Joint modeling of cell and nuclear shape variation

Gregory R. Johnson[a], Taraz E. Buck[a], Devin P. Sullivan[a], Gustavo K. Rohde[a,b], and Robert F. Murphy[a,b,c,d]

[a]Computational Biology Department and Center for Bioimage Informatics, [b]Department of Biomedical Engineering, and [c]Departments of Biological Sciences and Machine Learning, Carnegie Mellon University, Pittsburgh, PA 15213; [d]Freiburg Institute for Advanced Studies and Faculty of Biology, Albert Ludwig University of Freiburg, 79104 Freiburg im Breisgau, Germany

ABSTRACT **Modeling cell shape variation is critical to our understanding of cell biology. Previous work has demonstrated the utility of nonrigid image registration methods for the construction of nonparametric nuclear shape models in which pairwise deformation distances are measured between all shapes and are embedded into a low-dimensional shape space. Using these methods, we explore the relationship between cell shape and nuclear shape. We find that these are frequently dependent on each other and use this as the motivation for the development of combined cell and nuclear shape space models, extending nonparametric cell representations to multiple-component three-dimensional cellular shapes and identifying modes of joint shape variation. We learn a first-order dynamics model to predict cell and nuclear shapes, given shapes at a previous time point. We use this to determine the effects of endogenous protein tags or drugs on the shape dynamics of cell lines and show that tagged C1QBP reduces the correlation between cell and nuclear shape. To reduce the computational cost of learning these models, we demonstrate the ability to reconstruct shape spaces using a fraction of computed pairwise distances. The open-source tools provide a powerful basis for future studies of the molecular basis of cell organization.**

Monitoring Editor
Leah Edelstein-Keshet
University of British Columbia

Received: Jun 15, 2015
Revised: Aug 28, 2015
Accepted: Aug 28, 2015

INTRODUCTION

Understanding the relationship between cell and nuclear shape is an important problem in cell biology. Changes in cell and nuclear shape occur during development, in various pathologies, with addition of drugs, and after changes in gene expression. Although some work has been done to develop mechanistic models for cell and nuclear shape variation (Dahl *et al.*, 2006; Khatau *et al.*, 2009; Kihara *et al.*, 2011; Elliott *et al.*, 2015), efforts have been largely confined to assessing the effects of specific drugs or gene knockdowns to implicate particular molecules in shape regulation. For images of cells under various conditions, analysis has typically consisted of calculating descriptive features, such as cell shape, intensity, and texture features, to measure how shape correlates with condition (Yin *et al.*, 2008, 2009; Tsygankov *et al.*, 2014). Recently cell shape features have been used to identify discrete cell shape categories and the frequency of transitions between these categories and to learn how disruptions in signaling networks alter these transitions (Yin *et al.*, 2013; Sailem *et al.*, 2014).

These studies typically learn a probability distribution (either explicitly or implicitly) over cell or nuclear shapes, automatically determining which shapes are more or less likely. However, the models remain *descriptive*, in that they cannot readily be used to synthesize new shapes drawn from these probability distributions. As an alternative, parameters of functions that can generate shapes can be used instead of descriptive features, and the probability distributions learned over these parameters form a statistical *generative* framework over shapes (Pincus and Theriot, 2007; Zhao and Murphy, 2007; Peng and Murphy, 2011). This allows novel shapes to be created that are representative of the learned distribution.

Past analysis and modeling have typically not considered the *covariation* of cell or nuclear shape within a population. As part of an overall framework for capturing cell organization (Murphy, 2012), parametric approaches for modeling the relationship between cell

This article was published online ahead of print in MBoC in Press (http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E15-06-0370) on September 9, 2015.

Address correspondence to: Robert F. Murphy (murphy@cmu.edu).

Abbreviations used: ANOVA, analysis of variance; DAPI, 4′,6-diamidino-2-phenylindole; MDS, multidimensional scaling; MSE, mean squared error; RFP, red fluorescent protein; YFP, yellow fluorescent protein.

and nuclear shape for both two-dimensional (2D; Zhao and Murphy, 2007) and three-dimensional (3D; Peng and Murphy, 2011) images have been described. These models, however, require that the shapes to be modeled obey strict topological constraints (i.e., cell projections do not curve back toward the cell).

An alternative statistical generative framework that is not limited by shape assumptions has been presented for nuclear shape (Rohde *et al.*, 2008a,b; Peng *et al.*, 2009). It uses a nonrigid deformation method, large-deformation diffeomorphic metric mapping (LDDMM; Beg *et al.*, 2005), to measure distances between shapes. A similar approach has been used for comparing populations of cell shapes (Hagwood *et al.*, 2013). Given distances between all pairs of shapes in a collection, a map (a *shape space*) can be created that places each shape at coordinates such that its distance to the other shapes matches the measured distances as closely as possible (this is analogous to creating a map given only distances between cities). This can be accomplished using multidimensional scaling (MDS); the higher the dimensionality at which this map is created (mathematically, referred to as finding an embedding of that number of dimensions), the more accurately are the distances matched. Any coordinate in the shape space has a corresponding shape assigned to it. These shape spaces naturally encode variation across shapes, as more similar shapes are closer to each other, and less similar pairs of shapes are further away. If the properties of the cell and nuclear shapes linearly covary with each other (i.e., cells are always big and round or small and elongated), this covariation will be preserved in the low-dimensional embedding.

A generative model over the observed cell shapes can be constructed by fitting a probability density to the low-dimensional shape-space coordinates, as was done using kernel density estimation for nuclear shapes (Peng *et al.*, 2009). This provides an estimate of the probability density of any shape in the shape space, including shapes that have not actually been observed. Given a shape coordinate, triangulation methods can be used to deform neighboring shapes into the shape corresponding to the sampled location (Peng *et al.*, 2009).

These nonparametric models were constructed to represent single 2D shapes. Because cells and their components are 3D, realistic modeling should represent 3D variation in the shapes. In the work described here, we extend the nonparametric models to 3D shapes and to the combination of cell and nuclear shapes. This eliminates the need to model explicitly the conditional dependence of one shape on the other, in contrast with the previous parametric models (Zhao and Murphy, 2007; Peng and Murphy, 2011). We also develop generative models of the dynamics of cell and nuclear shape.

RESULTS

Determining the dependence of cell and nuclear shape on each other

An overview of our analysis and modeling pipeline is shown in Figure 1. To determine the relationship between the cell and nuclear shape, we applied this pipeline to 175 segmented 3D HeLa cell boundary and nuclear shapes and trained a shape space for each collection of shapes independently (see *Materials and Methods*). Our approach consists of using the cell shapes of the neighbors of a given nuclear shape to predict a cell shape for that nuclear shape. Using hold-one-out cross-validation, we learned a kernel function to predict cell shape from nuclear shape (and vice versa) that minimized the sum of squared errors between the actual shape and the predicted shape over all but the held-out image. Using that learned kernel, we measured the error in predicting the held-out cell shape from its nuclear shape (as described in *Materials and Methods*). We used two methods to evaluate the quality of our shape predictions. The first was by measuring the frequency at which the error of the hold-out shape prediction was less than that from of a model trained

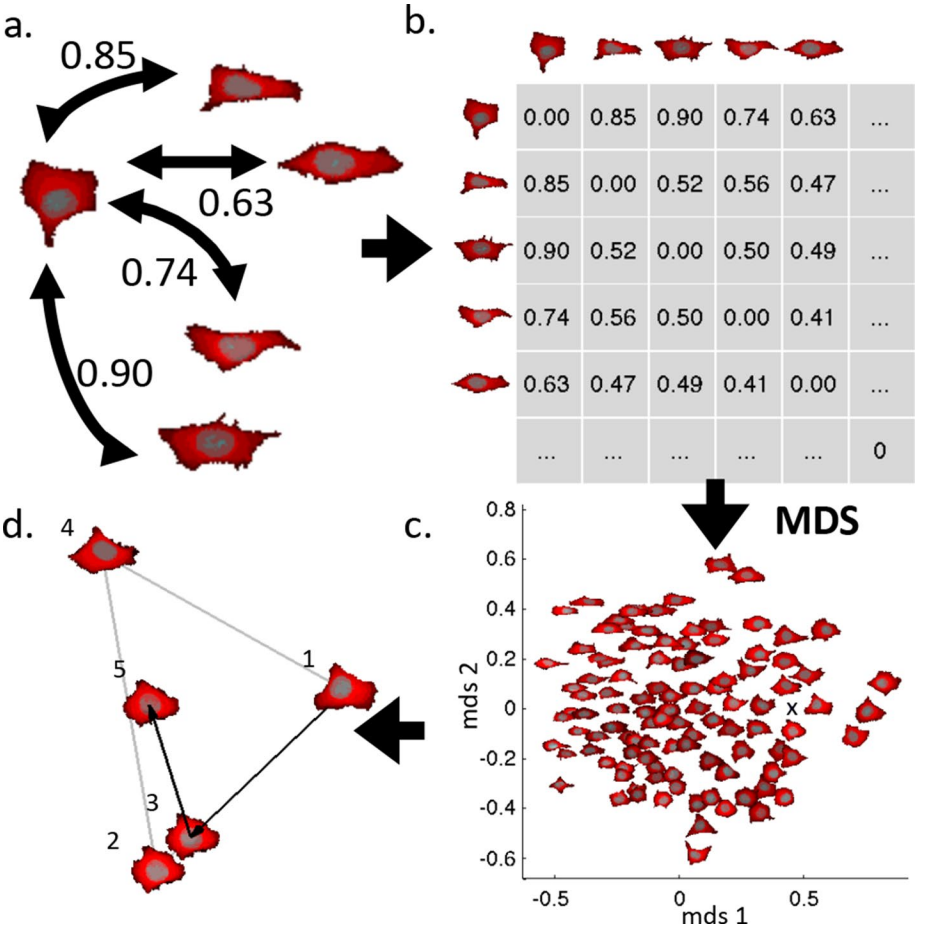

FIGURE 1: Shape-space-modeling pipeline. Diffeomorphic distances (a) are computed between each pair of images in a collection and loaded into a matrix (b). The distance matrix is embedded into a lower-dimensional space via multidimensional scaling (c). A shape can be synthesized (d) to correspond to any point in this space, as indicated with a black X in c. The shapes forming a simplex containing the target location (1, 2, 4) are iteratively interpolated (interpolate between shapes 1 and 2 to get shape 3, and between shapes 3 and 4 to get shape 5) to generate the target shape (5). The illustrations shown are for combined cell and nuclear shapes, but the process can equally be applied to just cell or nuclear shapes.

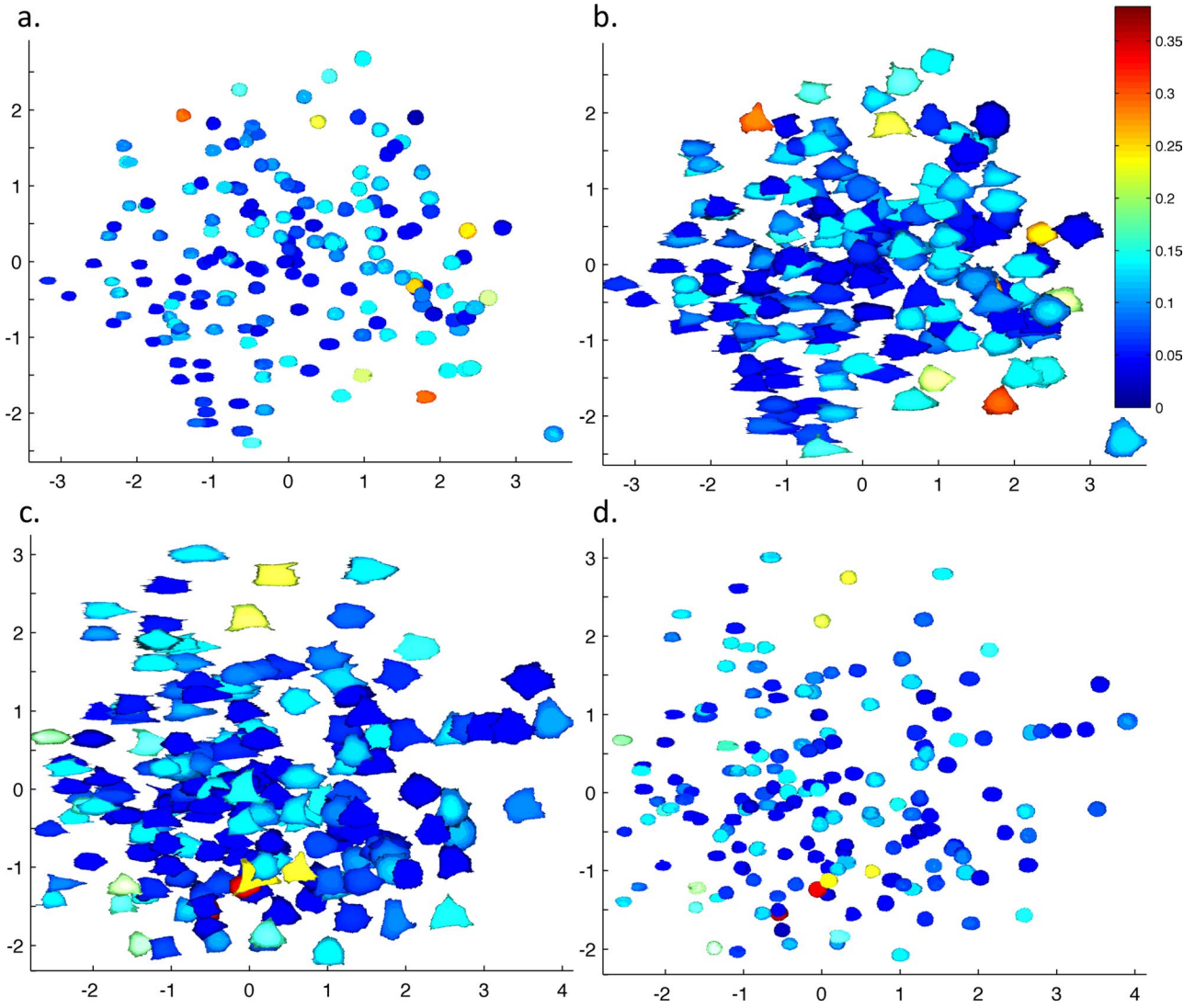

FIGURE 2: Predictive relationships between cell and nuclear shapes. (a) Shape space of 3D HeLa nuclear shapes, colored by *p* value estimated by the density method to show the significance of the ability to predict position of the nuclear shape corresponding to a given cell shape, where blue indicates strong predictive ability and red indicates poor predictive ability. (b) The cell shapes corresponding to the nuclei in a, plotted on the same coordinate space as a. (c) A shape space similar to the one in a, but for predicting cell shape from nuclear shape. (d) Nuclear shapes corresponding to the position of each cell in c.

on randomly matched cell–nuclear shape pairs over multiple permutations (which we call the permutation method). The second, more conservative approach tested the frequency at which the error of the shape prediction was less than what we would expect if we were to draw a shape from the approximate probability density of the to-be-predicted shape space at random. This was measured by comparing the error of the predicted shape to the distances between the target shape and all shapes in the collection (including itself; we call this the density method). With both methods, the mean squared error (MSE) for each condition was normalized to the expected MSE from that method (Supplemental Dataset S1 contains values used for the calculations). With the permutation model, the *p* values were bimodal; individual cells either showed a strong predictive relationship or they did not (Supplemental Figure S1A shows examples of accurate and inaccurate predictions). The normalized MSE across all predictions was determined to be 0.816 (with 77% of the predictions determined to be statistically significant at a 0.05 level) for predicting nuclear shape from cell shape and 0.835 (with 73% of the predictions significant) for predicting cell shape from nuclear shape. Thus the cell shape of most cells can be accurately predicted from its nuclear shape and vice versa. We also evaluated the predictions by the density method. This gave a normalized error of 0.398 when predicting cell shape from nuclear shape and 0.447 in the other direction, both of which are dramatically less than the value of 1 expected at random using this method. Figure 2 shows the results for the density method; shapes are colored by *p* value, with hotter colors indicating less predictive ability. It is important to note that for all of this analysis, the cell and nuclear shapes were segmented by independent methods, so that the correlation between the shapes observed for HeLa cells was not a result of the influence of the segmentation of one shape on the segmentation of the other.

Learning a joint model of 3D cell and nuclear shape for HeLa cells

Given our confirmation that cell and nuclear shape are dependent on each other, we constructed a joint cell-and-nucleus shape space

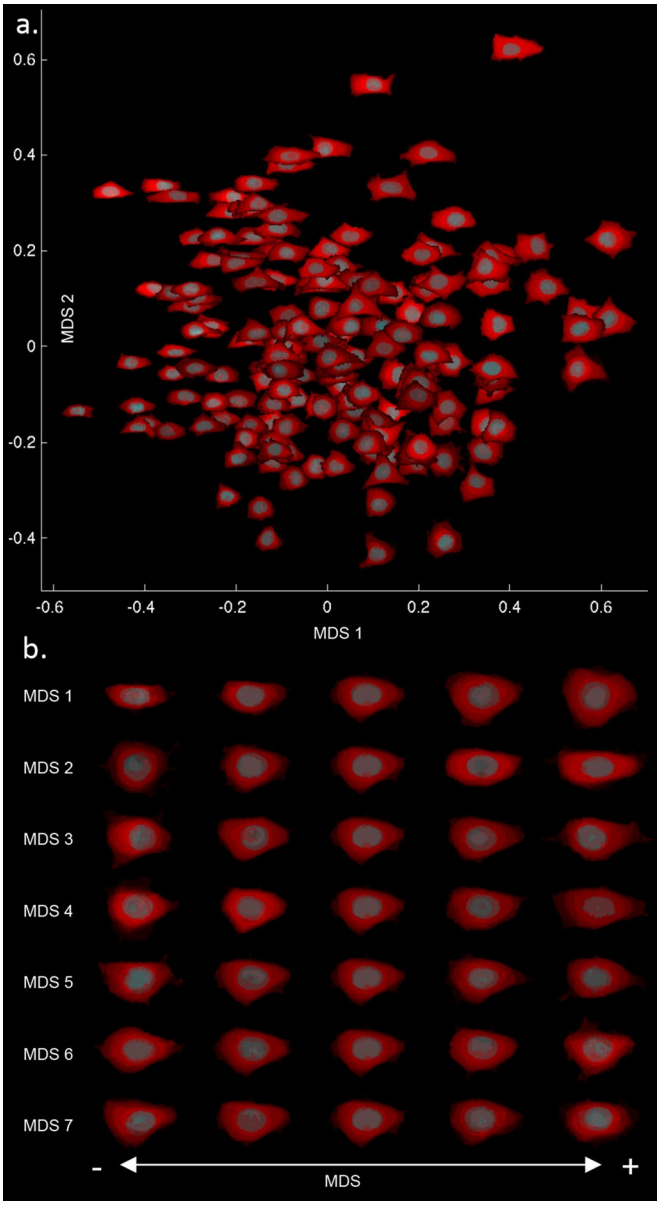

FIGURE 3: Shape space for 3D images of HeLa cells. (a) First two dimensions of a seven-dimensional HeLa shape space. (b) Synthesized cell and nuclear shapes across the principal dimensions of the completed shape space. Each row is a dimension of shape space with projections of the cell shape in the *z*-dimension, with the cell shape in red and the nuclear shape in cyan. Color intensity represents the relative thickness of that region.

for the 3D HeLa cell images. (Note that we use the term shape space somewhat loosely; rather than just measuring shape, our spaces encompass both size and shape.) Figure 3a shows the positions of joint cell and nuclear shapes in the first two dimensions of this shape space (see *Materials and Methods* for a discussion of how the dimensionality was chosen). Analogously to principal components, these two dimensions represent the two largest sources of variation within the shape space; these are typically referred to as the "major modes" of variation. We can see that the shapes vary smoothly across the dimensions of the figure, capturing variation in the size and eccentricity of cell shapes. To provide a visual representation of these major modes of variation across the shape space, Figure 3b shows joint shapes for points uniformly sampled across each dimension. We can see that the first dimension moves from smaller, eccentric cell shapes toward larger, rounder cell shapes. The second dimension starts with small, round cell shapes and moves to large, eccentric cell shapes. The interpretation of the other dimensions is less obvious.

Another way of interpreting these major modes of variation is by measuring the correlation between each mode and various interpretable, descriptive features. Supplemental Dataset S2 contains the results of such an analysis. It can be seen that the features showing correlation with the first major mode are primarily related to size and eccentricity.

Reducing the cost of shape-space computation

Construction of this shape space involved calculation of the pairwise distances for 175 cells. When seeking to construct shape spaces for larger cell image collections, the cost of computing the full distance matrix increases quadratically. An alternative is to estimate the shape space using only the distances of all shapes to a small set of "landmark" shapes (de Silva and Tenenbaum, 2004). (The idea is that one should be able to construct a map given distances of all cities to only a few cities and not need the distances between all pairs of cities.) To evaluate the performance of this distance completion procedure, we computed a complete distance matrix with the HeLa cell shapes for 106 cells. We simulated the matrix reconstruction procedure by randomly sampling a subset of "landmark" shapes, for which we measured the distance to all other shapes. For each set of randomly chosen landmarks, we found an approximate embedding according to Eq. 10 (see *Materials and Methods*) and measured the sum of squared errors between the true distances and estimated distances between the embedded shapes,

$$\mathrm{SSE}_D = \sum_{i,j}^{m,n} \left(d(x_i, x_j) - D_{i,j} \right)^2 \qquad (1)$$

where $D_{m,n}$ is the matrix of known pairwise distances, and $x_1,\ldots, x_n$ are the coordinates of the embedded positions of points $1,\ldots, n$. This measures how close the distances found using the landmarks were to the actual distances. We performed this analysis 10 times using randomly chosen landmark sets of different sizes, at each iteration embedding into Euclidean spaces from one to 15 dimensions, as well as the "full" embedding with one fewer dimension than there are shapes. Supplemental Figure S2 shows the mean and error of SSE_D, as well as the residual variance. We see that the error quickly drops when using at least 10% of the shapes as landmarks. We therefore used this percentage of shapes to build approximate shape spaces for the larger image collections given later.

Measuring alterations in the dependence of cell and nuclear shape for MCF7 cells

We next asked whether our conclusion that cell and nuclear shapes of HeLa cells are dependent on each other also applies to other cell types and whether we could identify drugs that alter this dependence. For this, we used 2D images of MCF7 from the Broad Biomage Benchmark Collection (Caie *et al.*, 2010). The collection contains cells treated with 113 compounds, each with one of 12 previously identified mechanisms of action. We selected one compound from each of the 12 mechanisms to perform our analysis (Supplemental Table S1). This gave us a total of 1639 cells for the 12 compounds and control; because this number was too large to compute a complete shape space, and given the success of landmark MDS described earlier, we used 272 landmark shapes to construct a seven-dimensional shape space using cell shapes and

264 landmark shapes to construct a seven-dimensional shape space using nuclear shapes, using the same methods as described for HeLa cells.

After normalizing the prediction errors via the density method described earlier, we compared the means of the errors across drug conditions with the pooled remaining conditions via ANOVA and Tukey's post hoc test (Tukey, 1949). As shown in Table 1, the average prediction error for the actin disruptor cytochalasin B was higher than average when predicting cell shape, suggesting a diversification of cell-shape phenotypes. The Aurora kinase inhibitor AZ-A had the opposite effect, with the predictive error decreasing with respect to nuclear shape prediction. We tested for all pairwise differences of mean; Figure 4A shows the results.

Measuring alterations in the dependence of cell and nuclear shape for H1299 cells

To extend the observed shape relationship to a third cell line and examine whether specific gene products could affect it, we used 2D images of H1299 non–small cell lung carcinoma from the Kahn Dynamic Proteomics Database (Sigal *et al.*, 2006, 2007). The H1299 data set contains movies for cell lines expressing different proteins tagged with yellow fluorescent protein (YFP). We used 28 movies for seven tagged proteins (four movies per protein). The movies show cells before and after addition of various drugs, but we used only the frames of the movies corresponding to the first 20 h of culture, before the addition of any drugs. Cell and nuclear shape masks were created using only the red fluorescent protein (RFP) channel (which primarily stains the nucleus but also shows mild cytoplasmic staining) so that they would not be affected by fluorescence from the protein that was tagged. A subset of frames that contained single cells was chosen as described in *Materials and Methods*.

As before, we trained independent cell and nuclear shape spaces at a dimensionality of seven using landmark MDS with 102 cell shape landmarks and 82 nuclear shape landmarks each containing 6515 shapes. Figure 5 shows the first two dimensions of the joint cell and nuclear shape space for H1299 cells, colored by protein label. The first two modes of the shape space account for morphological changes as a result of the protein label; the COX7C-labeled cells take a smaller, round conformation, and, on the opposite extreme, the C1QBP label results in larger cell shapes (although other dimensions contribute to the separability of the clones). The differences in shape among the different cell lines are highlighted by considering the regions that contain the most cells (where the probability density is highest), as shown in Figure 5b.

Because protein tagging may alter the function of the tagged protein, as well as change downstream interactions, we examined whether the presence of any of the protein tags altered the relationship between cell and the nucleus shapes (Table 2). As observed for HeLa cells, there is a significant degree of dependence between cell and nuclear shape. However, as shown in Figure 4, compared with tagging the other proteins, tagging C1QBP increases the error of both cell and nuclear shape prediction, suggesting a decorrelation of cell and nuclear shape and a broader range of shape phenotypes. On the other hand, tagging by COX7C does not seem to affect cell shape prediction, but it drastically reduces the error of nuclear shape, indicating a smaller range of nuclear shape phenotypes.

Modeling kinetics of cell and nuclear shape in H1299 cells

Because the images for these tagged lines are in fact movies, we can also ask how the evolution of cell and nuclear shape occurs over time. Because most of the cells that we analyzed were in G1, and hence the shape space distribution was dominated by G1, we chose to construct a cell shape transition model for cells within G1. To do this, we estimated the cell cycle phase of each cell by computing the integrated DNA intensity under the nuclear shape mask and clustered the cells into three groups (G1, S, G2) using *k*-means. We used only the cells that belonged to the lowest-intensity cluster centroid.

	E(cell|nuclear)		*E*(nuclear|cell)	
	Normalized MSE	*p* value avg err differs pooled population	Normalized MSE	*p* value avg err differs pooled population
Dimethyl sulfoxide (vehicle)	0.41	0.90	0.43	0.45
Cytochalasin B	0.44	0.02	0.45	0.78
AZ-A	0.41	0.98	0.38	0.00
Simvastatin	0.43	0.35	0.42	0.21
Etoposide	0.42	0.36	0.47	0.07
Floxuridine	0.44	0.17	0.46	0.29
AZ138	0.38	0.08	0.43	0.47
AZ-J	0.38	0.09	0.46	0.26
PD-169316	0.39	0.36	0.45	0.72
Demecolcine	0.38	0.07	0.45	0.85
Taxol	0.41	0.77	0.46	0.25
ALLN	0.40	0.76	0.46	0.30
Cycloheximide	0.42	0.45	0.44	0.67

See Supplemental Dataset S1 for supporting details. *E*(cell|nuclear) indicates the models in which cell shape is predicted from nuclear shape, and *E*(nuclear|cell) indicates the reverse. Normalized MSE was calculated as described in *Materials and Methods*. The *p* values shown are for the hypothesis that the average error for that condition is the same as the average error from all other conditions. ALLN, *N*-acetyl-leucine-leucine, norleucinyl.

TABLE 1: Statistical relationship between cell and nuclear shape for MCF7 cells.

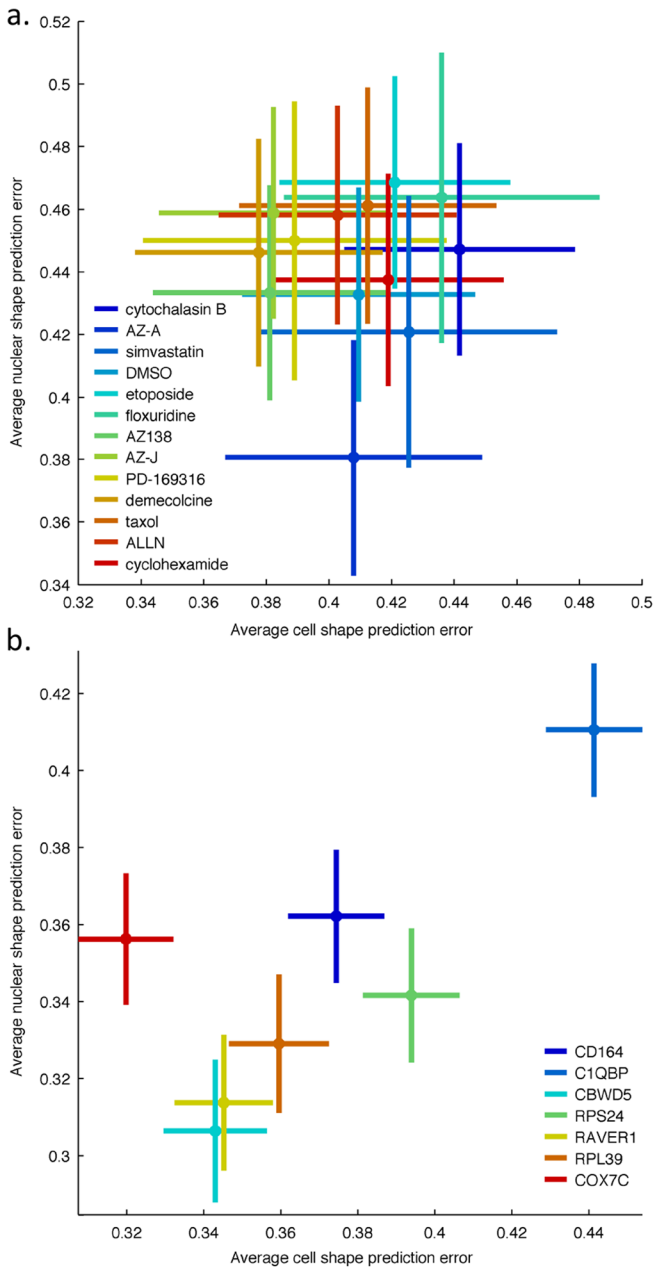

FIGURE 4: Testing of significance of changes in predictability of cell and nuclear shapes. Analysis of variance and Tukey's post hoc analysis on the means of the predictive error of cell shape and nuclear shape across all conditions for MCF7 (a) and H1299 (b) cells. Nonoverlapping bars indicate statistically significant differences.

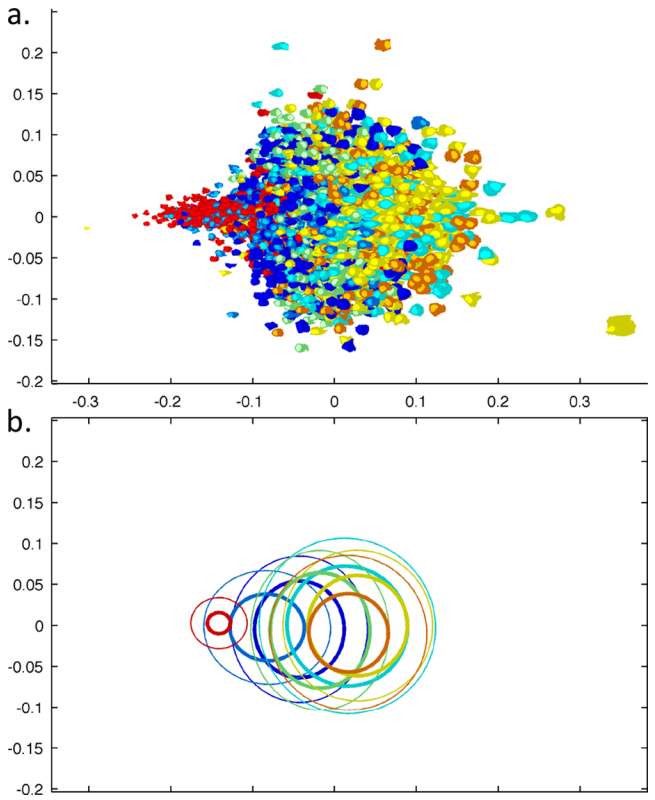

FIGURE 5: Shape analysis for H1299 cell clones. (a) H1299 shape space, colored by protein label. The labels dark blue, light blue, cyan, green, yellow, orange, and red correspond to CD164, C1QBP, CBWD5, RPS24, RAVER1, RPL39, and COX7C, respectively. (b) Contour lines for the 50th and 90th percentiles of probability density obtained via kernel density estimates are shown for each cell line with thin and thick lines, respectively.

We constructed a seven-dimensional shape space from these combined cell and nuclear shapes using landmark MDS with 102 landmarks.

To give an indication of the way in which shape evolves during the G1 phase, Figure 6 shows the direction in which cells are expected to move in the shape space using vectors showing the expected displacement in the first two dimensions and coloring for expected displacement magnitude in the third dimension. For each tagged line, the vectors all converge upon an average shape; however, this shape is different for each clone. Using these maps, we created a simple model of a walk through G1 by modeling the first frame of all the cells as a Gaussian distribution. We modeled the cell shape distribution as normal and randomly chose a starting cell shape and generated a walk through the cell shape space by taking steps in a directed random walk, using the expected displacement and covariance for the starting shape and the subsequent sampled positions (the step size was equal to the 20-min spacing in the original movies). We generated images corresponding to five shapes along each step, resulting in a movie with 4 min between each simulated frame. An example is included as Supplemental Video S1, and individual frames are shown in Figure 7.

DISCUSSION

A major goal of systems biology is to be able to create in silico models that reproduce the behaviors of eukaryotic cells. To do so, those models need to incorporate information on the spatial relationships between cellular components and the ways in which those relationships may change. Those spatial relationships include how the shape or position of one cellular component organelle influences the shape or position of others.

As a small step toward that end, we carried out the first characterization of the interdependence of cell and nuclear shape and demonstrated that the relationship is significant in both HeLa and H1299 cells. The majority of cells at any given time show a correlation between cell and nuclear shape. For HeLa and MCF7, both directions of prediction had similar accuracies. However, for H1299, generally nuclear shape could be predicted from cell shape better than the other way around. This asymmetry of prediction suggests

Tagged protein	*E*(cell|nuclear) Normalized MSE	*E*(cell|nuclear) *p* value avg err differs pooled population	*E*(nuclear|cell) Normalized MSE	*E*(nuclear|cell) *p* value avg err differs pooled population
CD164	0.37	0.30	0.36	0.05
C1QBP	0.44	0.00*	0.41	0.00*
CBWD5	0.34	0.00*	0.31	0.00*
RPS24	0.39	0.00*	0.34	0.51
RAVER1	0.35	0.00*	0.31	0.00*
RPL39	0.36	0.12	0.33	0.03*
COX7C	0.32	0.00*	0.36	0.21

See Table 1 footnote for definitions. Asterisks indicate cases where the *p* value is less than 0.05.

TABLE 2: Statistical relationship between cell and nuclear shape for H1299 cells.

the presence of subpopulations of cells with similar nuclear shapes but different cell shapes.

On the basis of these results, we created nonparametric, generative models that capture the relationships between cell and nuclear shape better than previous parametric approaches, which required unrealistic assumptions about allowable shapes. It will be of interest to see whether these models provide a more powerful framework for linking shape to molecular mechanisms than approaches based on descriptive features.

We also assessed the accuracy of constructing cell shape spaces for large image collections without computing all pairwise distances, using standard approaches for estimating full distance matrices.

Using images from the Broad Bioimage Benchmark Collection and the Kahn Dynamic Proteomics Database, we demonstrated that drug treatment or protein tagging not only can change cell shape, but can also disrupt the *relationship* between cell and nuclear shape. Specifically, we found that cytochalasin B causes a significant decorrelation of cell and nuclear shape, decreasing the ability to predict cell shape, whereas Aurora kinase inhibitors enhance this relationship (increasing the ability to predict nuclear shape). Inhibiting the mechanism of action of Aurora kinase has been known to disrupt the cell cycle, and it is likely that we are seeing those effects here (Vader and Lens, 2008). Furthermore, we found that tagging protein C1QBP led to a weaker association between the two shapes. C1QBP is a multifunctional, multicompartmental protein involved in a variety of processes. It is primarily a mitochondrial protein, and evidence has been presented that changes in expression of the normal protein affect apoptosis, cell proliferation, and migration (McGee *et al.*, 2011). We conjecture that tagging C1QBP leads to slower proliferation and larger cells, disrupting the normal relationship between cell and nuclear shape.

Finally, we presented the first image-derived generative model of cell shape kinetics, and used it to create synthetic movies of cell and nuclear shape change. Such models do not require tracks of single cells over extended periods and are therefore simple and efficient to create. The models capture how much variation in shape is permitted for cells of a given type or under a given condition.

Although the nonparametric, generative models we used have some significant advantages, these do not come without a cost. The primary disadvantage compared with either descriptive approaches or parametric generative models is the cost of computing large numbers of pairwise diffeomorphisms. Although this can be reduced by using landmark MDS, it is still significant. For the images used in our studies, the cost of computing one pairwise distance is ~13 s for a 144 × 144 × 14–pixel 3D image and 3 s for a 148 × 148–pixel 2D image. This compares to ~1 s for determining the cell-shape parameterization of equivalent-sized images with our previous 3D and 2D parametric models and <1 s for calculation of the descriptive features in Supplemental Dataset S2. The advantages include the potential ability to observe changes not captured by descriptive features (as suggested by the observation of shape modes that do not correlate with descriptive features in Supplemental Dataset S2), directly model joint relationships between cellular components, and build models of predictive relationships that allow for prediction of the actual shape rather than shape descriptors.

The software used here for both training of and synthesis from 2D and 3D joint shape space models has been added to the open-source CellOrganizer system for cell modeling (Buck *et al.*, 2012; Murphy, 2012; http://cellorganizer.org/). A curated, open-access repository for public deposition of models created for other cell types or conditions is also available. These models can be combined with models for other cell components and with biochemical models to explore the relationships among shape, organelle distribution, and cellular biochemistry.

MATERIALS AND METHODS

Image collections

We used three image collections for the studies described here. For analysis of 3D shapes, we used a previously described (Velliste and Murphy, 2002) collection of 3D images of HeLa cells obtained via confocal microscopy (available from http://murphylab.web.cmu.edu/data/3Dhela_images.html). We used only the nuclear and total protein channels for 175 cells from this collection. For analysis of 2D shapes and their dynamics, we used 2D time series images of seven different labeled protein lines of H1229 cells from the Khan Dynamic Proteomics Database (Sigal *et al.*, 2006, 2007; available at http://www.weizmann.ac.il/mcb/UriAlon/DynamProt). The movies we used were from cells tagged in CD164, C1QBP, CBWD5, RPS24, RAVER1, RPL39, and COX7C and contained channels for the tagged protein (YFP) and also a nuclear marker (RFP). From each movie, we chose 125 transition pairs (cells that appeared in two adjacent frames) and built a shape space of 6515 segmented cell and nuclear shapes (the same shape can be used in one or two transition pairs). We also used images of MCF7 cells treated with compounds identified as having 1 of 12 different mechanisms of action provided from the Broad Bioimage Benchmark Collection 21v1 (Caie *et al.*, 2010). Supplemental Table S1 gives a complete list of conditions and mechanisms of action.

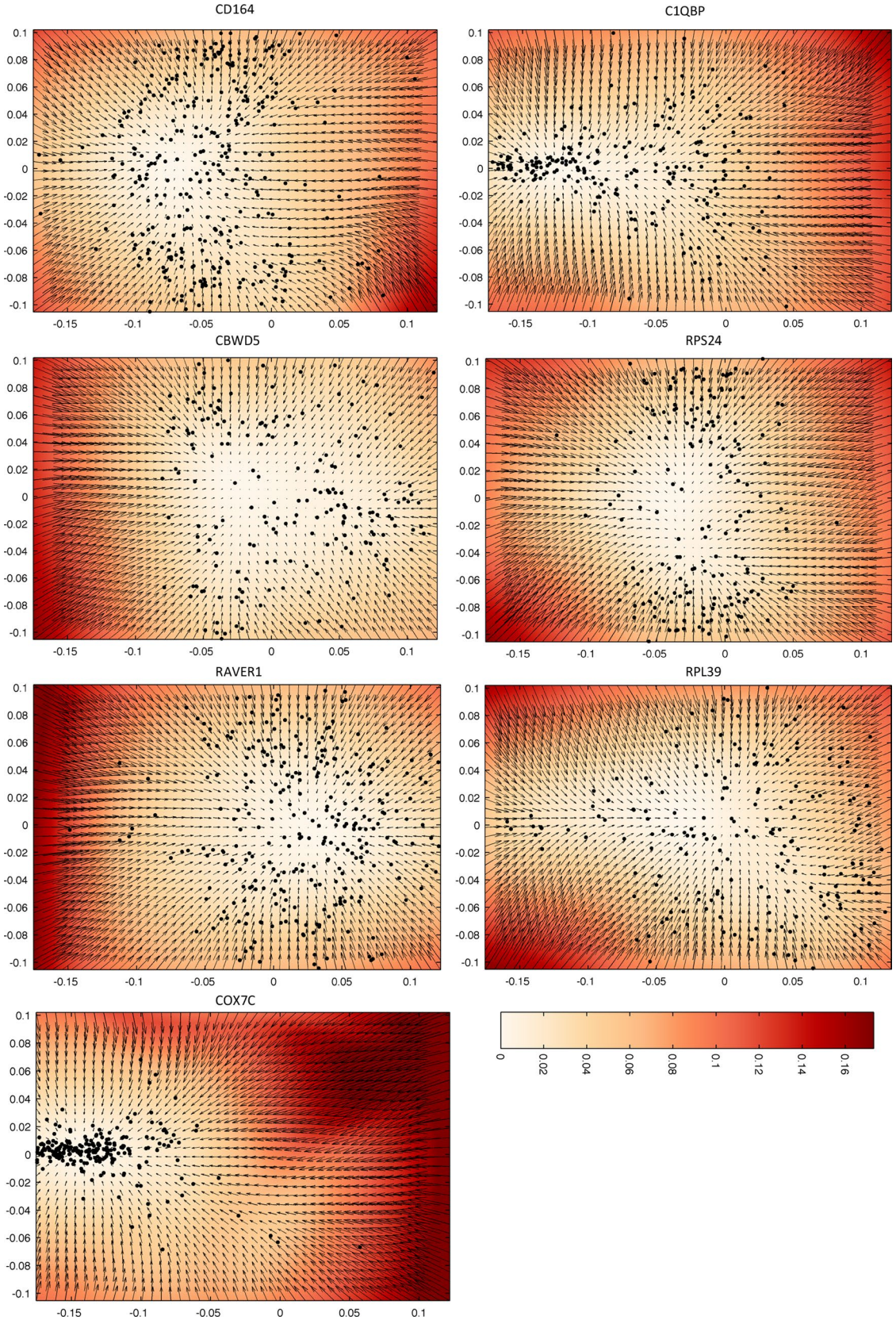

FIGURE 6: Expected H1299 cell shape displacement between consecutive frames for cells in G1. Displacement fields are shown in the first two dimensions of the shape space; the direction of highest probability displacement at each location is indicated using vectors for the *x*- and *y*-dimensions, and velocity is indicated by color temperature. The vectors have been scaled for visualization purposes. Note that locations in the shape space where the vectors have zero length and the color is white represent steady states.

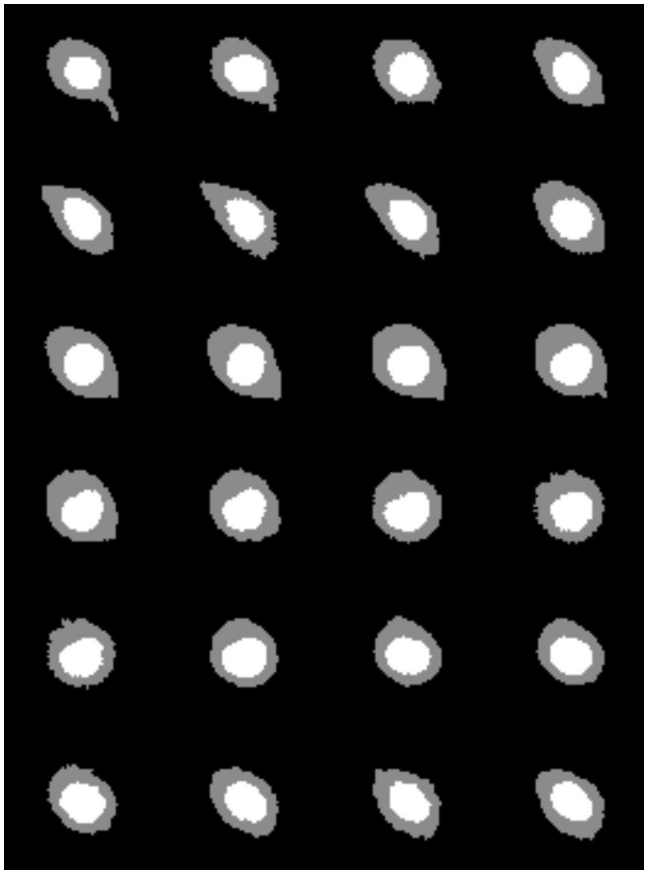

FIGURE 7: Individual frames from a movie of simulated cell shape changes. The first 24 frames (corresponding to every 4 min) are shown, with the frames ordered from left to right and top to bottom.

Image processing and segmentation

Owing to variation in microscope imaging parameters, protein labels, and the nature of the imaging experiments, we used data set–specific segmentation methods. Example images from our data set and corresponding segmentations are shown in Supplemental Figure S3.

HeLa cell and nucleus segmentation

Masks for each cell and nucleus were created using the total protein (Cy5) channel and the DNA (propidium iodide) channel by first blurring both the protein and DNA channels with a Gaussian filter with SD of 0.5 pixel. The nuclear shapes were determined by Ridler–Calvard thresholding of the DNA channel (Ridler and Calvard, 1978). Cell-shape pixels in the total protein channel were determined as all pixels with intensity above the most populous bin in a 124-bin histogram of the protein channel pixel intensities, and only the largest connected component was retained. This resulted in independently determined cell and nuclear shapes but did not address segmentation into single-cell regions. However, because most of the images in this data set contained only one cell, we avoided this by simply discarding the small number of cell shapes that contained more than one nucleus (more than one connected component after nuclear thresholding).

MCF7 cell and nucleus segmentation

4′,6-Diamidino-2-phenylindole (DAPI), actin, and tubulin images were blurred with a Gaussian filter with SD of 1, and intensity was rescaled from 0 to 1 based on the minimum and maximum intensity of the image, respectively. A DNA "edginess" image was created via Frangi filter (Frangi *et al.*, 1998) with SD of 6 on the DAPI image. This was then thresholded via the Ridler–Calvard method (Ridler and Calvard, 1978) to find individual nuclear regions. Refined nucleus–nucleus boundaries were found by using the regions as seeds for seeded watershed, and background was set to be all pixels less than an intensity of 0.1. A combined image was created by adding the blurred DAPI, actin, and tubulin and rescaling the resulting image. Because the nuclei are always inside the cell, the previously found nuclear regions were used as seeds in seeded watershed segmentation on this combined image to determine the cell–cell boundaries. All pixels less than an intensity of 0.1 were set as background in the combined image as well.

H1229 cell and nucleus segmentation

Masks were created using the RFP channel only by first blurring the image with a Gaussian filter at 1 SD. A background-adjusted intensity image was created by subtracting the result of processing the blurred image with a 100-pixel-wide averaging filter. Nuclear regions were determined by Otsu thresholding of the background-adjusted image (Otsu, 1979), and touching nuclei were separated by distance transforming the threshold image and watershed transforming the result. Cell regions were defined as connected components of pixels 3 bins above the most populous bin of a 32-bin histogram of the background-adjusted image. Cell regions containing more than one nuclear region or a nuclear region with an area less than that of a five-pixel-radius circle were discarded. To avoid complications due to possible influences of cell–cell contact, we selected frames containing cells that were not touching any other cells in two consecutive frames, as well as where the cell regions overlap each other across those frames.

Diffeomorphic distances and shape-space generation

The LDDMM is a measure of distance over a smooth, invertible, nonrigid deformation (a diffeomorphism) between a pair of images (Beg *et al.*, 2005). Given an image pair I_0 and I_1, the transformation that maps each location in I_0 to a corresponding location in I_1 is computed from the flow of a time-dependent vector field v,

$$\frac{\delta g(x;t)}{\delta t} = v(g(x;t);t) \tag{2}$$

where $g(x,0) = x$. The vector field is determined as a minimization of the function

$$\int_0^1 \| Lv(x;t) \|^2 \, dt + \int_\Omega | I_0(g(x,1)) - I_1(x) |^2 \, dx \tag{3}$$

which minimizes the time-dependent vector field deforming one shape into another, v_t, and the difference of the two images after alignment, $\int_\Omega | I_0(g(x,1)) - I_1(x) |^2 \, dx$. In Eq. 3, L represents a linear differential operator such as $\nabla^2 + \lambda I$. Given the resulting vector field that minimizes Eq. 3, the diffeomorphic distance between the two shapes is defined as

$$d(I_0, I_1) = \int_0^1 \| Lv_t \| \, dt \tag{4}$$

To find these differences, we used the approach used previously for characterizing nuclear shapes (Rohde *et al.*, 2008a; Peng *et al.*, 2009). This involves a greedy optimization method for finding the diffeomorphism (Beg *et al.*, 2005) and a symmetric version of the problem in Eq. 3 that iteratively deforms images toward each other to enforce symmetry in the distance measure (Avants and Gee, 2004; Joshi *et al.*, 2004).

Given a collection of n shapes, we computed an $n \times n$ distance matrix, D, where each entry of the distance matrix corresponds to the diffeomorphic distance between two shapes, $D_{i,j} = \mathrm{d}(I_i, I_j)$. Using MDS, we embedded this distance matrix into a Euclidean space where each image I_i was assigned a coordinate x_i such that the distances between the embedded image coordinates approximate the distances in the distance matrix,

$$\{x_1, \ldots, x_n\} = \arg\min_{\{x_1,\ldots,x_n\}} \sum_{i<j\leq n} (d(x_i, x_j) - D_{i,j})^2 \quad (5)$$

In other words, given the measured distances between shapes, we can reconstruct an underlying space from which the shapes were observed that captures their relative similarities and differences. When constructing joint cell and nuclear shape spaces, we encoded each segmented cell as a ternary image, with a pixel value of 0, indicating that this pixel was outside the cell; a value of 1, indicating that it was inside the cell but not inside the nucleus; and a value of 2, indicating that it was inside the nucleus.

Shape-space dimensionality

To determine the embedding dimensionality of the shape space, we calculated the residual variance, $1 - R^2(D, D')$, between the diffeomorphic distance matrix and an approximate distance matrix, D', using the embedded coordinates from Eq. 5. The so-called "intrinsic dimensionality" of the shape space was determined to be the dimensionality at which an "elbow" occurs when plotting the residual variance as a function of embedding dimension (Tenenbaum *et al.*, 2000; Rohde *et al.*, 2008b). For our experiments, we chose a dimensionality of seven, as it is the approximate position of this "elbow," sufficient to reconstruct known images, and is also a practical limitation due to the computational cost of computing a Delaunay triangulation (see earlier discussion) and the size of the output.

Shape synthesis

Given a point in the convex hull of our embedded space, we can synthesize a shape corresponding to that point by deforming the known shapes that form the simplex containing that point (Peng *et al.*, 2009). Simplices were determined by a Delaunay triangulation. We determined a probability density function via kernel density estimation over the embedded space of shapes to permit novel shapes to be sampled representative of the training image distribution.

Relationship between cell and nuclear shape

Given a shape space of cell shape and a shape space of their corresponding nuclei, we can build a model that allows us to predict the cell shape of a cell from its nuclear shape and vice versa. Given a collection of n shape pairs, $\left\{\left\{x_1^0, x_1^1\right\}, \ldots, \left\{x_n^0, x_n^1\right\}\right\}$, where x^0 is the shape space coordinate for a cell shape, and x^1 is the shape space coordinate for the corresponding nuclear shape, we constructed a predictive model of nuclear shape given cell shape (or vice versa) using the weighted average nuclear shapes of neighbors in cell shape space,

$$E(x^1 \mid x^0) = \frac{\sum_i^n w_i x_i^1}{\sum_i^n w_i} \quad (6)$$

where w_i is a weighting function,

$$w_i = k_h(x^0 - x_i^0) \quad (7)$$

Here we chose $k_h(x^0 - x_i^0)$ to be a Gaussian kernel weighting function with bandwidth h. We determined the size of the neighborhood that gives the best approximation of the corresponding shape by learning a kernel bandwidth that minimizes the mean squared error after hold-one-out cross-validation,

$$\mathrm{MSE}_h = \min_h \frac{1}{n} \sum_i^n \left\| x_i^1 - \frac{\sum_{j\neq i}^n w_i x_j^1}{\sum_{j\neq i}^n w_i} \right\|_2^2 \quad (8)$$

Given the bandwidth h for each held-out cell, we predicted its shape and thereby determined the MSE across all shapes. (The chosen bandwidths varied among cell types but were quite similar within a cell type; their average and SDs are listed in Supplemental Dataset S1.) We used a permutation test to construct a p value on this error compared with the null hypothesis that there was no correlation between cell and nuclear shape. In addition, we computed a normalized MSE by dividing by the average MSE from the permutation tests (thus measuring how much better than random the average prediction was; values <1 are better). We also used a second, more conservative statistical test in which we compared the prediction error to the distribution of errors that we would expect by randomly sampling according to the probability density of the shape space by measuring the error between the held-out shape and the predicted shape and measuring the frequency at which that was less than the pairwise distances between the held-out and all shapes (including itself). We also normalized MSE according to the average MSE from the probability density method.

Shape dynamics

To model the expected shape in the next frame given a current shape, we used kernel density estimation over shape transitions at subsequent time points. Given a collection of n sequential shape transitions, $\left\{\left\{x_1^{s0}, x_1^{s1}\right\}\right\}, \ldots, \left\{\left\{x_n^{s0}, x_n^{s1}\right\}\right\}$, where x^{s0} is a shape space coordinate of type s (0 = cell, 1 = nucleus, 2 = both) at a given time point (not necessarily time 0 of the movie) and x^{s1} is the shape-space coordinate at the subsequent time point, we computed the expectation of the shape at the next time step as a weighted average of the transitions of its neighbors using Eq. 6 and learning a bandwidth similarly to Eq. 8 (but $w_i = K_h(x_i^{s1} - x_i^{s0})$). In addition, we can model the variance of the step with the foregoing kernel by computing a covariance matrix from the residuals of the neighbors,

$$\mathrm{COV}_{j,k}\left(x^{s1} \mid x^{s0}\right) = \frac{\sum_i^n w_i}{\left(\sum_i^n w_i\right)^2 - \sum_i^n w_i^2} \sum_i^n w_i \left(x_{i,j}^{s1} - E\left(x^1 \mid x^0\right)_j\right) \times \left(x_{i,k}^1 - E\left(x^1 \mid x^0\right)_k\right) \quad (9)$$

This is a relatively simple, first-order model, and dynamics models can increase in sophistication with the availability of data.

Shape spaces for incomplete data

To construct a shape space, given values for only a subset of the full distance matrix $D_{i,j}$, we found an embedding that satisfies

$$\{x_1, \ldots, x_n\} = \arg\min_{\{x_1,\ldots,x_n\}} \sum_{i<j\leq n} w_{i,j} \left(d\left(x_i, x_j\right) - D_{i,j}\right)^2 \quad (10)$$

where $\{x_1, \ldots, x_n\}$ are the coordinates of the embedded shapes 1 through n, and $w_{i,j}$ is a weight indicating the relative importance of

that distance observation. As in typical MDS implementations, this weighting is an indicator that is 1 if the $d_{i,j}$ is observed and 0 otherwise. Here $d(a, b)$ is the Euclidean distance between vectors a and b. Due to the presence of the weight matrix W, we do not need to observe all pairs of distances as long as $D_{i,j}$ does not comprise disjoint subgraphs, and there are at least $N - 1$ unique paths from any subset to any other subset of points, where N is the dimensionality of the embedding.

ACKNOWLEDGMENTS

We thank Gaudenz Danuser and Armaghan Naik for helpful suggestions. This work was supported in part by National Institutes of Health Grants GM090033, GM103712, and EB009403.

REFERENCES

Avants B, Gee JC (2004). Geodesic estimation for large deformation anatomical shape averaging and interpolation. NeuroImage 23(Suppl 1), S139–S150.

Beg MF, Miller MI, Trouve A, Younes L (2005). Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int J Comput Vis 61, 139–157.

Buck TE, Li J, Rohde GK, Murphy RF (2012). Toward the virtual cell: automated approaches to building models of subcellular organization "learned" from microscopy images. BioEssays 34, 791–799.

Caie PD, Walls RE, Ingleston-Orme A, Daya S, Houslay T, Eagle R, Roberts ME, Carragher NO (2010). High-content phenotypic profiling of drug response signatures across distinct cancer cells. Mol Cancer Ther 9, 1913–1926.

Dahl KN, Scaffidi P, Islam MF, Yodh AG, Wilson KL, Misteli T (2006). Distinct structural and mechanical properties of the nuclear lamina in Hutchinson-Gilford progeria syndrome. Proc Natl Acad Sci USA 103, 10271–10276.

de Silva V, Tenenbaum JB (2004). Sparse multidimensional scaling using landmark points. Technical Report, Stanford University, Stanford, CA. Available at http://pages.pomona.edu/~vds04747/public/papers/landmarks.pdf (accessed 15 June 2015).

Elliott H, Fischer RS, Myers KA, Desai RA, Gao L, Chen CS, Adelstein RS, Waterman CM, Danuser G (2015). Myosin II controls cellular branching morphogenesis and migration in three dimensions by minimizing cell-surface curvature. Nat Cell Biol 17, 137–147.

Frangi AF, Niessen WJ, Vincken KL, Viergever MA (1998). Multiscale vessel enhancement filtering. In: Medical Image Computing and Computer-Assisted Interventation—MICCAI'98, Berlin: Springer, 130–137.

Hagwood C, Bernal J, Halter M, Elliott J, Brennan T (2013). Testing equality of cell populations based on shape and geodesic distance. IEEE Trans Med Imaging 32, 2230–2237.

Joshi S, Davis B, Jomier M, Gerig G (2004). Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage 23(Suppl 1), S151–S160.

Khatau SB, Hale CM, Stewart-Hutchinson PJ, Patel MS, Stewart CL, Searson PC, Hodzic D, Wirtz D (2009). A perinuclear actin cap regulates nuclear shape. Proc Natl Acad Sci USA 106, 19017–19022.

Kihara T, Haghparast SM, Shimizu Y, Yuba S, Miyake J (2011). Physical properties of mesenchymal stem cells are coordinated by the perinuclear actin cap. Biochem Biophys Res Commun 409, 1–6.

McGee AM, Douglas DL, Liang Y, Hyder SM, Baines CP (2011). The mitochondrial protein C1qbp promotes cell proliferation, migration and resistance to cell death. Cell Cycle 10, 4119–4127.

Murphy RF (2012). CellOrganizer: image-derived models of subcellular organization and protein distribution. Methods Cell Biol 110, 179–193.

Otsu N (1979). A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybernet 9, 62–66.

Peng T, Murphy RF (2011). Image-derived, three-dimensional generative models of cellular organization. Cytometry A 79A, 383–391.

Peng T, Wang W, Rohde GK, Murphy RF (2009). Instance-based generative biological shape modeling. Proc IEEE Int Symp Biomed Imaging 5193141, 690–693.

Pincus Z, Theriot JA (2007). Comparison of quantitative methods for cell-shape analysis. J Microsc 227, 140–156.

Ridler TW, Calvard S (1978). Picture thresholding using an iterative selection method. IEEE Trans Syst Man Cybernet SMC-8, 630–632.

Rohde GK, Ribeiro AJ, Dahl KN, Murphy RF (2008a). Deformation-based nuclear morphometry: capturing nuclear shape variation in HeLa cells. Cytometry A 73, 341–350.

Rohde GK, Wang W, Peng T, Murphy RF (2008b). Deformation-based nonlinear dimension reduction: applications to nuclear morphometry. Proc 2008 Int Symp Biomed Imaging 2008, 500–503.

Sailem H, Bousgouni V, Cooper S, Bakal C (2014). Cross-talk between Rho and Rac GTPases drives deterministic exploration of cellular shape space and morphological heterogeneity. Open Biol 4, 130132.

Sigal A, Danon T, Cohen A, Milo R, Geva-Zatorsky N, Lustig G, Liron Y, Alon U, Perzov N (2007). Generation of a fluorescently labeled endogenous protein library in living human cells. Nat Protoc 2, 1515–1527.

Sigal A, Milo R, Cohen A, Geva-Zatorsky N, Klein Y, Alaluf I, Swerdlin N, Perzov N, Danon T, Liron Y, *et al.* (2006). Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins. Nat Methods 3, 525–531.

Tenenbaum JB, de Silva V, Langford JC (2000). A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323.

Tsygankov D, Bilancia CG, Vitriol EA, Hahn KM, Peifer M, Elston TC (2014). CellGeo: a computational platform for the analysis of shape changes in cells with complex geometries. J Cell Biol 204, 443–460.

Tukey JW (1949). Comparing individual means in the analysis of variance. Biometrics 5, 99–114.

Vader G, Lens SM (2008). The Aurora kinase family in cell division and cancer. Biochim Biophys Acta 1786, 60–72.

Velliste M, Murphy RF (2002). Automated determination of protein subcellular locations from 3D fluorescence microscope images. Proc 2002 IEEE Int Symp Biomed Imaging 2002, 867–870.

Yin Z, Sadok A, Sailem H, McCarthy A, Xia X, Li F, Garcia MA, Evans L, Barr AR, Perrimon N, *et al.* (2013). A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes. Nat Cell Biol 15, 860–871.

Yin Z, Zhou X, Bakal C, Li F, Sun Y, Perrimon N, Wong ST (2008). Using iterative cluster merging with improved gap statistics to perform online phenotype discovery in the context of high-throughput RNAi screens. BMC Bioinformatics 9, 264.

Yu D, Pham TD, Zhou X, Wong SC (2009). Recognition and analysis of cell nuclear phases for high-content screening based on morphological features. Pattern Recogn 42, 498–508.

Zhao T, Murphy RF (2007). Automated learning of generative models for subcellular location: building blocks for systems biology. Cytometry A 71A, 978–990.