization pattern independently from the local interaction of two proteins. Such global bias is often neglected when interpreting colocalization. Here, we describe DeBias, a computational method to quantify and decouple global bias from local interactions by modeling the observed colocalization between coupled variables as the cumulative contribution of a global and a local component. We demonstrate applications of DeBias in three areas of cell biology at different scales: Analysis of the (1) alignment of vimentin fibers and microtubules in the context of polarized cells; (2) alignment of cell velocity and traction stress during collective migration; and (3) specific recruitment of transmembrane receptors to clathrin-coated pits during endocytosis. The DeBias software package is freely accessible online via a web-server at https://debias.biohpc.swmed.edu[1].

[1] Web server is not online yet, waiting for UTSW information resources approval. Until then, please use the less friendly GITHUB repository https://git.biohpc.swmed.edu/ydu/debias/tree/master.

Introduction

Interpretation of the relations of coupled variables is a classic problem that appears in many flavors of cell biology. One example is the spatiotemporal colocalization of molecules – a critical clue to interaction between molecular components; another example is alignment of orientational molecular activities, such as between different filamentous networks. However, colocalization or alignment may also occur because two components are associated with a third effector. For example, the internal components of a polarized cell are organized along the polarization axis, making it difficult to quantify how much of the observed alignment between two filamentous networks is related to common organizational constraints imposed by the polarity cue, i.e., both networks are independently biased by global effects, and how much of it is indeed caused by direct interaction between the filaments. The combined effects of any *global bias* with *local interactions* are manifested in the joint distribution of the spatially coupled variables. The contribution of global bias to this joint distribution can be recognized by the deviation of the marginal distributions of each of the two variables from an (un-biased) uniform distribution. Another example is introduced with molecular platforms and scaffolds for protein localization: different proteins may tightly colocalize, without necessarily directly interacting. Although global bias can significantly mislead the interpretation of colocalization measurements, most colocalization studies do not account for this effect (Adler and Parmryd, 2010; Bolte and Cordelieres, 2006; Costes et al., 2004; Das et al., 2015; Dunn et al., 2011; Helmuth et al., 2010; Kalaidzidis et al., 2015; Pearson, 1901; Rizk et al., 2014; Serra-Picamal et al., 2012; Tambe et al., 2011). Previous approaches assessed spatial correlations (e.g., (Drew et al., 2015; Karlon et al., 1999)) or variants of mutual information (e.g., (Krishnaswamy et al., 2014; Reshef et al., 2011)) but did not explicitly quantify the contribution of a common global bias.

Here, we present *DeBias* as a method to decouple the global bias (expressed by a *global index*) from the bona fide local interaction (expressed by a *local index*) in colocalization of two independently-measured spatial variables. The decoupling offers two major advantages over existing colocalization methods: First, it enables the distinction of global mechanisms that constrain positioning of events from specific local interactions between spatially-matched variables. Second, it permits a more accurate assessment of the level of local interactions by exploiting the representation of the observed colocalization by the global and local indices. Our method is dubbed DeBias because it **De**couples the global **bias** and local interaction from the observed colocalization.

To highlight its capabilities, *DeBias* was applied on data from 3 different areas in cell biology: (1) Analysis of the alignment of vimentin fibers and microtubules in the context of polarized cells; (2) Analysis of the alignment of cell velocity and traction stress during collective cell migration; and (3) Analysis of the specific recruitment of transmembrane receptors to clathrin-coated pits during endocytosis. These applications, ranging in scale from macromolecular to intercellular, demonstrate the generalization of the method for a wide range of applications.

Results

Similar observed colocalization originating from different mechanisms

The issue of separating contributions from global bias and local interactions to the apparent colocalization of two entities is best illustrated with the alignment of two sets of variables that carry orientational information. Examples of 'orientational colocalization' include the alignment of two filament networks (Drew et al., 2015; Nieuwenhuizen et al., 2015), or the alignment of cell velocity and traction stress, a phenomenon referred to as *plithotaxis* (Das et al., 2015; Tambe et al., 2011; Trepat and Fredberg, 2011). In these systems, global bias imposes a preferred axis of orientation on the two variables, which is independent of the local interactions between the two variables (Fig. 1A).

Similar observed alignments may arise from different levels of global bias and local interactions. Hence, different mechanisms can lead to a similar outcome. This is demonstrated by simulation of two independent orientational random variables X and Y (Fig. 1B, left), from which pairs of samples x_i and y_i are drawn to form an alignment angle θ_i (Fig. 1B, middle). Then, a local interaction between the two variables is modeled by co-aligning θ_i by a degree of ζ_i, resulting in two variables x'_i and y'_i with an observed alignment $\theta_i - \zeta_i$ (Fig. 1B, right).

We show the joint distribution of X, Y for 4 simulations (Fig. 1C) where X and Y are normal distributions with identical means but different standard deviations (σ) and magnitudes of local interactions (ζ). The latter is defined as $\zeta = \alpha\theta$ (Fig. 1B, $\alpha=1$ for perfect alignment). Throughout the simulations both the standard deviation σ and α are gradually increased (Fig. 1C, left-to-right), implying that the global bias in the orientational variables is reduced while their local interactions get stronger. As a result, all simulations display similar observed alignment

(mean values, 18.9°-19.5°). Fig. 1D visualizes 100 samples from each of the two most distinct scenarios: low σ and no local interaction (σ = 17°, α = 0) leads to tendency of X and Y to align independently to one direction (left); higher variance together with increased interaction (σ = 40°, α = 0.5) leads to more diverse orientations of X and Y (right), while maintaining similar mean alignment. This simple example highlights the possibility of observing similar alignment arising from different mechanisms of global bias and local interactions.

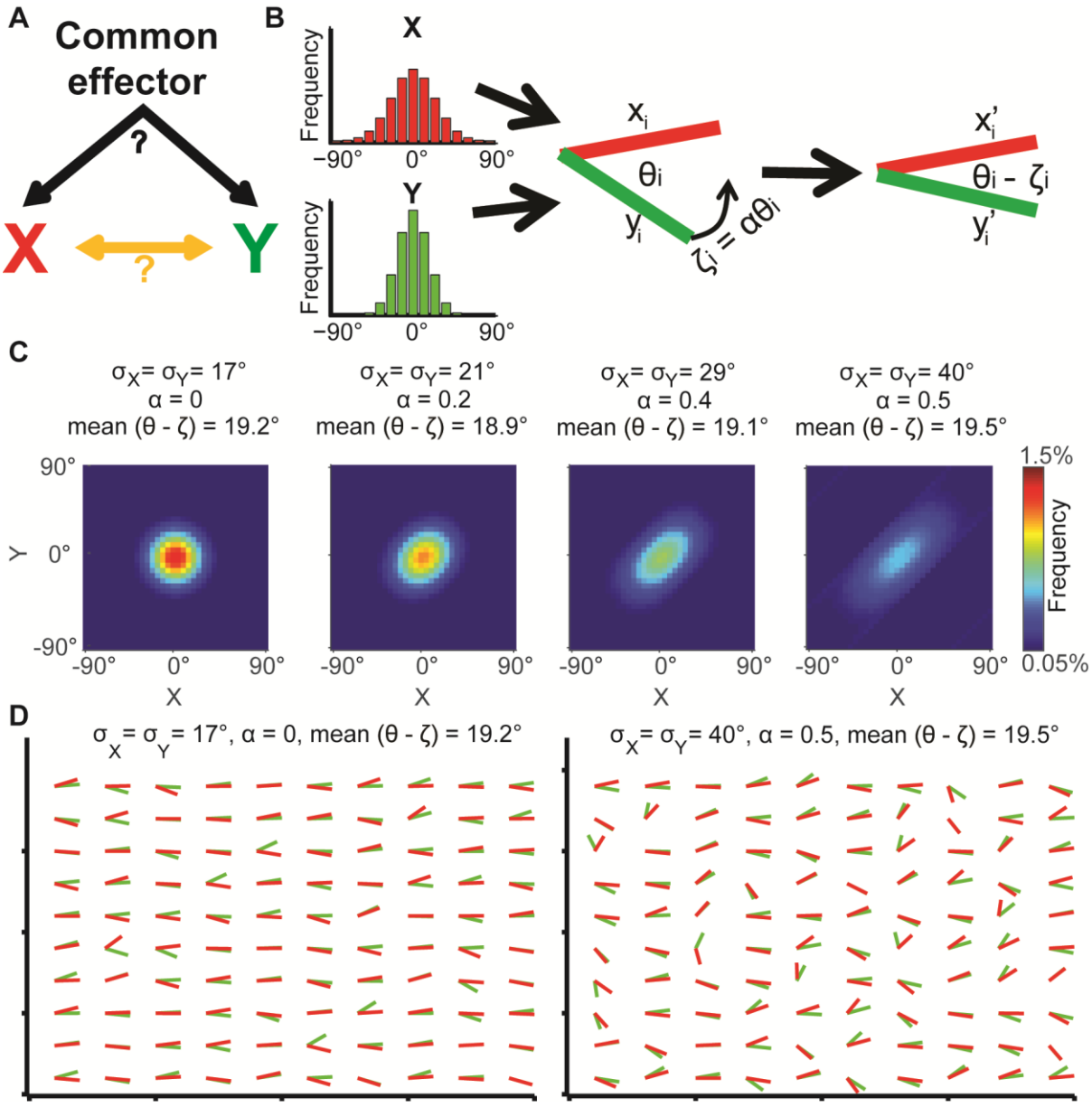


Figure 1: (A) The relation between two variables X, Y can be explained from a combination of direct interactions (orange) and a common effector/s. (B) Simulation. Given two distributions X, Y pairs of coupled variables are constructed by drawing sample pairs (x_i,y_i) from X, Y and transforming them to (x_i',y_i') by a correction parameter $\zeta_i = \alpha\theta_i$, which represents the effect of a local interaction between x_i and y_i. α is constant for each of these simulations. (C) Simulated joint distributions. X, Y normal distributions with mean 0 and $\sigma_X = \sigma_Y$. Shown are the joint distributions of 4 simulations with reduced global bias (i.e., increased standard deviation σ_X, σ_Y) and increased local interaction (left-to-right), all scenarios have similar observed mean alignment of ~19°. (D) Example of 100 random coupled orientational variables from the two most extreme scenarios in panel C. Most orientations are aligned with the x-axis when the global bias is high and no local interaction exists (left), while the orientations are less aligned with the x-axis but maintain the mean alignment between (x_i',y_i') pairs for reduced global bias and increased local interaction (right).

DeBias: a method to assess the global and local contribution to observed co-alignment

DeBias models the observed marginal distributions X' and Y' as the sum of the contributions by a common effector, i.e., the global bias, and by local interactions that effect the co-alignment of the two variables in every data point (Fig. 2A).

In a scenario without any global bias or local interaction between X' and Y', the observed alignment would be uniformly distributed (denoted *uniform*). Hence any deviation from the uniform distribution would reflect contributions from both the global bias and the local interactions. To extract the contribution of the global bias we constructed a resampled alignment distribution (denoted *resampled*) from independent samples of the marginal distributions X' and Y', which decouples matched pairs (x'_i, y'_i), and thus excludes their local interactions. The global bias is defined as the dissimilarity between the uniform and resampled distributions and accordingly, describes to what extent elements of X' and Y' are aligned without local interaction (Fig. 2B). If a local interaction exists then the distribution of the observed alignment angles will differ from independently resampled alignment distribution. Hence, the uniform distribution will be less similar to the experimentally observed alignment distribution (denoted *observed*) than to

the resampled distribution. Accordingly, the local interaction is defined by the difference in dissimilarities between the observed and uniform distributions and between the resampled and uniform distributions (Fig. 2B).

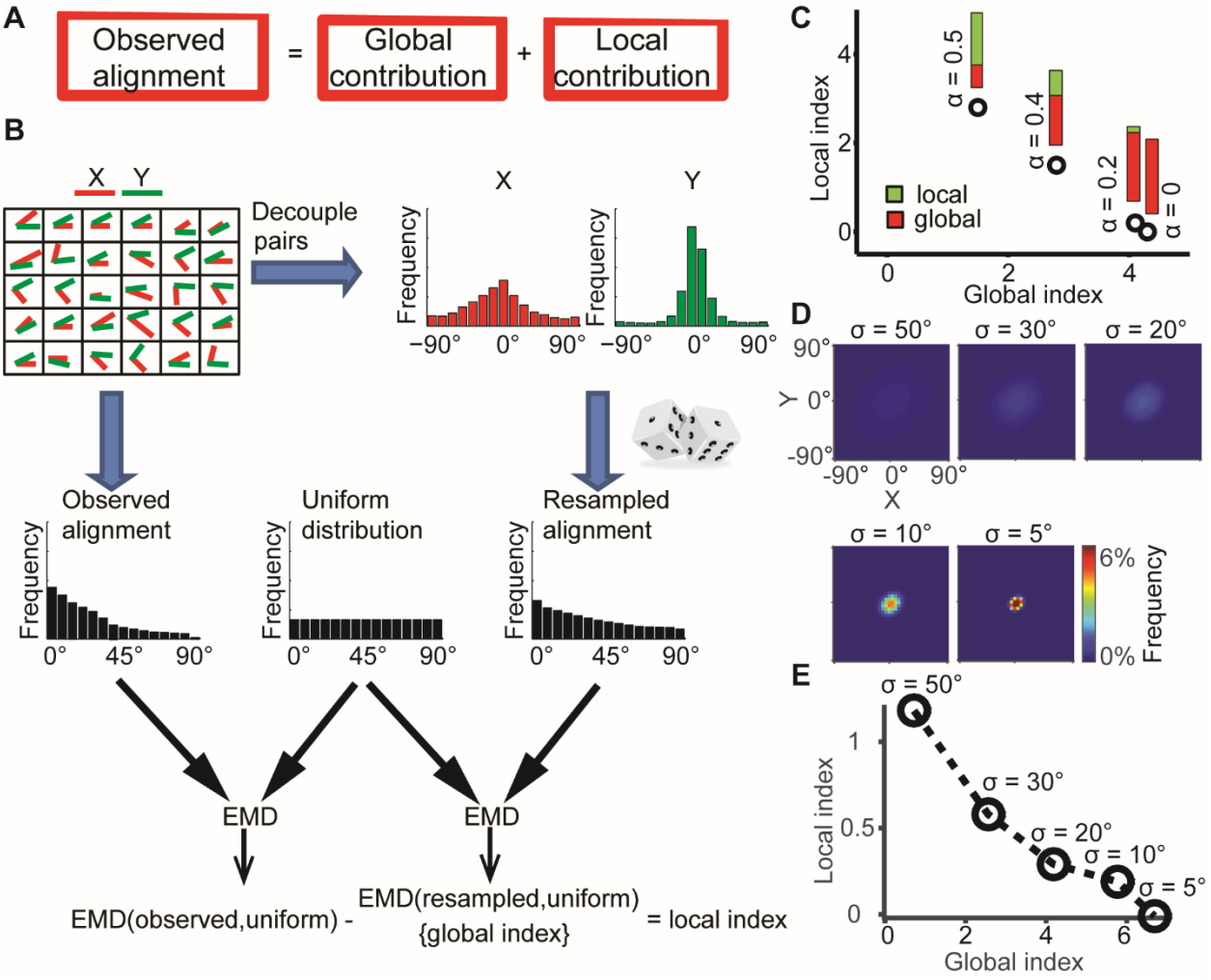


Figure 2: DeBias: local and global indices. (A) Underlying assumption: the observed relation is a cumulative process of a global bias and a local interaction component. (B) Quantifying local and global indices: sample from the marginal distributions X, Y to construct the resampled distribution. The global index (GI) is calculated as the Earth Movers Distance (EMD) between the uniform and the resampled distributions. The local index (LI) is calculated as the subtraction of the GI from the EMD between the uniform and the observed distribution. (C) Local and global indices calculated for the examples from Fig. 1C. Black circles represent the (GI,LI) value for the corresponding example in Fig. 1C, bars represent the relative contribution of the local (green) and global (red) index to the observed alignment. (D-E) Simulation results with a constant interaction parameter $\alpha = 0.2$ and varying standard deviation of X, Y, $\sigma = 50°$ to $5°$. (D) Joint distributions. Correlation between X and Y is (subjectively) becoming less obvious for increasing global bias (decreasing σ). (E) GI and LI are negatively correlated: decreased σ enhances

GI and reduces LI. The change in GI is ~4 fold larger compared to the change in LI indicating that the GI has a limited effect on LI values.

The Earth Mover's Distance (EMD) (Peleg et al., 1989; Rubner et al., 2000) was used as the distance metric to calculate dissimilarities between distributions. The EMD of 1-dimensional distributions is defined as the minimal cost to transform one distribution into the other (Kantorovich and Rubinstein, 1958). This cost is proportional to the minimal accumulated number of moving observations to adjacent histogram bins needed to complete the transformation. Formally, we calculate $EMD(A,B) = \sum_{i=1,..,K} | \sum_{j=1,...,i} a_j - \sum_{j=1,...,i} b_j |$, with a straight-forward implementation for 1-dimensional distributions. Introducing the EMD defines scalar values for the dissimilarities and allows us to define the EMD between resampled and uniform alignment distributions as the *global index* (GI) and the difference between EMD between observed and uniform and the GI as the *local index* (LI). Fig. 2C, demonstrates how the GI and LI recognize the global bias and local interactions between the matched variable pairs (x'_i, y'_i) established in Fig. 1C. For a scenario with no local interaction ($\alpha = 0$) DeBias correctly reports a LI of ~0 and a GI ~3. For a scenario with gradually wider distributions X,Y, i.e., less global bias, and gradually stronger local interactions ($\alpha > 0$), the LI increases while the GI decreases.

In the previous illustration, changes in spread of the distributions X and Y were compensated by changes in the local interactions in order to indicate that similar orientational alignments between variables can arise from significantly different influences from global and local cues. However, leaving the interaction parameter α constant while changing the spread of X and Y revealed a relatively weak, intrinsically negative correlation between LI and GI: changing σ prominently alters the GI, but has a secondary effect on the LI (Fig. 2D-E). Another approach to measure

local interactions, termed Mutual information (Cover and Thomas, 2012), suffered from the same problem (Supplementary Fig. S1). Thus, while the DeBias framework can correctly distinguish between scenarios with substantial shifts from global bias to local interactions, the precise numerical values estimating the contribution of local interactions (LI or mutual information) varies between scenarios with a low versus high global bias. Therefore, we suggest that for a particular set of experiments the biological variation between experiments should be exploited to model the relation between LI and GI in order to accurately estimate the local interaction. This procedure is demonstrated in one of the following case studies.

Theoretical results, limiting cases and assessment with synthetic data

To obtain intuition on the basic properties of the DeBias approach we used theoretical statistical reasoning. To simplify notations for this section, we define X, Y as the marginal distributions of the observed paired variables. The first limit is set by the case in which observations from X and Y are independent. The expected values of the observed and resampled alignments are identical; accordingly, LI converges to 0 for large N (Method: Theory, Theorem 1). The second limit is set by the case in which X and Y are both uniform. The corresponding resampled alignment is also uniform; accordingly, GI converges to 0 for a large N (Method: Theory, Theorem 2). The third limiting case occurs with perfect alignment, i.e., $x_i = y_i$ for all i. In this case the observed alignment distribution is concentrated in the bin containing $\theta = 0$. We examine two opposite scenarios of perfect alignment: (1) When all the local matched measurements are identical ($x_i = y_j$ for all i, j), the resampled distribution is also concentrated in the bin $\theta = 0$ implying that LI = 0 and GI has maximal possible value: $GI = \frac{1}{K}\sum_{i=1,..,K}(i-1) = \frac{K-1}{2}$, where K is the number of quantization bins (Method: Theory, Theorem 3.I). (2) When X, Y are uniform (and $x_i = y_i$ for all i), the resampled distribution is uniform, thus GI = 0 and LI reaches its maximum value:

$LI = \frac{1}{K}\sum_{i=1,..,K}(i-1) = \frac{K-1}{2}$, Method: Theory, Theorem 3.II). Generalizing this limiting case, we prove that LI is a lower bound for the actual contribution of the local interaction to the observed alignment (Method: Theory, Theorem 4). Complementarily, GI is an upper bound for the contribution of the global bias to the observed alignment.

Last, we show that when X and Y are truncated normal distributions, or when the alignment distribution is truncated normal, GI reduces to a limit of 0 as $\sigma \rightarrow \infty$, when σ is the standard deviation of the normal distribution before truncation (Method: Theory, Theorem 5). Simulations complement this result demonstrating that σ and GI are negatively associated, i.e., GI decreases with increasing σ (Fig. 2E). This final property is intuitive, because resampling from more biased distributions (smaller σ) tends to generate high agreement between (x_i, y_i) leading to reduced alignment angles and increased GI.

The modeling of the observed alignment as the sum of GI and LI allowed us to assess the performance of DeBias from synthetic data. By using a constant local interaction parameter ζ (ζ = c), we were able to retrieve the portion of the observed alignment that is attributed to the local interaction and to compare it with the true predefined ζ (Methods, Supplementary Fig. S2). These simulations demonstrated again the GI-dependent interpretation of LI (first shown in Fig. 2E). Simulations were also performed to assess how the choice of the quantization parameter K (i.e., number of histogram bins) and number of observations N affect GI and LI (Methods, Supplementary Fig. S3-S4). In summary, by combining theoretical considerations and simulations we demonstrated the properties and limiting cases of DeBias in decoupling paired matching variables.

Local alignment of Vimentin and Microtubule filaments

We applied DeBias to investigate the degree of alignment between vimentin intermediate filaments and microtubules in polarized cells. Recent work in our lab analyzed single cells in a wound healing assay to suggest that vimentin provides a structural template for microtubule growth, and through this maintains cell polarity (Gan, Ding and Burckhardt et al., in review). The effect was strongest in cells at the wound front where both vimentin and microtubule networks collaboratively align with the direction of migration. An open question remains as to how much of this alignment is caused by the extrinsic directional bias associated with the general migration of cells into the wound as opposed to a local interaction between the two cytoskeleton systems.

Wound healing experiments were performed using Retinal Pigment Epithelial (RPE) cells that express endogenous levels of fluorescently tagged vimentin and α-tubulin. A second set of experiments was performed in cells transfected with an shRNA construct against vimentin. We confirmed knock-down by ~75% and showed impairment of cell polarity maintenance (Gan, Ding and Burckhardt et al., in review). Cells were fixed and stained 90 minutes after scratching the monolayer and images of the vimentin and microtubule filaments were acquired. The filaments of both vimentin and microtubule networks were extracted by computational image analyses. Subsequently, the filaments of the two networks were spatially matched to generate a list of pairwise corresponding orientations (Methods).

Visual inspection suggested that WT cells at the wound edge were polarized, whereas the polarity seemed to decay in cells 2-3 rows inside the cell monolayer sheet. Vimentin-depleted cells looked less polarized also at the wound edge (Fig. 3A). These observations were accompanied by lower alignment between vimentin and microtubule filaments in WT cells away from the wound edge and between the residual vimentin filaments and microtubules in vimentin-

depleted cells at the wound edge (Fig. 3B). Analysis of the GI and LI revealed that most of the discrepancy in vimentin-microtubule alignment between these cases originated from a shift in the global bias (Fig. 3C), suggesting that the local interaction between the two cytoskeletons is unaffected by the cell position or knock-down of vimentin. Instead, the reduced alignment between the two cytoskeletons is caused by a loss of cell polarity in cells away from the wound edge, probably associated with the reduced geometric constraints imposed by the wound edge. In a similar fashion, reduction of vimentin expression relaxes global cell polarity cues that tend to impose alignment.

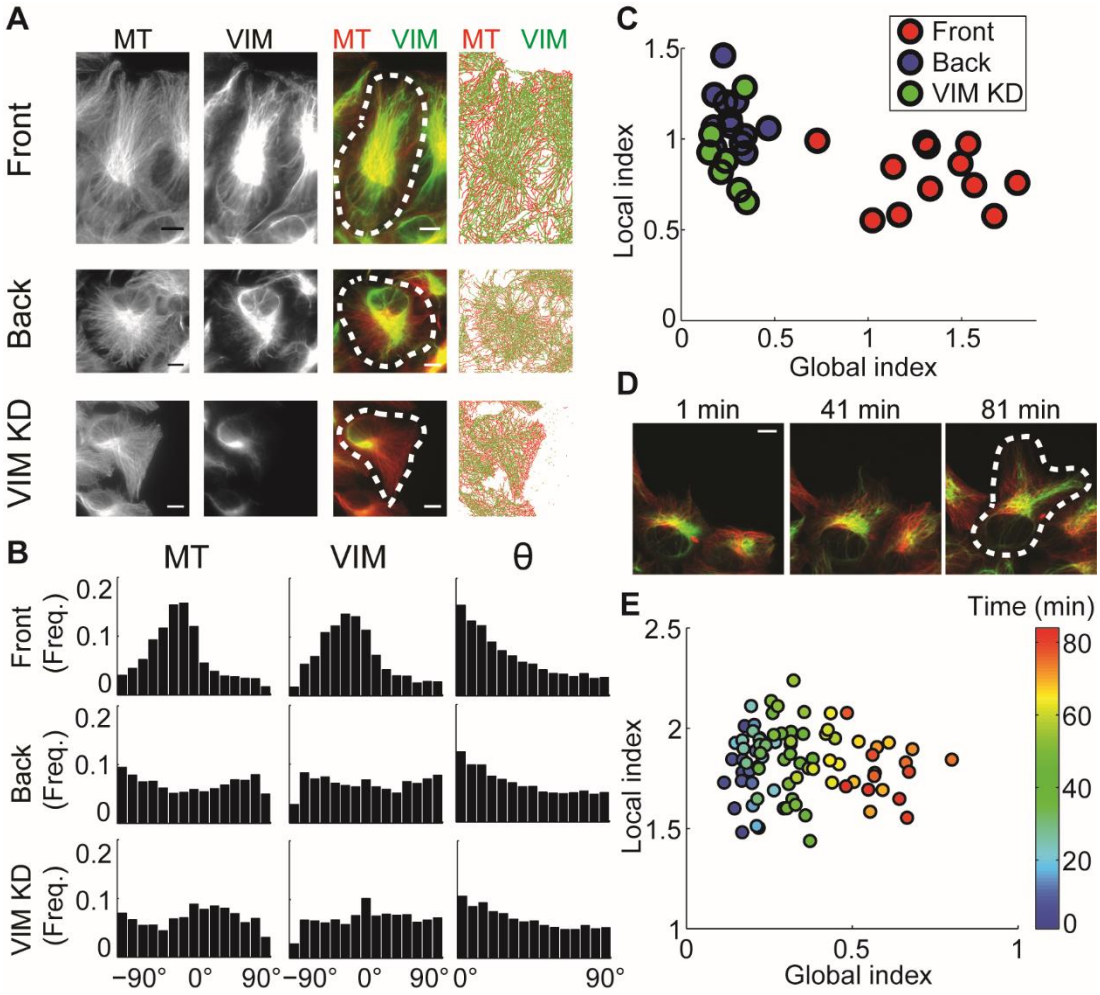


Figure 3: Alignment of microtubule and vimentin intermediate filaments in the context of cell polarity. (A) RPE cells expressing TagRFP α-tubulin (MT) and mEmerald-vimentin (VIM) at endogenous levels.

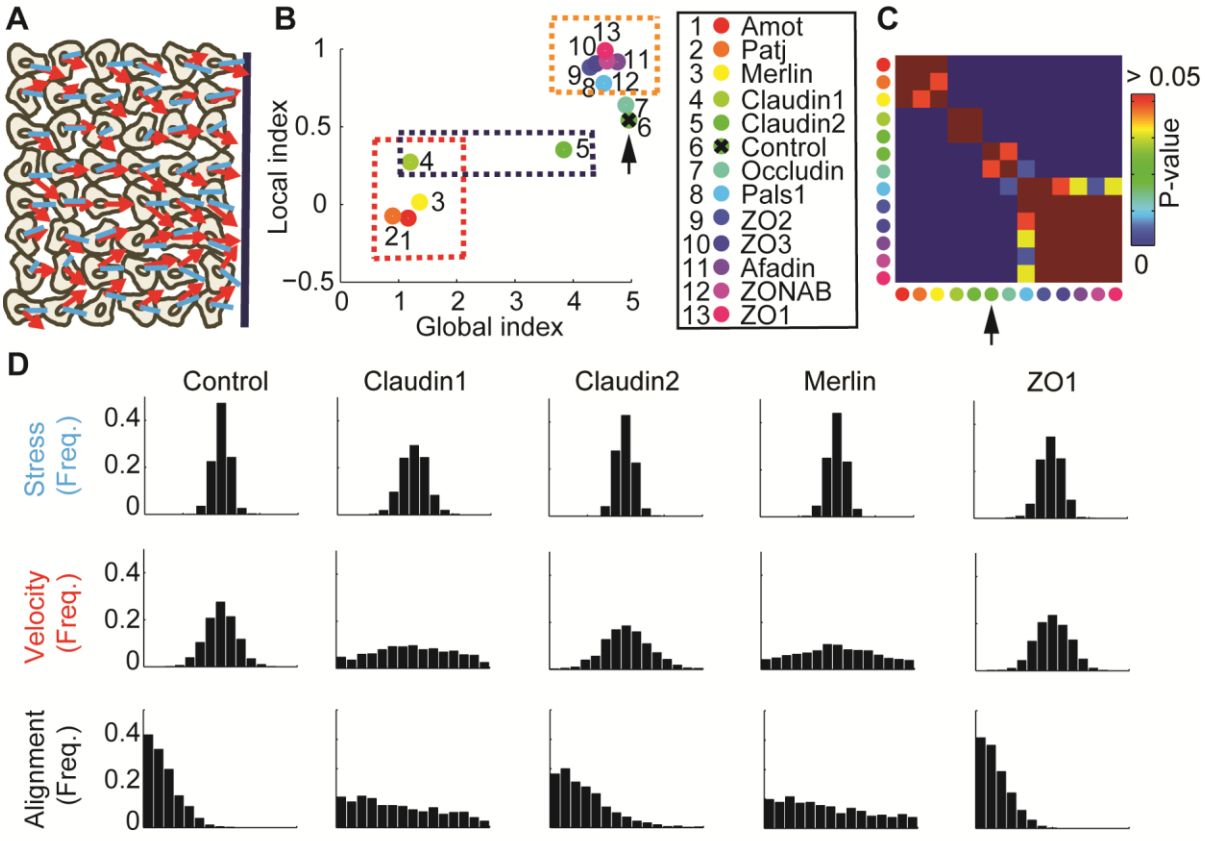
Right-most column, computer segmented filaments of both cytoskeleton systems. Top row, cells located at the wound edge ('Front'); Middle row, cells located 2-3 rows away from the wound edge ('Back'); Bottom row, cells located at the wound edge partially with shRNA knock-down of vimentin. Images were acquired 90 minutes after scratching. Scale bar 10 μm. (B) Orientation distribution of microtubules (left) and vimentin filaments (middle). Vimentin-microtubule alignment distributions (right). (C) Scatterplot of GI vs LI derived by DeBias. The GI is significantly higher in WT cells at the wound edge ('Front', n = 12) compared to cells inside the monolayer ('Back', n = 12, fold change = 4.8, $p < 0.0001$) or to vimentin-depleted cells at the wound edge ('VIM KD', n = 7, fold change = 5.2, $p < 0.0001$). Statistics based on Wilcoxon rank sum test. (D) Polarization of RPE cells at the wound edge at different time points after scratching. Scale bar 10 μm. (E) Representative experiment showing the progression of LI and GI as function of time after scratching (see color code). Correlation between GI and time ~0.90, $p < 10^{-30}$ (n time points = 83). N = 5 independent experiments were conducted of which 4 experiments showed a gradual increase in GI with increased observed polarity.

To corroborate our conclusion that the global state of cell polarity is encoded by the GI, we performed a live cell imaging experiment, in which single cells at the edge of a freshly inflicted wound in a RPE monolayer were monitored for 80 minutes after scratching. The same image analysis pipeline as described above for fixed-cell data was used to extract matching vimentin and microtubule filament orientations in each time point of the movie (frames were sampled at an interval of 1 minute) and DeBias was applied to calculate a time sequence of LI and GI. Cells at the wound edge tended to gradually increase their polarity and started migrating during the imaging time frame (Fig. 3D, Supplementary Video S1). Accordingly, the GI increased over time (Fig. 3E). This demonstrates the capacity of DeBias to distinguish fundamentally different effectors of cytoskeleton alignment.

Identifying new molecular players in alignment of cell velocity and mechanical forces during collective cell migration

Collective cell migration requires intercellular coordination, achieved by mechanical and chemical information transfer between cells. One mechanism for cell-cell communication is plithotaxis, the tendency of individual cells to align their velocity with the maximum principal

stress orientation (He et al., 2015; Tambe et al., 2011; Zaritsky et al., 2015). As in the previous example of vimentin and microtubule interaction, much of this alignment may be associated with a general directionality of velocity and stress field parallel to the axis of collective migration rather than a local coupling of velocity and stress. Recently, applying an early version of the DeBias approach, we decoupled the local and global contribution to velocity-stress alignment and found, indeed, that the global contribution plays the predominant role in inducing motion-stress alignment. However, a smaller, local component (plithotaxis) exists, and plays a crucial role in cell-cell information transfer underlying collective migration (Zaritsky et al., 2015).



	Control	Claudin1	Claudin2	Merlin	ZO1
Align.	11.78°	36.25°	19.77°	36.73°	11.55°
LI	0.54	0.27	0.35	0.02	0.99
GI	4.96	1.20	3.84	1.36	4.55

Figure 4: Alignment of stress orientation and velocity direction during collective cell migration. (A) Assay illustration. Wound healing assay of MDCK cells. Particle image velocimetry was applied to

calculate velocity vectors (red) and monolayer stress microscopy to reconstruct stresses (blue). Alignment of velocity direction and stress orientation was assessed. (B) Mini-screen that includes depletion of 11 tight-junction proteins and Merlin. Shown are GI and LI values, legend is sorted by the LI values (control is ranked 6^{th}, pointed by the black arrow). Each dot was calculated from accumulation of 3 independent experiments (N = 925-1539 for each condition). Three groups of tight junction proteins are highlighted by dashed rectangles: red - low LI and GI compared to control, purple – different GI but similar LI, orange – high LI. Data from (Das et al., 2015), where effective depletion was demonstrated. (C) Pair-wise statistical significance for LI values. P-values were calculated via a permutation-test on the velocity and stress data (Methods). Red – none significant ($p \geq 0.05$) change in LI values, blue – highly significant (< 0.01) change in LI values. Conditions color coding as in panel B. (D) Highlighted hits: Claudin1, Claudin2, Merlin and ZO1. Top: Distribution of stress orientation (top), velocity direction (middle) and motion-stress alignment (bottom). Bottom: table of mean alignment angle, LI and GI. Claudin1 and Claudin2 have similar mechanisms for transforming stress to aligned velocity. ZO1 depletion enhances alignment of velocity by stress.

Using a wound healing assay, Das et al. (Das et al., 2015) screened 11 tight-junction proteins to identify molecular components and pathways that promote motion-stress alignment (Fig. 4A). Knockdown of Merlin, Claudin1, Patj and Angiomotin (Amot) reduces the alignment of velocity direction and stress orientation (Das et al., 2015). Further inspection of these hits showed that the stress orientation remains stable upon depletion of these proteins, but the velocity direction distribution is much less biased towards the wound edge (Zaritsky et al., 2015). Here, we further analyze this data to demonstrate the capacity of DeBias to pinpoint tight-junction proteins that alter specifically the global or local components that induce velocity-stress alignment.

By distinguishing GI and LI we generate a much more refined annotation of the functional alteration that depletion of these tight-junction components causes in mechanical coordination of collectively migrating cells (Fig. 4B-C). First, we confirm that the four hits reported by (Das et al., 2015) massively reduce the GI, consistent with the notion that absence of these proteins diminished the general alignment of velocity to the direction induced by the migrating sheet (Fig. 4B, red dashed rectangle). Merlin, Patj and Angiomotin reduced the LI to values close to 0, suggesting that the local dependency between stress orientation and velocity direction was lost.

Depletion of Claudin1, or of its paralog Claudin2, which was not reported as a hit in the Das et al. screen, reduced the LI to a lesser extent, similarly for both proteins, but had very different effects on the GI (Fig. 4B, purple dashed rectangle). This suggests that the analysis by (Das et al., 2015) missed effects that do not alter the general alignment of stress or motion, and implies the existence of a local velocity-stress alignment mechanism that does not immediately change the collective aspect of cell migration but may have implications on the mechanical interaction between individual cells.

When assessing the marginal distributions of stress orientation and velocity direction we observed that depletion of Claudin1 reduced the organization of stress orientations and of velocity direction, while Claudin2 reduced only the latter while maintaining similar LI (Fig. 4D). Merlin depletion is characterized by an even lower LI and marginal distributions with aligned stress orientation and almost uniform alignment distribution (Fig. 4D). Since we think that aligned stress is transformed to aligned motion (He et al., 2015; Zaritsky et al., 2015), we speculate that the LI quantifies the effect of local mechanical communication on parallelizing the velocity among neighboring cells and thus suggest that this stress-motion transmission mechanism is impaired to a similar extent by reduction of Claudin1 and Claudin2, albeit less than by reduction of Merlin.

Using LI as a discriminative measure also allowed us to identify a group of new hits (Fig. 4C). ZO1, ZO2, ZO3, Occludin and ZONAB are all characterized by small reductions in GI but a substantial increase in LI relative to control (Fig. 4B, orange dashed rectangle). A quantitative comparison of control and ZO1 depleted cells provides a good example for the type of information DeBias can extract: both conditions yield similar observed alignment distributions with nearly identical means, yet ZO1 depletion has an 83% increase in LI and 8% reduction in

GI, i.e., the mild loss in the marginal alignment of velocity or stress is compensated by enhanced local alignment in ZO1 depleted cells (Fig. 4D). This might point to a mechanism in which stress orientation is reduced by tight-junction depletion, but enhanced by transmission of stress orientation into motion orientation, leading to comparable alignment. Notably, all paralogs, ZO1, ZO2 and ZO3 fall into the same cluster of elevated LI and slightly reduced GI relative to control experiments. This phenotype is in agreement with the outcomes of a screen that found ZO1 depletion to increase both motility and cell-junctional forces (Bazellières et al., 2015).

Analysis of recruitment of transmembrane receptors to clathrin coated pits during clathrin-mediated endocytosis

Assessing protein-protein colocalization is ubiquitous example of correlating spatially matched variables in cell biology. To quantify GI and LI for protein-protein colocalization data we normalized each channel to the [0,1] range and the alignment θi of matched observations (x_i,y_i) then referred to the difference in normalized fluorescent intensities $x_i - y_i$ (Methods).

In the following, we demonstrate how the decoupling of global and local contributions to the overall intensity alignment of fluorescent channels improves the sensitivity of colocalization analysis in clathrin-mediated endocytosis (CME). In CME receptors and ligands are recruited to clathrin-coated pits (CCPs) via interaction with adaptor proteins (Traub, 2009). We exploited established knowledge on the recruitment of transmembrane receptors to CCPs to (1) validate DeBias capabilities and (2) demonstrate that the GI is complementary to correlation measurements for more accurate assessment of the local interaction in colocalization analysis. Together, DeBias provides enhanced discriminatory capabilities between different degrees of protein-protein interaction.

We used immunofluorescence images of fixed RPE cells overexpressing a fluorescent fusion of clathrin light chain (CLC, as a CCP marker), previously shown to preserve CME frequency and dynamics of CCP formation (Aguet et al., 2013). We used Total Internal Reflection Fluorescence microscopy (TIRFM) to image cells in which CCPs were reported in the EGFP-CLC channel (Aguet et al., 2013; Loerke et al., 2009) and specific cargo molecules were visualized in a second channel. For single cells, the location of fluorescent signals of CLC and the cargo molecule were recorded and the data was pooled and processed by DeBias (Methods).

The transferrin receptor (TfnR) is a well-studied CME-dependent cargo and was used here for validation of DeBias. It has been previously shown that TfnR accumulation in CCPs is ligand independent (Lamaze et al., 1993). The recruitment of TfnR into CCPs requires interaction of its tyrosine-based internalization motif YTRF, with the heterotetrameric adaptor protein AP2. By mutating key residues in the μ subunit of the AP2 complex (Höning et al., 2005), we have established an AP2 mutant cell line (denoted AP2μ^{cargo-}) that is deficient in TfnR recruitment into CCPs. By subjective visual assessment, we observed colocalization of fluorescently labeled transferrin ligand with EGFP-CLC punctate in control cells, and, as expected, a diffuse pattern of transferrin in AP2μ^{cargo-} (Fig. 5A). Quantitatively, the LI values of EGFP-CLC-TfnR in WT cells (2.4 ± 1.04) were over 7 fold higher than those observed for the mutated cells (0.31 ± 0.51) (Fig. 5B-C).

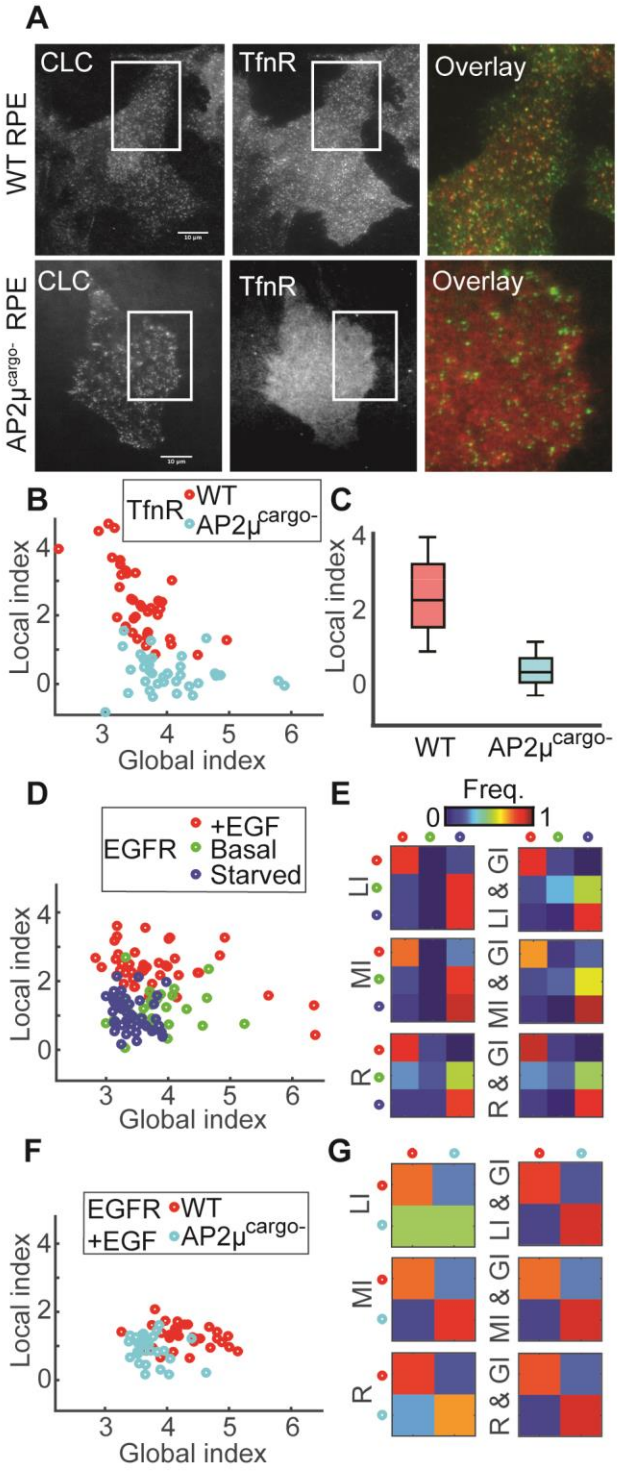


Figure 5: Recruitment of transmembrane receptors to CCPs during CME. (A) RPE cells expressing CLC and TfnR ligands. Top row, representative WT cell (TfnR ligand, GI = 4.8, LI = 2.3). Bottom row, representative AP2μ^{cargo-} mutated cell (TfnR ligand, GI = 3.3, LI = 0.3). Scale bar 10 μm. (B) LI and GI of CLC-TfnR colocalization for WT (red, number of cells N = 38, number of CCPs M = 71,112) and

AP2μ^{cargo-} mutated cells (cyan, N = 38, M = 25,164). Every point represents a single cell. (C) Boxplot of WT vs. AP2μ^{cargo-} cells ($p < 10^{-12}$, Wilcoxon rank sum test). (D) LI and GI for CLC-EGFR for 3 experimental conditions: with EGF (red, N = 44, M = 74,376); basal, with serum that includes some growth factor (green, N = 30, M = 44,782); in starvation media (blue, N = 44, M = 62,731). LDA single-cell classification accuracy was 74% ($p < 0.001$, using bootstrapping against the random classifier, Methods) and was significant between any two conditions (+EGF vs. Starved: $p < 0.001$, Basal vs. Starved: $p < 0.001$, +EGF vs. Basal: $p = 0.003$). (E) Using the GI enhanced classification accuracy. Confusion matrix of single cell classification for alternative colocalization measures. LDA was trained for the 3 conditions of EGFR recruitment to CCPs presented in panel D. Bin (*i*,*j*) record the fraction of cells from treatment *i* that were classified as from treatment *j*. Red circle: +EGF, blue circle: basal cells (with serum), blue circles: starved cells (no serum). Comparison of single cell prediction accuracy with (top) versus without using the GI (left-to-right): Local index (**74%** vs. 64%, $p = 0.001$, bootstrapping, Methods), Mutual information (67% vs. 62%, $p = 0.023$), Pearson's Rho (73% vs. 69%, $p = 0.01$). (F) Recruitment of EGFR to CCPs is AP2-dependent. LI and GI for CLC-EGFR in EGF-treated cells: WT (red, N = 32, M = 28,468) versus AP2μ^{cargo-} (cyan, N = 28, M = 27,228). LDA single-cell classification achieved 75% accuracy ($p < 0.001$). (G) Confusion matrices for single cell classification. Red: +EGF, cyan: AP2μ^{cargo-}+EGF. Single cell prediction accuracy for using (top) versus not using (bottom) the GI (left-to-right): Local index (**85%** vs. 63%, $p < 0.001$), Mutual information (82% vs. 80%, $p = 0.27$), Pearson's Rho (83% vs. 77%, $p = 0.047$).

To test the pair (GI, LI) as a 2-dimensional measure of colocalization we applied Linear Discriminative Analysis (LDA) classification (Methods). Briefly, each cell's (GI,LI) values were labeled based on their experimental condition as either WT or AP2μ^{cargo-}, a LDA-linear model was trained based on this data and the model cell-classification accuracy and statistical significance was recorded. For the WT versus AP2μ^{cargo-} experiment, LDA reached classification accuracy of 88% ($p < 0.001$, bootstrapping, against the null hypothesis of arbitrary classification, Methods) providing validation for DeBias' capabilities to distinguish between obvious alterations in colocalization.

In contrast to the TfnR, the epidermal growth factor receptor (EGFR) is recruited to CCPs only after stimulation with EGF at nanomolar concentrations (Sigismund et al., 2008). The amount of TfnR in CCPs is proportional to CLC (Taylor et al., 2011); however, whether the same is true for EGFR and if this is proportional to the amplitude of stimulation is not known. Therefore, we tested CCP/EGFR association in wild type cells for 3 stimuli-dependent conditions: (1) cells

starved for 1h at 0.1% FBS, (2) grown under basal conditions, (3) stimulated for 5 min with 20 ng/*ml* of EGF. EGFR was detected by a neutral antibody that does not interfere with EGF binding and hence does not pose a stimulatory or inhibitory effect (Chung et al., 2010; Kawashima et al., 2010). Discrimination between these different conditions appears more challenging (Fig. 5D), and thus we exploited the heterogeneity in GI values and the negative association between LI and GI values (Fig. 5B, D, as simulated in Fig. 2E) to improve the sensitivity of our assay.

A dose-dependent response in (GI,LI) was observed for the experimental conditions: the (GI,LI)-trend of EGF-treated cells was located above those with basal-level stimulation or in starvation (Fig. 5D). This observation suggests that using the scalar LI as the discriminative measure will be less effective than a model based on the (GI,LI) pair that can correct for the influence of the global bias on LI. This was validated by applying LDA classification which achieved 74% accuracy in classification of single cells ($p < 0.001$, legend Fig. 5D). LDA was also able to distinguish between any two of the experimental conditions (legend Fig. 5D) suggesting that EGFR recruitment to CCPs is proportional to EGFP-CLC in an EGF-dose dependent manner.

Next we demonstrate that the GI encodes complementary information to local measures that allows a more accurate assessment of colocalization. LDA classification was applied to assess the classification accuracy for different colocalization measures, scalar or paired with the GI, validating that enhanced discrimination between experimental conditions is achieved by pairing the GI to various local measures (Fig. 5E). The results are displayed in a confusion matrix (Methods): Each row represents the LDA classification of cells (colored circles). Each column represents the cells' true condition. Thus, the bin at row *i* and column *j* hold the fraction of cells from condition *j* that were classified as originating from condition *i*. Accordingly, the values of

the top-left to bottom-right diagonal represent classification accuracy for each of the three conditions – higher values in the diagonal (and low values elsewhere) reflect better prediction accuracy (Fig. 5E). For example, when comparing LI to the pair (GI,LI), the correct classification accuracy of basal cells increased by 30% (Fig. 5E, top-left), while the misclassification of basal to starved cells decreases by 33% (Fig. 5E, top-right). We compared classification accuracy of LI and two widely used scalar measures (i.e., 1-dimensional numerical measures) to their pairing with the GI (Fig. 5E, left versus right). Local index, mutual information (denoted MI) or Pearson's correlation coefficient (denoted R) classification accuracy was significantly improved by the 2-dimensional pairing with the GI, proving enhanced sensitivity (legend Fig. 5E).

It has been shown that multiple, concomitant mechanisms regulate EGFR endocytosis (Goh et al., 2010; Sigismund et al., 2008; Villaseñor et al., 2015). Biochemical studies revealed that the EGFR internalization motif YRAL is required for direct EGFP/AP2 interaction (Goh et al., 2010). However, in intact cells, mutating this internalization motif does not seem to affect EGFR endocytosis upon EGF stimulation (Goh et al., 2010; Nesterov et al., 1999). To resolve this discrepancy, we applied DeBias to test whether EGFR affinity is reduced for CCPs in AP2μ^{cargo-} cells. The (GI, LI) scatter plot showed a tendency of the WT cells to cluster above the trend of (GI,LI) of the AP2μ^{cargo-} cells (Fig. 5F). LDA-based single cell classification accuracy reached 75% ($p < 0.001$), implying that EGFR recruitment to CCPs is AP2- and hence YRAL-motif dependent. The absence of a measurable shift in cargo uptake between these two conditions thus may be explained by higher initiation density of clathrin coated pits upon EGF stimulation and/or development of compensatory internalization pathways. Comparison of scalar colocalization measures to paired measurements demonstrated again that by accounting for GI, single cell

classification accuracy surpasses scalar-based measures (Fig. 5G). Notably, the combination (GI,LI) outperforms all alternatives (legend Fig. 5E, G).

Together, we conclude that accounting for GI values and their cell-to-cell variation enables a more precise estimation of the local interaction to ultimately better distinguish between functional states of single cells.

Discussion

We introduce DeBias as a new method to assess global bias and local interactions between spatially matched variables in biased biological systems. A software package implementation is publically available via a web-based platform implementation, https://debias.biohpc.swmed.edu. The website also provides detailed instruction for the operation of the user interface. The distinction of global and local contributions to the level of coupling of two jointly observed variables addresses two issues that are often neglected in the interpretation of the coupling of cell biological variables: (1) Global bias is an additional, and often dominant, mechanism conferring agreement between variables (Figs. 1-2), (2) The level of local interactions is strongly influenced by the global bias; thus, to interpret local interactions they have to be analyzed in the context of the global bias (Fig. 2D-E). To accomplish such coupled interpretation we propose a two-dimensional measure for colocalization instead of the scalar measures. Using this framework, we demonstrated that cell polarity is the main factor behind the observed alignment of vimentin and microtubule filaments and that their local interaction is polarity-independent (Fig. 3); that some tight junction proteins mediate oriented stress, while others mediate the transmission of stress to aligned velocity (Fig. 4); and that the recruitment of EGFR to CCPs is AP2-dependent (Fig. 5).

Also, we demonstrated that the GI improves sensitivity over alternative correlation-based colocalization measures (Fig. 5). Together, DeBias provides a powerful tool to study molecular colocalization or other coupled measurements, as a preliminary step for more direct assays of interactions (e.g., (Clegg, 1995; Langer-Safer et al., 1982; Piston and Kremers, 2007; Schwille et al., 1997)).

Alternative quantification methods

Many methods have been developed to quantify protein-protein colocalization. These are traditionally divided into pixel-based and object-based methods. Pixel-based methods measure pixel-wise correlation coefficients (Adler and Parmryd, 2010; Bolte and Cordelieres, 2006; Costes et al., 2004; Manders et al., 1993; Pearson, 1901), exploiting the notion that fluorescent levels of colocalized proteins are correlated, but suffering from background noise. Object-based methods first detect objects of interest and then assess colocalization based on second-order statistics of the spatial distributions of the detections (Helmuth et al., 2010; Kalaidzidis et al., 2015; Lagache et al., 2015; Rizk et al., 2014). Object-based methods remove noise in colocalization attributed to background pixels but lose the information contained by the fluorescence levels at the detected objects. Thus, object-based methods are best applicable for the colocalization of binary signals, but not for applications in which colocalization accounts for coupling of molecular counts on a continuous spectrum. Moreover, object-based methods require detection of objects in both channels, which often limits their applicability. In our examples of receptor-CCP colocalization we demonstrated a hybrid of the two approaches: colocalization analysis by DeBias is focused on the intensity of fluorescent readouts within detected CCPs. Importantly, DeBias could be applied with the same advantages to two-channel images of more diffusely localized proteins.

An important step in revealing local interactions masked by global biases was recently made by (Krishnaswamy et al., 2014) for applications to single cell mass cytometry data. They developed a measure referred to as DREMI to quantify the influence of a protein X on protein Y based on the conditional probability P(Y|X). DREMI takes advantage of the abundant mass cytometry data to equally weigh data at different intervals along the range of X values using >10,000 cells per experimental condition. This approach is less reliable when limited data is available, because of the low confidence in the conditional probability of observations with low data abundance. Thus, DREMI is not well suited for image data, which typically has a limited volume of observations.

Limitations of DeBias

LI and GI are measures of association, as are the vast majority of colocalization measures. Thus, they cannot predict causality between the variables. For example, in our analysis of the VIM-MT alignment (Fig. 3), we cannot conclude whether vimentin aligns microtubules or vice versa. To determine such hierarchical relations, asymmetric (Krishnaswamy et al., 2014) or time-resolved measurements (Welf and Danuser, 2014) are necessary. A second limitation occurs with molecular platforms and scaffolds that can induce a pseudo-correlation between two variables although they interact with the platform via independent mechanisms. One example is intracellular spatial patterns defined by subcellular compartments that may induce correlations without interactions between variables such as the nuclear speckles, adhesion sites or endosomes. This is a general problem in colocalization analysis that should be addressed by other means (Costantino et al., 2005; Wu et al., 2010).

Sources of global bias in protein-protein colocalization

Global bias in protein-protein colocalization (Fig. 5) is reflected in the difference of the two fluorescent intensity patterns of the two proteins' marginal distributions from a uniform distribution. The origins of such global bias is less intuitive to grasp than with co-aligned orientational variables, where global bias is induced, for example, by cell polarity (Fig. 3) or by a preferred direction of migration (Fig. 4). To provide some intuition for the global bias in protein-protein co-localization we use again the example of CME: the number of clathrin molecules recruited to CCPs is associated with CCP size (Ehrlich et al., 2004). Overall, the size distribution of CCPs is non-uniform, i.e., a global bias exists, which introduces correlation between the abundances of two proteins recruited to CCPs even when they do not interact. This is illustrated in Fig. 6, where we show a small population of small pits, a large population of mid-sized pits, and again a small population of large pits. Hence, the marginal protein distributions peak at values that correspond to these CCP sizes. The resampled distribution deviates from the uniform distribution, which increases the GI and renders the LI hard to interpret. Local interactions between a pair of molecules are recognized by co-fluctuations about their mean abundances. However, without accounting for the global effects, these behaviors are masked. In Fig. 6 we illustrate this phenomenon by observations where one protein deviates from the size-defined mean abundance, and due to local interactions, the paired protein tends to fluctuate in the same direction (red arrows, left column). Such co-fluctuations are absent in an example with no local interaction (right column). Accordingly, the fluctuations encoded by the LI allow the distinction between different experimental conditions that enhance / reduce the local interactions.

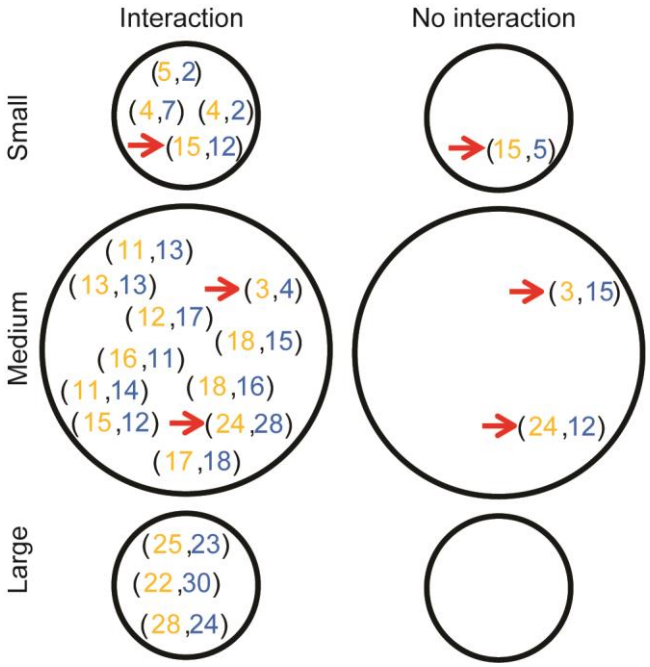


Figure 6: One possible source for global bias in protein-protein colocalization. CCP sizes are small, medium or large. Most structures are of medium size. Paired numbers (orange, blue) represent molecular counts in a CCP. Protein recruitment to structures is associated with the latter's size: ~0-10 for small structures, ~10-20 for medium structures and ~20-30 for large structures. Thus, non-uniform CCP size distribution induces global bias in the molecules distribution. Red arrows highlight outliers in the molecular counts of the orange protein that co-fluctuate with the blue protein (left column) indicating direct interaction. Such co-fluctuations are missing when the proteins do not interact (right column, showing only the outliers).

Methods

DeBias procedure

The DeBias procedure is depicted in Fig. 2A. The marginal distributions X and Y are estimated from the experimental data, $\forall i, x_i, y_i \in [0,90°]$. The experimentally observed alignment distribution (denoted *observed*) is calculated from the alignment angles θ_i of matched (x_i,y_i) paired variables, for all i.

$$\theta_i = \begin{cases} |x_i - y_i| & |x_i - y_i| \leq 90 \\ 180 - |x_i - y_i| & |x_i - y_i| > 90 \end{cases}$$

The resampled alignment distribution (denoted *resampled*) is constructed by independent sampling from X and Y. N random observations (where N = |X| is the original sample size) from X and Y are independently sampled with replacement, arbitrarily matched and their alignment angles calculated to define the resampled alignment. This type of resampling precludes the local dependencies between the originally matched (x_i,y_i) paired variables.

The uniform alignment distribution (denoted *uniform*) is used as a baseline for comparison between distributions. This is the expected alignment distribution when neither global bias (reflected by uniform X, Y distributions) nor local interactions exist. The Earth Mover's Distance (EMD) (Peleg et al., 1989; Rubner et al., 2000) was used as a distance metric between alignment distributions. The EMD for two distributions, A and B, is defined as follows:

$EMD(A,B) = \sum_{i=1,..,K} | \sum_{j=1,...,i} a_j - \sum_{j=1,...,i} b_j |$, where a_j and b_j are the frequencies of observations in bins j of the histograms of distributions A and B, respectively, each containing K bins.

The global index (GI) is defined as the EMD between the uniform distribution and the resampled alignment:

GI = EMD(*uniform*,*resampled*)

The local index is determined by subtraction of the global index from the EMD between the uniform distribution and the experimentally observed alignment distribution:

LI = EMD(*uniform*,*observed*) - global index

DeBias for protein-protein colocalization: The following adjustments to this procedure are implemented to allow DeBias to quantify protein-protein colocalization:

1. Levels of fluorescence are not comparable between different channels due to different expression levels and imaging parameters. Thus, each channel is normalized to [0,1] by the 5^{th} and 95^{th} percentiles of the corresponding fluorescence intensities.
2. The alignment angle θ_i of the matched observation (x_i,y_i) is calculated as the difference in normalized fluorescence intensities x_i - y_i and the alignment distribution is thus defined on the interval [-1,1].
3. Cells with GI above a given threshold (set to 7) where excluded from the analysis because extreme global bias values obscure the ability to differentiate between different experimental conditions.

The number of histogram bins used to represent the marginal distributions was 89 for angular data and 39 for colocalization data. The corresponding number of histogram bins for the alignment distributions (observed, resampled and uniform) was 15 and 40, respectively.

Simulating synthetic data

Let us define X and Y as the angular probability distribution functions, with angle instances denoted x_i and y_i, respectively. When simulating local relations, for each pair of angles, one of the angles will be shifted towards the other by ζ degrees, unless $|x_i - y_i| < \zeta$, in which case it will be shifted by $|x_i - y_i|$ degrees. The angle to be shifted (either x_i or y_i) is chosen by a Bernoulli random variable, p, with probability 0.5. The observed angles for pixel i will therefore be

$$x_i' = \begin{cases} x_i & p = 1 \\ \max(x_i - \zeta, y_i) & y_i \le x_i \wedge p = 0 \\ \min(x_i + \zeta, y_i) & y_i > x_i \wedge p = 0 \end{cases}$$

and

$$y_i' = \begin{cases} y_i' & p = 0 \\ \max(y_i - \zeta, x_i) & x_i \le y_i \wedge p = 1 \\ \min(y_i + \zeta, x_i) & x_i > y_i \wedge p = 1 \end{cases}$$

The alignment of angles at pixel i will be:

$$\theta_i = \begin{cases} |x_i' - y_i'| & |x_i' - y_i'| \le 90 \\ 180 - |x_i' - y_i'| & |x_i' - y_i'| > 90 \end{cases}$$

For example, for our simulations we choose X, Y to be truncated normal distributions on $(-90,90)$ with $\mu = 0$ and varying values of σ.

ζ is modeled in two ways: either as a constant value, e.g. $\zeta = 5°$ (Supplementary Data S1), or as a varying value dependent on $|x_i - y_i|$ (Figs. 1-2). For the latter, ζ it is defined as a fraction $0 < \alpha < 1$ from $|x_i - y_i|$ for each pair of observations; namely, $\zeta_i = \alpha|x_i - y_i|$ (see Fig. 1B). Note, that the observed marginal distributions X', Y' may be slightly different from X, Y.

Theoretical results

Terms and definitions: Let X, Y be the distribution functions of two random variables representing angles on $[-90°, 90°]$. Spatially matched random variables from these distributions are denoted x_i and y_i, $i = 1,2 \ldots N$, where N is the number of observations. x_i^* and y_i^* are random variables sampled from X and Y independently (without considering the spatial matching). The observed and resampled alignment distributions are denoted A and A^*, respectively. The alignment distributions represent angles on $[0°, 90°]$, and are functions of X, Y,

the interaction between X and Y, and N. Random variables from A are denoted θ_i, $i = 1,2 \dots N$. Random variables from A^* are denoted θ_i^*. Let K be the number of bins in the alignment histogram. Histogram bins are denoted $bin_i, i = 0, \dots, K-1$ where bin_0 contains the lowest values (including 0°) and bin_{k-1} the highest values (including 90°). U denotes the uniform distribution on $[0°, 90°]$with the same K bins as A, A^*.

Theorem 1: Local index of independent variables

If X, Y are independent, then $E(LI(X,Y)) = 0$

Proof:

$$E(LI) = E(EMD(A,U) - EMD(A^*,U))$$

$$= E(EMD(A,U)) - E(EMD(A^*,U)) \underbrace{=}_{*} E(EMD(A,U)) - E(EMD(A,U)) = 0$$

* Resampling does not change the expectation of the difference of two independent variables.

Theorem 2: Global index of uniform distributions

Let X, Y be uniform distributions, then $\lim_{N\to\infty} GI = 0$.

Proof:

Since X and Y are uniform distributions, the density of the resampled distributions are $f_{x_i^*}(\omega) = f_{y_i^*}(\omega) = \frac{1}{180}, -90 < \omega < 90$. We can compute the distribution of the difference by

$$f_{x_i^*-y_i^*}(\omega) = \int_{-\infty}^{\infty} f_{x_i^*}(x) f_{y_i^*}(x-\omega)\,dx = \int_{-\infty}^{\infty} \frac{1}{180^2} I_{x\in(-90,90)} I_{x-\omega\in(-90,90)}\,dx =$$

$$\begin{cases} \frac{1}{180^2}\int_{-90}^{90+\omega} 1\,dx & 0<\omega<180 \\ \frac{1}{180^2}\int_{-90+\omega}^{90} 1\,dx & 0<\omega<180 \end{cases} = \begin{cases} \frac{1}{180^2}(180+\theta) & -180<\omega<0 \\ \frac{1}{180^2}(180-\theta) & 0<\omega<180 \end{cases}$$

We conclude that the distribution of the difference is

$$f_{x_i^*-y_i^*}(\omega) = \begin{cases} \frac{1}{180}\left(1+\frac{\omega}{180}\right) & -180<\omega<0 \\ \frac{1}{180}\left(1-\frac{\omega}{180}\right) & 0<\omega<180 \end{cases}$$

Therefore, when taking the absolute value of the angle difference we get:

$$f_{|x_i^*-y_i^*|}(\omega) = \begin{cases} \frac{1}{90}\left(1-\frac{\omega}{180}\right) & \omega \geq 0 \\ 0 & \omega<0 \end{cases}$$

Finally, since the alignment is limited to [0°, 90°] we apply the function $g(\omega) = \begin{cases} 180-\omega & 90<\omega\leq 180 \\ \omega & 0\leq\omega\leq 90 \end{cases}$ so that

$$f_{g(|x_i^*-y_i^*|)}(\omega) = \begin{cases} \frac{1}{90}\left(1-\frac{\omega}{180}\right)+\frac{1}{90}\left(1-\frac{180-\omega}{180}\right) & 0\leq\omega\leq 90 \\ 0 & else \end{cases}$$

$$= \begin{cases} \frac{1}{90} & 0\leq\omega\leq 90 \\ 0 & else \end{cases}$$

Therefore $f_{g\left(\left|x_i^*-y_j^*\right|\right)} = A^*$ is a uniform distribution.

For N sufficiently large, the histogram has approximately $\frac{1}{K}$ of the observations in each of the K equally spaced intervals between 0 and 90 and thus $\lim_{N\to\infty} GI = \lim_{N\to\infty} EMD(A^*, U) = 0$.

Theorem 3: Perfect alignment

(I) If $\forall i,j, x_i = y_j$ then $GI = \frac{(K-1)}{2}, LI = 0$

(II) If X and Y are uniform distributions, and $\forall i, x_i = y_i$ then $GI = 0, LI = \frac{(K-1)}{2}$

Proof:

(I)

$A = A^*$ because $\forall i,j\ a_i = a_j^* = 0$. For large N the random variable drawn from the alignment distribution A will be approximately:

$$\theta_i = \begin{cases} 1 & \theta_i \in bin_0 \\ 0 & else \end{cases}, \forall i.$$

The EMD of the alignment from the uniform distribution is therefore simply 'moving' $\frac{1}{K}$ observations from bin_0 to every other bin, which sums up to

$$\lim_{N\to\infty} EMD(A,U) = \frac{1}{K} * 1 + \frac{1}{K} * 2 + \cdots + \frac{1}{K} * (K-1) = \frac{1}{K} * (K-1) * \frac{1+K-1}{2} = \frac{K-1}{2}$$

Therefore, for large N $LI = EMD(A,U) - EMD(A^*,U) = 0$, $GI = EMD(A^*,U) = \frac{K-1}{2}$.

(II)

Since $\forall i, x_i = y_i$ we get that, similarly to part (I), $EMD(A,U) = \frac{K-1}{2}$.

On the other hand, since X and Y are uniform distributions, we get from theorem 2 that $\lim_{N\to\infty} EMD(A^*, U) = 0$.

Therefore, for infinite observations, $LI = \frac{K-1}{2}$, $GI = 0$.

Theorem 4: *LI* is a lower bound for the local contribution to the observed alignment

Assuming that the observed alignment distribution A is cumulatively explained by a global bias and a local interaction, we construct a new alignment distribution $A_{-\zeta}$ encoding the true cumulative local contribution to the observed alignment and demonstrate that $LI \leq EMD(U, A_{-\zeta}) - GI$ to conclude that LI is a lower bound for the local contribution to the observed alignment.

Proof:

We first define A^-, the alignment distribution corresponding to A that does not include any local interaction. Thus, A^-, can be interpreted as an alignment distribution constructed from $X^-$and Y^-, denoting X and Y after elimination of the (unknown) alignment correction due to local interactions between the observations (x_i, y_i). The construction of $A_{-\zeta}$ is based on the corresponding matching pairs $(x_i^- \in X^-, y_i^- \in Y^-)$ with alignment correction by the local interaction ζ_i (see Fig. 1B for as a schematic depiction). Such local interaction exists in our model (although it might not be explicitly known) and can be represented as a vector $\zeta \in \mathbb{R}^N, \zeta_i \geq 0\ \forall i$. Note, that this construction supports different ζ_i values for every observation i and thus can provide a more detailed platform than the single measure LI that DeBias outputs

(which assumes $\zeta_i = \zeta_j \ \forall i,j$). Also note, that when $\zeta_i > \theta_i^-$ (θ_i^- is the alignment angle between (x_i^-, y_i^-)), then the observed alignment $\theta_i^- - \zeta_i < 0$.

Accordingly, $A_{-\zeta}$ is defined as the alignment distribution of $\theta_i^- - \zeta_i$. As described above, $A_{-\zeta}$ can contain negative values for $\zeta_i > \theta_i^-$. *A*, the experimentally observed alignment, thus can be generated from $A_{-\zeta}$ as well, by truncating the "saturated" observations (where $\zeta_i > \theta_i^-$) to the value 0. More formally, the elements in *A* are defined by

$$\begin{array}{ll} \theta_i^- - \zeta_i & \theta_i^- > \zeta_i \\ 0 & \theta_i^- \leq \zeta_i \end{array}$$

We can get an upper bound for $EMD(A, A^-)$ in the form of:

$$EMD(A, A^-) \leq EMD\left(A_{-\zeta}, A^-\right) \leq \sum_{i=1}^{N} \frac{1}{N} \left\lceil \frac{\zeta_i}{|bin|} \right\rceil \ .$$

Where $|bin|$ defines the size of the angular interval of a bin in the alignment histogram.

This equation is intuitively interpreted as every observation i is locally aligned by ζ_i, and therefore is translocated $\left\lceil \frac{\zeta_i}{|bin|} \right\rceil$ bins, at most.

Note that a decreased bin size reduces this bound as close as needed to the value of $EMD\left(A_{-\zeta}, A^-\right)$.

Finally,

$$LI \underset{*}{\leq} EMD(A, A^*) \approx EMD(A\ , A^-) \leq EMD\left(A_{-\zeta}\ , A^-\right) \leq \sum_{i=1}^{N} \frac{1}{N} \left\lceil \frac{\zeta_i}{|bin|} \right\rceil$$

Thus the LI is a lower bound on the contribution of the direct interaction between X and Y on the alignment distribution.

Additionally, we get that

$$GI = EMD(A^*, U) = EMD(A, U) - LI \underbrace{\geq}_{*} EMD(A, U) - EMD(A^*, A)$$

$$\geq EMD(A, U) - \sum_{i=1}^{N} \frac{1}{N} \left\lceil \frac{\zeta_i}{|bin|} \right\rceil$$

Implying that the GI is an upper bound of the contribution of the global bias.

* by corollary 2

Corollary 2: For any alignment distribution A, $LI \leq EMD(A, A^*)$

Proof:

Let A_i, A_i^*, U_i denote the relative frequency of observations in bin_i, $0 \leq i \leq k-1$ for A, A^*, U, respectively.

$$EMD(A, A^*) = \sum_{0 \leq i \leq K-1} |A_i - A_i^*|$$

$$= \sum_{0 \leq i \leq K-1} |A_i - U_i + U_i - A_i^*| \underbrace{\geq}_{*} \sum_{0 \leq i \leq K-1} (|A_i - U_i| - |A_i^* - U_i|)$$

$$= \sum_{0 \leq i \leq K-1} |A_i - U_i| - \sum_{0 \leq i \leq K-1} |A_i^* - U_i| = EMD(A, U) - EMD(A^*, U) = LI$$

* triangle inequality

Theorem 5: GI limits for highly variant truncated normal distributions

$\lim_{\substack{\sigma \to \infty \\ N \to \infty}} GI = 0$ for the following scenarios:

(I) The resampled alignment is a truncated normal distribution with variance parameter σ^2.

(II) X and Y are truncated normal distributions, each with variance parameter σ^2.

Proof:

(I)

Let A_σ^* be the truncated normal resampled alignment distribution, defined by the parameters $\mu = 0, \sigma,$ with the support interval (a, b), such that $a \leq \mu \leq b$. Let $\phi^\sigma(x)$ be the probability density function (PDF) of the truncated normal distribution and $u(x)$, U respectively, the PDF and CDF (cumulative distribution function)of the uniform distribution function on (a, b). The PDF and CDF of the normal distribution function is denoted in the standard notation of ϕ and Φ respectively.

First we prove that $\lim_{\sigma\to\infty} \phi^\sigma(x) = u(x)$ and use this to conclude that $\lim_{\substack{\sigma\to\infty \\ N\to\infty}} \mathrm{EMD}(A_\sigma^*, U) = \lim_{\substack{\sigma\to\infty \\ N\to\infty}} \mathrm{GI} = 0$.

$$\forall x_1, x_2 \in (a, b), \lim_{\sigma\to\infty} \frac{\phi^\sigma(x_1)}{\phi^\sigma(x_2)} = \lim_{\sigma\to\infty} \frac{\frac{\phi\left(\frac{x_1}{\sigma}\right)}{\sigma\left(\Phi\left(\frac{b}{\sigma}\right) - \Phi\left(\frac{a}{\sigma}\right)\right)}}{\frac{\phi\left(\frac{x_2}{\sigma}\right)}{\sigma\left(\Phi\left(\frac{b}{\sigma}\right) - \Phi\left(\frac{a}{\sigma}\right)\right)}} = \lim_{\sigma\to\infty} \frac{\phi\left(\frac{x_1}{\sigma}\right)}{\phi\left(\frac{x_2}{\sigma}\right)} = \lim_{\sigma\to\infty} \frac{e^{-\frac{x_1^2}{2\sigma^2}}}{e^{-\frac{x_2^2}{2\sigma^2}}}$$

$$= \lim_{\sigma\to\infty} e^{\frac{x_2^2 - x_1^2}{2\sigma^2}} = 1$$

Therefore, $\lim_{\sigma\to\infty} \phi_t^{\sigma}(x) = Constant$. Since the support of ϕ_t^{σ} is (a, b), the only constant satisfying that $\lim_{\sigma\to\infty} \phi_t^{\sigma}(x)$ is a probability distribution is $\frac{1}{b-a} = u(x)$. Therefore, $\lim_{\substack{\sigma\to\infty \\ N\to\infty}} \mathrm{EMD}(A_{\sigma}^{*}, U) = \lim_{\substack{\sigma\to\infty \\ N\to\infty}} GI = 0$.

(II)

Let X, Y be truncated normal distributions. In part (I) we prove that $\lim_{\sigma\to\infty} X = \lim_{\sigma\to\infty} Y = u(x)$. Theorem 2 implies that when X and Y are uniform distributions $\lim_{\substack{\sigma\to\infty \\ N\to\infty}} GI = 0$.

Simulations with constant ζ

To assess the performance of DeBias we tested its ability to retrieve a pre-determined local interaction parameter ζ (see Fig. 1B) from simulated synthetic data. X and Y were modeled as truncated normal distributions on (-90, 90), with μ=0 and changing σ_x, σ_y. Pairs of (x_i,y_i) were sampled from X, Y and shifted towards each other by ζ degrees (similar to Fig. 1B, but with a constant cumulative ζ) to construct the observed alignment angles. To avoid confusion we denote X, Y, σ_x, σ_y as the observed values post-simulation. For a given constant ζ, we exhaustively explored the σ_x, σ_y space. For each σ_x, σ_y, we performed 20 independent simulations with N= 1600 observations (x_i,y_i). For each simulation we constructed the resampled distribution 10 times based on 400 observations drawn from the marginal X, Y distributions, and used the mean GI, LI. The final recorded GI, LI were averaged over the independent simulations.

The expected mean alignment when neither global bias nor local interactions exist is 45°. We begin by examining the deviation of the mean observed alignment from this value (45° - θ_{mean}). Better alignment is reflected by higher 45° - θ_{mean} values implying a larger deviation from the

unbiased and no-interactions scenario. Low standard deviations σ, correspond to better alignment, improving with growing ζ, as expected (Supplementary Fig. S2A). The GI follows a similar pattern and remains relatively stable for small changes in ζ (Supplementary Fig. S2B). The similar patterns between Fig. S2A and B indicate that the global bias has a prominent role in determining the observed alignment.

The LI grows with ζ (Supplementary Fig. S2C) and its relative contribution to the observed alignment grow with increasing ζ (Fig. S2D, quantified by LI/(LI+GI)), as expected. This relative contribution can be harnessed to restore an estimated ζ as the corresponding fraction from $(45° - \theta_{mean})$ (Supplementary Fig. S2E). The estimated ζ is a lower bound for the actual value (Method: Theory, Theorem 4). Estimation is more accurate for larger ζ and for large σ (Supplementary Fig. S2F). These results again highlight the importance of exploiting the GI for better interpretation of the LI (first introduced in Fig. 2D-E).

We also investigated the effect of the choice of the number of bins K, used for sampling and computation of the EMD between distributions. Increased K induces linear growth in LI and GI values, as expected (Data S1 Theorem 5, Supplementary Fig. S3A-B) and stabilized its accuracy in predicting ζ for $K \geq 11$ (Supplementary Fig. S3 C-E). Large K will require more observations to estimate the true distribution. Using a constant K for a specific application assures fair comparison between different cases. Varying N, the number of observations, did not have a major effect on these measurements (Supplementary Fig. S4A-D), but increasing N reduced the noise which increased the accuracy in predicting ζ (Supplementary Fig. S4E).

Vimentin and Microtubule filaments experiments and analysis

Cell model: hTERT-RPE-1 cells were TALEN-genome edited to endogenously label vimentin with mEmerald and α-tubulin with mTagRFPt. These cells were validated for protein expression levels (Gan, Ding and Burckhardt et al., in review). Cells were stably transfected with shRNA against vimentin to knock down vimentin and the knockdown efficiency was validated as ~75% (Gan, Ding and Burckhardt et al., in review).

Fixed cell imaging: hTERT-RPE-1 mEmerald-vimentin/mTagRFPt-α-tubulin cells expressing shRNA-VIM or control shRNA Scr were plated into MatTek (Ashland, MA) 35 mm glass-bottom dishes (P35G-0-20-C) coated with 5 µg/mL fibronectin. Cells were incubated overnight to allow them to adhere and form monolayers. Monolayers were scratched with a pipette tip to form a wound. Cells were incubated for 90 minutes, washed briefly and fixed with methanol at -20°C for 15 minutes. Cells were imaged at the wound edge (denoted "front" cells), and at 2-3 cell rows from the wound edge (denoted "back" cells, only for control condition). Images were acquired using a Nikon Eclipse Ti microscope, equipped with a Nikon Plan Apo Lambda 100x/1.45 N.A. objective. Images were recorded with a Hamamatsu ORCA Flash 4.0 with 6.45 µm pixel size (physical pixel size: 0.0645 x 0.0645 µm). All microscope components were controlled by Mciro-manager software.

Live cell imaging: hTERT-RPE-1 mEmerald-vimentin/mTagRFPt-α-tubulin cells expressing control shRNA Scr were plated into MatTek (Ashland, MA) 35 mm glass-bottom dishes (P35G-0-20-C) coated with 5 µg/mL fibronectin. Cells were incubated overnight to allow them to adhere and form monolayers. Monolayers were scratched with a pipette tip to form a wound. Imaging started 30 minutes after scratching with an Andor Revolution XD spinning disk microscope mounted on a Nikon Eclipse Ti stand equipped with Perfect Focus, a Nikon Apo 60x 1.49 N.A. oil objective and a 1.5x optovar for further magnification. Images were recorded with

an Andor IXON Ultra EMCCD camera with 16 μm pixel size (physical pixel size: 0.16 x 0.16 μm). Lasers with 488 nm and 561 nm light emission were used for exciting mEmerald and mTagRFPt, respectively. The output powers of the 488 nm and 561 nm lasers were set to 10% and 20% of the maximal output (37 mW and 23 mW, respectively). The exposure time was 300 ms per frame for both channels and images were collected at a frame rate of 1 frame per minute. During acquisition, cells were kept in an onboard environmental control chamber as described above. All microscope components were controlled by Metamorph software.

Filaments extraction and spatial matching: We applied the filament reconstruction algorithm reported in (Gan, Ding and Burckhardt et al., in review; Ding and Danuser, in review). Briefly, multi-scale steerable filtering is used to enhance curvilinear image structures, centerlines of candidate filament fragments are detected, clustered to high and low confidence sets and iterative graph matching is applied to connect fragments into complete filaments. Each filament is represented by an ordered chain of pixels and the local filament orientation derived from the steerable filter response. Spatial matching was performed as follows: each pixel belonging to a filament detected in the MT channel is recorded to the closest pixel that belongs to a filament in the VIM channel. If the distance between the two pixels is less than 20 pixels, then the pair of VIM and MT orientations at this pixel is recorded for analysis. The same process is repeated to record matched pixels from VIM to MT filaments.

Collective cell migration experiments and analysis

Coupled measurements of velocity direction and stress orientation were taken from the data originally published by Tamal Das et al. (Das et al., 2015). Particle image velocimetry (PIV) was

applied to calculate velocity vectors, monolayer stress microscopy (Tambe et al., 2011) to extract stress orientations. These measures were recorded 3 hours after collective migration was induced by lifting off the culture-insert in which the cells have grown to confluence. Validated siRNAs were used for gene screening. Detailed experimental settings can be found in the original paper (Das et al., 2015).

Statistical test: We devised a permutation test to determine statistical significance of different LI values (Fig. 4C). For a given experiment (condition) 50% of the velocity-stress observations were randomly selected and the LI (and GI) calculated from this subsampling. For every pair of conditions (i,j), where $LI_i < LI_j$ (when considering all observations, as in Fig. 4B) this procedure was repeated for 100 iterations and the p-value was recorded based on the number of iterations in which the subsampled LI value for condition i was higher than that for condition j. 0 such occurrences thus implies $p < 0.01$.

Clathrin mediated endocytosis experiments

Cells and cell culture: Retinal pigment epithelial (RPE) cells stably expressing EGFP-CLC were generated as previously described (Aguet et al., 2013). RPE cells stably expressing µ2 subunit harboring $AP2^{WT}$ or $AP2\mu^{cargo-}$ were generated as described in (Aguet et al., 2013). cDNA of $AP2^{WT}$ or $AP2\mu^{cargo-}$ µ adaptin harboring silent mutations that confer resistance to siRNA , was kindly provided by M.S. Robinson (Motley et al., 2006). The cDNA was subcloned into the pMIEG3-IRES-mTagBFP retroviral vector. $AP2^{WT}$ or $AP2\mu^{cargo-}$ µ adaptin pMIEG3-mTagBFP construct was used to generate retroviruses. Expression of µ-adaptins within each stable cell line, sorted by FACS for low expression of BFP, was determined by Western blotting

with µ-adaptin antibody (Supplementary Fig. S5). All cells were grown under 5% CO2 at 37°C in DMEM medium supplemented with 10% (v/v) and fetal calf serum (FCS, HyClone).

siRNA transfection: To silence endogenous µ-adaptin, RPE cells were transfected with a previously established siRNA sequence (Motley et al., 2006) using RNAiMAX (LifeTechnologies, Carlsbad, CA) following the manufacturer's instructions. Briefly, 200 pmol of µ-adaptin siRNA and 15 µl of RNAiMAX reagent were added in 2 ml of OptiMEM in 6-cm dish of RPE cells for 4 hours. Transfection was performed three times, 96h, 72h and 48h, prior to experiments.

Fluorescence microscopy: Total internal reflection fluorescence (TIRF) microscopy was performed as previously described (Reis et al., 2015). RPE cells expressing EGFP-CLC and wild type or mutant AP2 µ subunit were imaged using a 100 Å~ 1.49 NA Apo TIRF objective (Nikon) mounted on a Ti-Eclipse inverted microscope equipped with the Perfect Focus System (Nikon). Cell surface TfnR labeling: Cells were incubated for 5min at 20 µg/ml of Transferrin-Alexa fluor 568 or 647 (TfnR) at room temperature. Cells were washed 3 times with ice cold PBS and fixed following a protocol previously described (Mettlen et al., 2010).

EGF pulse and EGFR immunofluorescence: Cells were stimulated for 5 minutes at room temperature with 20 ng/*ml* of EGF and EGFR was labeled simultaneously with 4 µg/*ml* of anti-EGFR monoclonal antibody AB11 (ThermoFisher Scientific) in DMEM. Starved cells or cells under basal conditions were incubated with a solution of 4 µg/*ml* anti-EGFR monoclonal antibody AB11 in DMEM for 5 minutes and cells were washed and fixed as described above.

Image analysis: Single cell masks were manually annotated in each field of view. We applied the approach described in (Aguet et al., 2013) to automatically detect CCPs from the CLC channel. Briefly, CLC fluorescence was modeled as a two-dimensional Gaussian approximation

of the microscope PSF above a spatially varying local background. CCP candidates were first detected via filtering, followed by a model-fitting for sub-pixel localization. The fluorescent intensity of the CLC and any other acquired channel were recorded in the detection coordinates to define the matched observations for DeBias. GI and LI were calculated independently for each single cell. Linear discriminant analysis (LDA) (Fisher, 1936) was applied to assess the accuracy of single cell classification. The confusion matrix (Fawcett, 2006) was used to visualize classification accuracy and errors. Briefly, the matrix displays a table where bin (i,j) corresponds to the percentage of cells from experimental condition i classified as condition j. The general classification-accuracy percentage was recorded (corresponding to the values on the table's diagonal) to assess the accuracy of different measures.

Assessing classification accuracy and statistics: Linear Discriminative Analysis (LDA) classification was applied to assess single-cell classification accuracy. Every cell constituted an observation, a label was assigned based on the experimental condition and the colocalization of the paired proteins was quantitatively represented by a scalar (local index, Pearson's coefficient or mutual information) or by a two-dimensional feature (the global index together with either of the scalar-representations). LDA classifier was trained on a labeled dataset consisting of 2 or 3 different experimental conditions and the classification accuracy was reported. Statistical significance for distinguishable different experimental conditions was calculated by a permutation test: the labels of the experimental conditions were permuted such that every observation was assigned an arbitrary label, an LDA classifier was trained based on these new labels and its classification accuracy was recorded. 1000 iterations were performed and the p-value was defined using the number of iterations at which the random labeling single cell classification accuracy exceeded that of the un-randomized labels. Confusion matrices were

calculated to visualize classification accuracy and evaluate sensitivity of alternative measures for colocalization: bin (i,j) holds the percentage of cells with label i that were classified with label j. Statistical significance for comparing classification performance of LDA classifiers that were trained for scalar measures with or without the GI was calculated by bootstrapping. The following process was repeated 1000 times and the frequency for which the scalar-based classifier outperformed the classifier trained on pairs of measures was reported as the p-value. Random resampling with replacement was performed to obtain a sample size identical to that of the observed dataset. The competing pre-trained LDA classifiers accuracy was assessed for this resampled dataset and recorded when the model that was trained without the GI predicted better.

Webserver

The DeBias code was implemented in Matlab, compiled with Matlab complier SDK and transferred to a web-based platform to allow public access for all users at https://debias.biohpc.swmed.edu[2]. The graphical user interface (GUI) was designed to be simple and easy to use. The user uploads one or more datasets to the DeBias webserver, selects "angles data" checkbox if the variables are angles and presses the "process" data. GI and LI values are displayed and the results can be downloaded or emailed to the user once the calculation is completed. The software's flow chart and a detailed user manual are available in online user manual.

[2] The web server will be available online upon UTSW information resources approval. Until then, please use the source code from the GITHUB repository https://git.biohpc.swmed.edu/ydu/debias/tree/master.

Acknowledgements

We thank the BioHPC team at UTSW and especially Liqiang Wang for the help in implementing the DeBias web-server and making it freely available for public use. We are grateful to Tamal Das and Joachim Spatz for providing us with the motion and stress data from their tight-junction screen, to Christoph Burckhardt for the endogenously labelled vimentin and α-tubulin RPE cells and to Liya Ding for the filaments segmentation software. We thank Sangyoon Han, Claudia Schaefer, Meghan Driscoll, Erik Welf, Phillippe Roudot and Marcel Mettlen for critically reading the manuscript and Marcel Mettlen for fruitful discussions and advice. This work was supported by the Cancer Prevention and Research Institute of Texas (CPRIT R1225 to GD), by NIH P01 GM103723 (to GD) and by GM713165 (to SLS and GD). UO was supported by a fellowship from the Manna Program in Food Safety and Security. ZK has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme under REA grant agreement n° PIOF-GA-2012-330268 and the Swiss National Science Foundation Fellowship for Prospective Researchers. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Abbreviations List

GUI - graphical user interface

EMD - Earth Mover's Distance

GI - global index

LI - local index

RPE - Retinal Pigment Epithelial

VIM - Vimentin

MT - Microtubule

CME - clathrin mediated endocytosis

CCPs - clathrin-coated pits

CLC - clathrin light chain

TIRFM - Total Internal Reflection Fluorescence microscopy

TfnR - transferrin receptor

LDA - Linear Discriminative Analysis

EGFR - epidermal growth factor receptor

PIV - Particle image velocimetry

Author Contribution

AZ and GD conceived the study. AZ designed experiments, performed simulations, analyzed the data, assisted in theoretical results and the web-server design and development and wrote the initial draft. UO devised the theoretical part and assisted in simulations. ZK conceived designed and implemented the endocytosis colocalization experiments. ZG designed and performed the experiments on vimentin-microtubule alignment. YD implemented the web-server and wrote its user manual. SLS and GD supervised the research. AZ, UO, ZK and GD wrote the paper. All authors read and edited the manuscript and approved of its content.

Competing Financial Interests

The authors declare that they have no competing interests.

References

Adler, J., and I. Parmryd. 2010. Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient. *Cytometry Part A*. 77:733-742.

Aguet, F., C.N. Antonescu, M. Mettlen, S.L. Schmid, and G. Danuser. 2013. Advances in analysis of low signal-to-noise images link dynamin and AP2 to the functions of an endocytic checkpoint. *Developmental cell*. 26:279-291.

Bazellières, E., V. Conte, A. Elosegui-Artola, X. Serra-Picamal, M. Bintanel-Morcillo, P. Roca-Cusachs, J.J. Muñoz, M. Sales-Pardo, R. Guimerà, and X. Trepat. 2015. Control of cell–cell forces and collective cell dynamics by the intercellular adhesome. *Nature cell biology*. 17:409-420.

Bolte, S., and F. Cordelieres. 2006. A guided tour into subcellular colocalization analysis in light microscopy. *Journal of microscopy*. 224:213-232.

Chung, I., R. Akita, R. Vandlen, D. Toomre, J. Schlessinger, and I. Mellman. 2010. Spatial control of EGF receptor activation by reversible dimerization on living cells. *Nature*. 464:783-787.

Clegg, R.M. 1995. Fluorescence resonance energy transfer. *Current opinion in biotechnology*. 6:103-110.

Costantino, S., J.W. Comeau, D.L. Kolin, and P.W. Wiseman. 2005. Accuracy and dynamic range of spatial image correlation and cross-correlation spectroscopy. *Biophysical journal*. 89:1251-1260.

Costes, S.V., D. Daelemans, E.H. Cho, Z. Dobbin, G. Pavlakis, and S. Lockett. 2004. Automatic and quantitative measurement of protein-protein colocalization in live cells. *Biophysical journal*. 86:3993-4003.

Cover, T.M., and J.A. Thomas. 2012. Elements of information theory. John Wiley & Sons.

Das, T., K. Safferling, S. Rausch, N. Grabe, H. Boehm, and J.P. Spatz. 2015. A molecular mechanotransduction pathway regulates collective migration of epithelial cells. *Nature cell biology*. 17:276-287.

Drew, N.K., M.A. Eagleson, D.B. Baldo Jr, K.K. Parker, and A. Grosberg. 2015. Metrics for Assessing Cytoskeletal Orientational Correlations and Consistency.

Dunn, K.W., M.M. Kamocka, and J.H. McDonald. 2011. A practical guide to evaluating colocalization in biological microscopy. *American Journal of Physiology-Cell Physiology*. 300:C723-C742.

Ehrlich, M., W. Boll, A. van Oijen, R. Hariharan, K. Chandran, M.L. Nibert, and T. Kirchhausen. 2004. Endocytosis by random initiation and stabilization of clathrin-coated pits. *Cell*. 118:591-605.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern recognition letters*. 27:861-874.

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics*. 7:179-188.

Goh, L.K., F. Huang, W. Kim, S. Gygi, and A. Sorkin. 2010. Multiple mechanisms collectively regulate clathrin-mediated endocytosis of the epidermal growth factor receptor. *The Journal of cell biology*. 189:871-883.

He, S., C. Liu, X. Li, S. Ma, B. Huo, and B. Ji. 2015. Dissecting Collective Cell Behavior in Polarization and Alignment on Micropatterned Substrates. *Biophysical journal*. 109:489-500.

Helmuth, J.A., G. Paul, and I.F. Sbalzarini. 2010. Beyond co-localization: inferring spatial interactions between sub-cellular structures from microscopy images. *BMC bioinformatics*. 11:372.

Höning, S., D. Ricotta, M. Krauss, K. Späte, B. Spolaore, A. Motley, M. Robinson, C. Robinson, V. Haucke, and D.J. Owen. 2005. Phosphatidylinositol-(4, 5)-bisphosphate regulates sorting signal recognition by the clathrin-associated adaptor complex AP2. *Molecular cell*. 18:519-531.

Kalaidzidis, Y., I. Kalaidzidis, and M. Zerial. 2015. A probabilistic method to quantify the colocalization of markers on intracellular vesicular structures visualized by light microscopy. *In* AIP Conference Proceedings. Vol. 1641. 580.

Kantorovich, L.V., and G.S. Rubinstein. 1958. On a space of completely additive functions. *Vestnik Leningrad. Univ*. 13:52-59.

Karlon, W.J., P.-P. Hsu, S. Li, S. Chien, A.D. McCulloch, and J.H. Omens. 1999. Measurement of orientation and distribution of cellular alignment and cytoskeletal organization. *Annals of biomedical engineering*. 27:712-720.

Kawashima, N., K. Nakayama, K. Itoh, T. Itoh, M. Ishikawa, and V. Biju. 2010. Reversible Dimerization of EGFR Revealed by Single-Molecule Fluorescence Imaging Using Quantum Dots. *Chemistry-A European Journal*. 16:1186-1192.

Krishnaswamy, S., M.H. Spitzer, M. Mingueneau, S.C. Bendall, O. Litvin, E. Stone, D. Pe'er, and G.P. Nolan. 2014. Conditional density-based analysis of T cell signaling in single-cell data. *Science*. 346:1250689.

Lagache, T., N. Sauvonnet, L. Danglot, and J.C. Olivo-Marin. 2015. Statistical analysis of molecule colocalization in bioimaging. *Cytometry Part A*. 87:568-579.

Lamaze, C., T. Baba, T.E. Redelmeier, and S.L. Schmid. 1993. Recruitment of epidermal growth factor and transferrin receptors into coated pits in vitro: differing biochemical requirements. *Molecular biology of the cell*. 4:715-727.

Langer-Safer, P.R., M. Levine, and D.C. Ward. 1982. Immunological method for mapping genes on Drosophila polytene chromosomes. *Proceedings of the National Academy of Sciences*. 79:4381-4385.

Loerke, D., M. Mettlen, D. Yarar, K. Jaqaman, H. Jaqaman, G. Danuser, and S.L. Schmid. 2009. Cargo and dynamin regulate clathrin-coated pit maturation.

Manders, E., F. Verbeek, and J. Aten. 1993. Measurement of co-localization of objects in dual-colour confocal images. *Journal of microscopy*. 169:375-382.

Mettlen, M., D. Loerke, D. Yarar, G. Danuser, and S.L. Schmid. 2010. Cargo-and adaptor-specific mechanisms regulate clathrin-mediated endocytosis. *The Journal of cell biology*. 188:919-933.

Motley, A.M., N. Berg, M.J. Taylor, D.A. Sahlender, J. Hirst, D.J. Owen, and M.S. Robinson. 2006. Functional analysis of AP-2 α and μ2 subunits. *Molecular biology of the cell*. 17:5298-5308.

Nesterov, A., R.E. Carter, T. Sorkina, G.N. Gill, and A. Sorkin. 1999. Inhibition of the receptor-binding function of clathrin adaptor protein AP-2 by dominant-negative mutant μ2 subunit and its effects on endocytosis. *The EMBO journal*. 18:2489-2499.

Nieuwenhuizen, R.P., L. Nahidiazar, E.M. Manders, K. Jalink, S. Stallinga, and B. Rieger. 2015. Co-Orientation: Quantifying Simultaneous Co-Localization and Orientational Alignment of Filaments in Light Microscopy. *PloS one*. 10.

Pearson, R.A. 1901. Section I, Social and Economic Science. *Science*. 14:912-926.

Peleg, S., M. Werman, and H. Rom. 1989. A unified approach to the change of resolution: Space and gray-level. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 11:739-742.

Piston, D.W., and G.-J. Kremers. 2007. Fluorescent protein FRET: the good, the bad and the ugly. *Trends in biochemical sciences*. 32:407-414.

Reis, C.R., P.H. Chen, S. Srinivasan, F. Aguet, M. Mettlen, and S.L. Schmid. 2015. Crosstalk between Akt/GSK3β signaling and dynamin-1 regulates clathrin-mediated endocytosis. *The EMBO journal*:e201591518.

Reshef, D.N., Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, and P.C. Sabeti. 2011. Detecting novel associations in large data sets. *Science*. 334:1518-1524.

Rizk, A., G. Paul, P. Incardona, M. Bugarski, M. Mansouri, A. Niemann, U. Ziegler, P. Berger, and I.F. Sbalzarini. 2014. Segmentation and quantification of subcellular structures in fluorescence microscopy images using Squassh. *Nature protocols*. 9:586-596.

Rubner, Y., C. Tomasi, and L.J. Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*. 40:99-121.

Schwille, P., F.-J. Meyer-Almes, and R. Rigler. 1997. Dual-color fluorescence cross-correlation spectroscopy for multicomponent diffusional analysis in solution. *Biophysical journal*. 72:1878.

Serra-Picamal, X., V. Conte, R. Vincent, E. Anon, D.T. Tambe, E. Bazellieres, J.P. Butler, J.J. Fredberg, and X. Trepat. 2012. Mechanical waves during tissue expansion. *Nature Physics*. 8:628-U666.

Sigismund, S., E. Argenzio, D. Tosoni, E. Cavallaro, S. Polo, and P.P. Di Fiore. 2008. Clathrin-mediated internalization is essential for sustained EGFR signaling but dispensable for degradation. *Developmental cell*. 15:209-219.

Tambe, D.T., C.C. Hardin, T.E. Angelini, K. Rajendran, C.Y. Park, X. Serra-Picamal, E.H.H. Zhou, M.H. Zaman, J.P. Butler, D.A. Weitz, J.J. Fredberg, and X. Trepat. 2011. Collective cell guidance by cooperative intercellular forces. *Nature materials*. 10:469-475.

Taylor, M.J., D. Perrais, and C.J. Merrifield. 2011. A high precision survey of the molecular dynamics of mammalian clathrin-mediated endocytosis. *PLoS-Biology*. 9:581.

Traub, L.M. 2009. Tickets to ride: selecting cargo for clathrin-regulated internalization. *Nature reviews Molecular cell biology*. 10:583-596.

Trepat, X., and J.J. Fredberg. 2011. Plithotaxis and emergent dynamics in collective cellular migration. *Trends in Cell Biology*. 21:638-646.

Villaseñor, R., H. Nonaka, P. Del Conte-Zerial, Y. Kalaidzidis, and M. Zerial. 2015. Regulation of EGFR signal transduction by analogue-to-digital conversion in endosomes. *eLife*. 4:e06156.

Welf, E.S., and G. Danuser. 2014. Using Fluctuation Analysis to Establish Causal Relations between Cellular Events without Experimental Perturbation. *Biophysical journal*. 107:2492-2498.

Wu, Y., M. Eghbali, J. Ou, R. Lu, L. Toro, and E. Stefani. 2010. Quantitative determination of spatial protein-protein correlations in fluorescence confocal microscopy. *Biophysical journal*. 98:493-504.

Zaritsky, A., E.S. Welf, Y.-Y. Tseng, M.A. Rabadán, X. Serra-Picamal, X. Trepat, and G. Danuser. 2015. Seeds of Locally Aligned Motion and Stress Coordinate a Collective Cell Migration. *Biophysical journal*. 109:2492-2500.

Supplementary information

- Supplementary figures
    - Supplementary Figure S1: Simulation of mutual information as function of GI
    - Supplementary Figure S2: Simulations of global bias and local interaction parameters
    - Supplementary Figure S3: Simulations of quantization parameter K
    - Supplementary Figure S4: Simulations of number of observations N
    - Supplementary Figure S5: Expression of µ-adaptins within cell lines
- Supplementary videos legends
    - Supplementary Video S1: Polarization of RPE cells at the monolayer edge over time

Supplementary figures

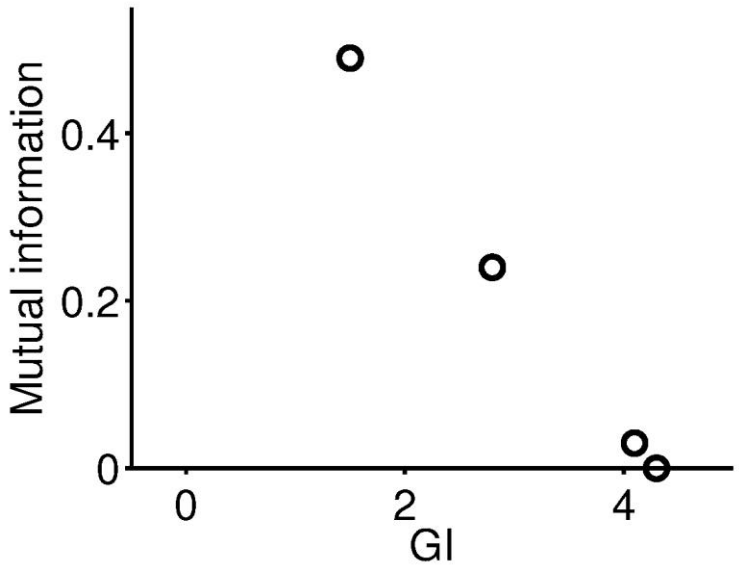


Figure S1: Mutual information as function of GI. MI is negatively associated with GI, similarly to the trend shown for LI (Fig. 3). Constant interaction parameter $\alpha = 0.2$ and varying standard deviation of X, Y, σ = 50°-5° (left –to-right). Simulated (GI,LI) for the corresponding σ: increased σ enhances GI and reduces LI.

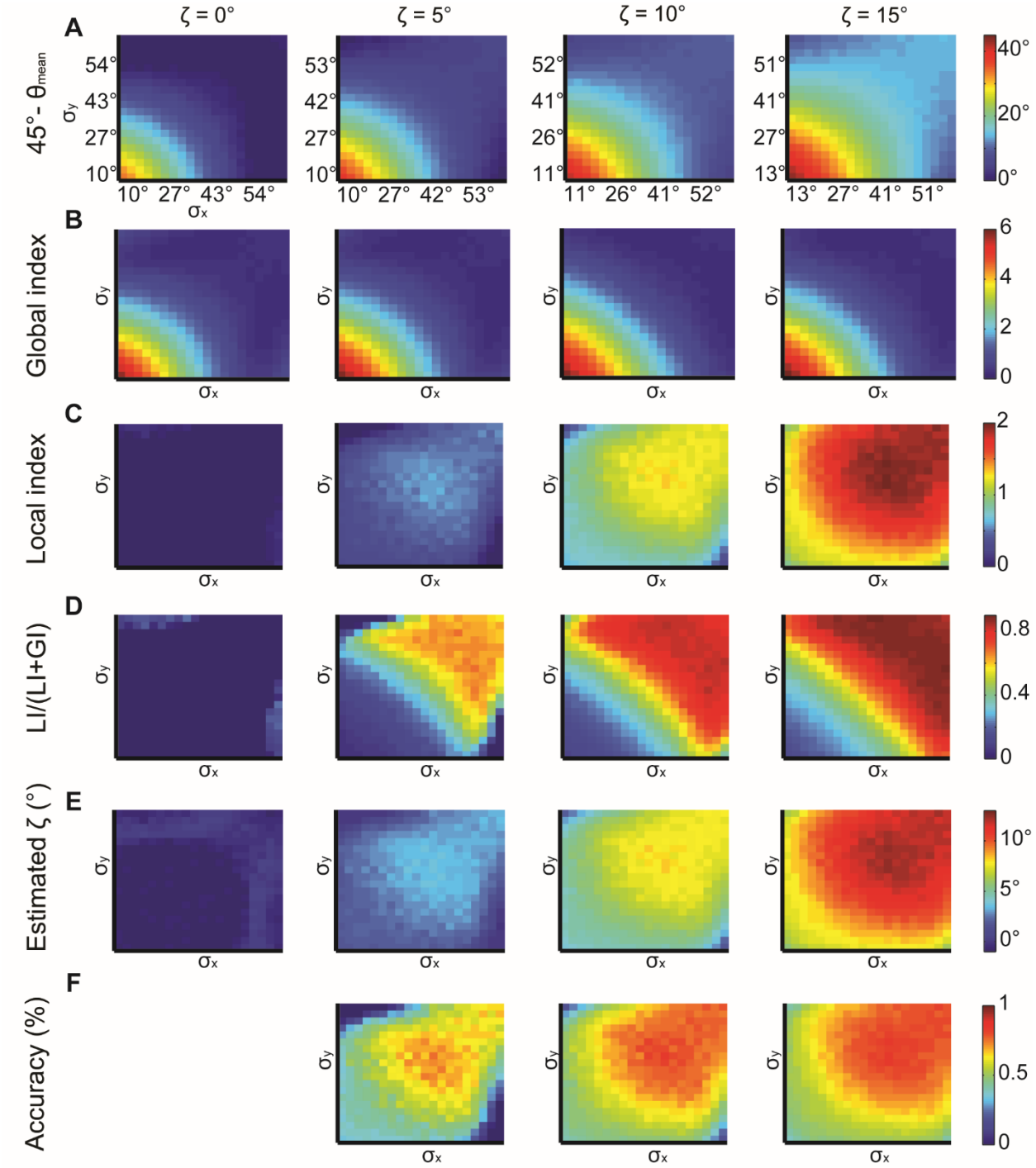


Figure S2: Simulations for X, Y normal distributions with different σ_x, σ_y and constant $\zeta = 0°, 5°, 10°, 15°$. (A) $45° - \theta_{mean}$ reflecting the cumulative effect of the global bias and the local interaction between X, Y (θ_{mean} is the mean observed alignment). Lower variance and higher ζ correspond to better alignment. (B) Global index. ζ has a small effect on GI. (C) Local index. ζ has a major effect on LI. (D) Relative contribution of LI to the observed alignment increases as function of ζ. (E) Retrieved estimated ζ calculated as the relative contribution of LI to the observed alignment (panel D) times the cumulative effect of the global bias and the local interaction (panel A). (F) Accuracy of estimated ζ grows with ζ and with lower σ_x, σ_y. Note, that this estimation is a lower bound for the true ζ (Method: Theory, Theorem 4). Accuracy cannot be measured for $\zeta = 0°$ hence the empty panel.

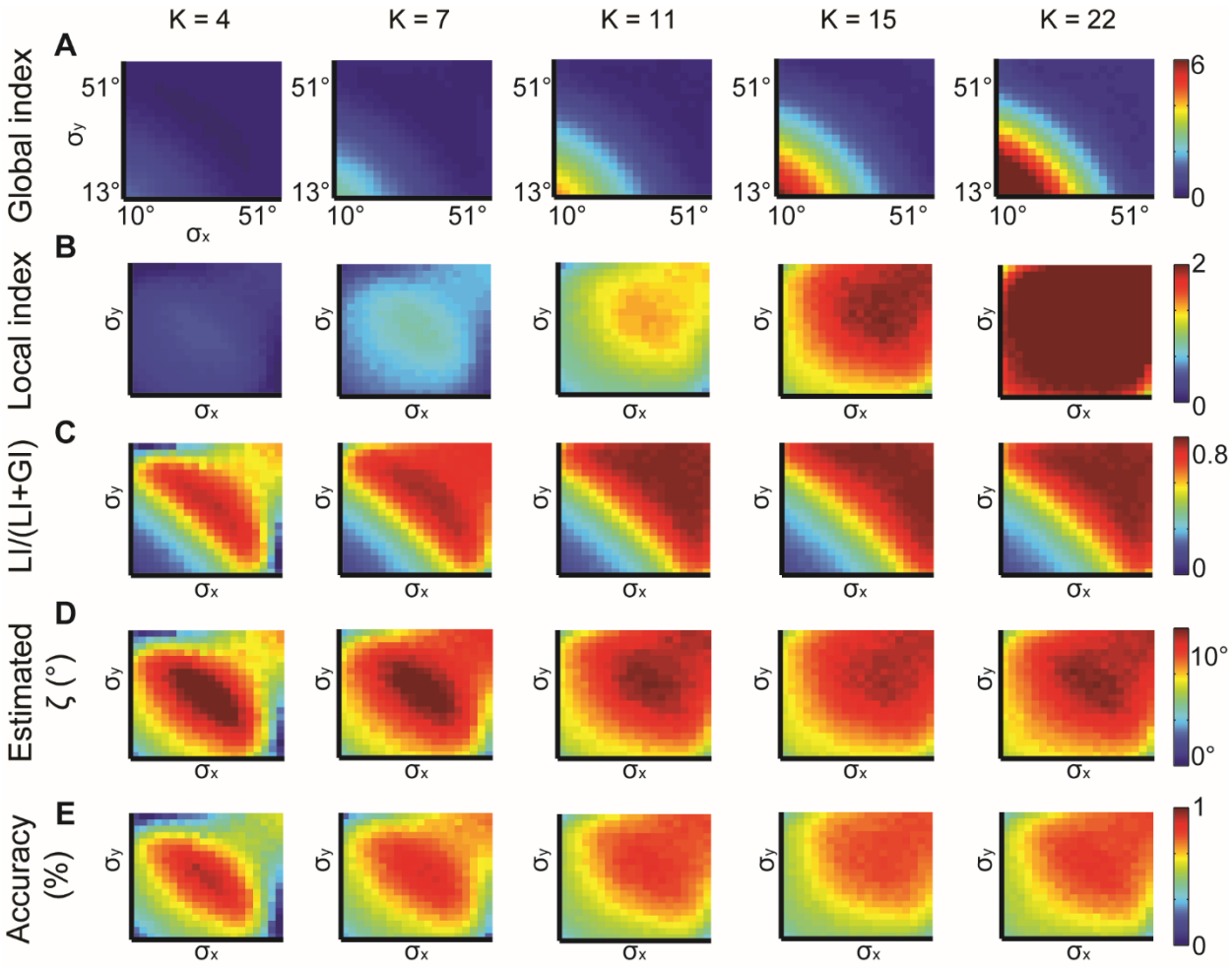


Figure S3: Simulations for different values of K, the number of bins in the alignment distribution. X, Y normal distributions with different σ_x, σ_y and constant $\zeta = 15°$. K = 4, 7, 11, 15, 22 were examined. (A-B) Global (A) and local (B) indices grow with K. (C-E) Relative contribution of LI to the observed alignment (C), Retrieved estimated ζ (D) and accuracy of estimated ζ (E) stabilizes for $K \geq 11$.

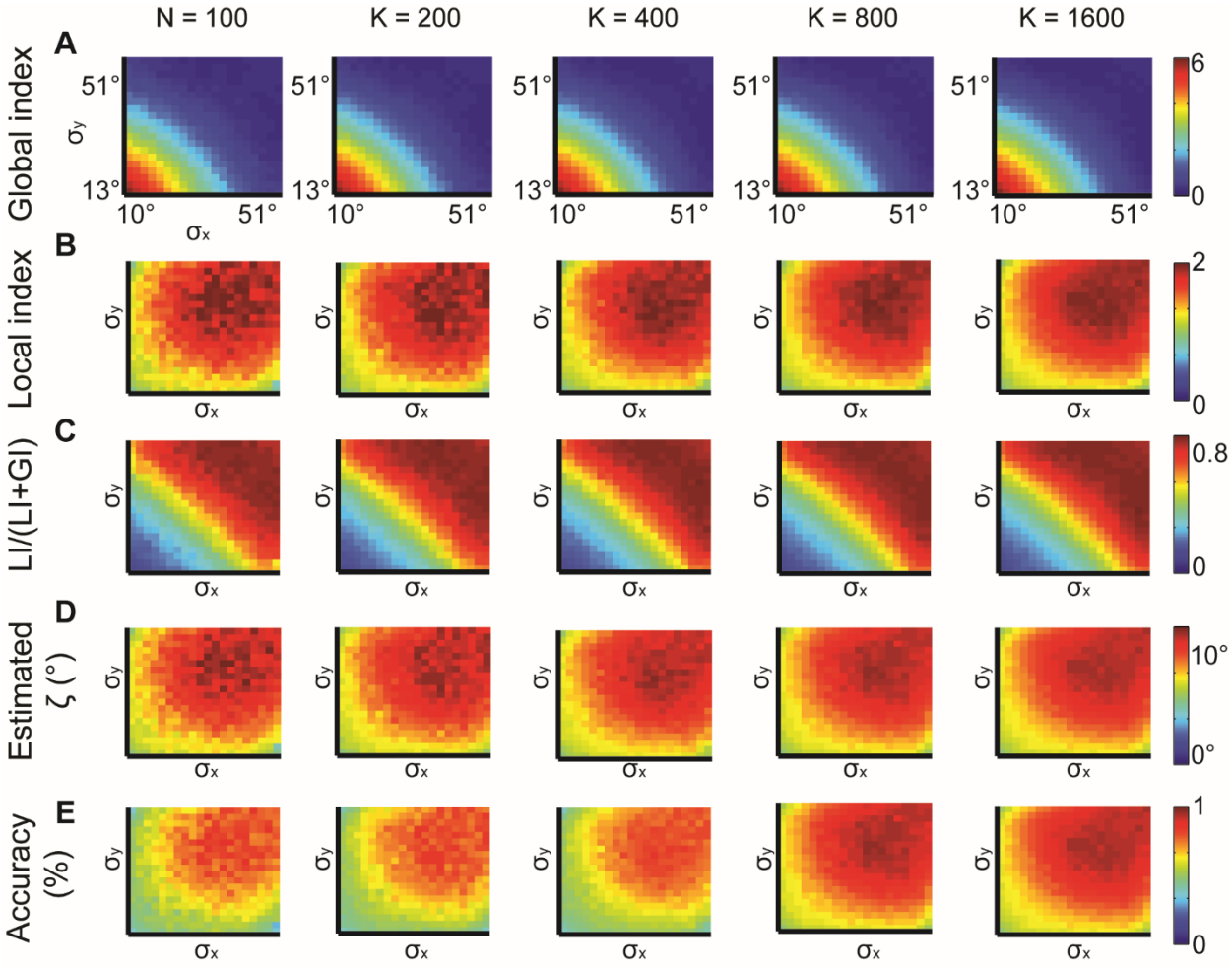


Figure S4: Simulations for different N, the number of observations. N = 100, 200, 400, 800, 16000 were examined. X, Y normal distributions with different σ_x, σ_y and constant $\zeta = 15°$. All measures provide similar information but are noisier for lower N. (A) Global index. (B) Local index. (C) Relative contribution of LI to the observed alignment. (D) Retrieved estimated ζ. (E) Accuracy of estimated ζ.

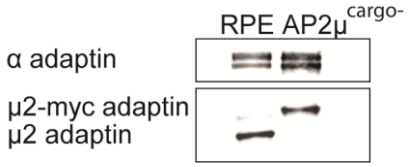


Figure S5: Immunoblot of AP2 μ subunit expression in AP2μ^{cargo-} mutant cell line treated with μ-adaptin siRNA to silence the expression of the endogenous protein.