A robust method to quantify cellular and nuclear morphology and phenotypic heterogeneity

Jude M. Phillip[1, 5,*], Kyu-Sang Han[1,*], Wei-Chiang Chen[1], Denis Wirtz[1-4, +], Pei-Hsun Wu[1, +]

[1] Department of Chemical and Biomolecular Engineering, Johns Hopkins Physical Sciences Oncology Center, Johns Hopkins Institute for Nanobiotechnology (INBT), Johns Hopkins University, Baltimore, Maryland 21218, USA; [2] Department of Pathology, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, USA; [3] Department of Oncology, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, USA; [4] Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, USA; [5] Department of Medicine, Division of Hematology and Oncology, Weill Cornell Medicine, New York, New York, 10065, USA;

[*] These authors contributed equally to this work;

[+]Correspondence should be addressed to: P-H. W pwu@jhu.edu and D.W. wirtz@jhu.edu

Keywords: cell morphology; single-cell phenotyping; shape modes; cellular heterogeneity; entropy

Abstract

Cell morphology encodes essential information on many underlying biological processes. It is commonly used by clinicians and researchers in the study, diagnosis, prognosis, and treatment of human diseases. Quantification of cell morphology has seen tremendous advances in recent years. However, effectively defining morphological shapes and evaluating the extent of morphological heterogeneity within cell populations remain challenging. Here we present a protocol and software for the analysis of cell and nuclear morphology using the VAMPIRE algorithm. This algorithm enables the profiling and classification of cells into shape modes based on equidistant points along cell and nuclear contours. Examining the distributions of cell morphologies across automatically identified shape modes provide an effective visualization scheme that relates cell shapes to cellular subtypes based on endogenous and exogenous cellular conditions. In addition, these shape mode distributions offer a direct and quantitative way to measure the extent of morphological heterogeneity within cell populations. This protocol is highly automated and fast, with the ability to quantify the morphologies from two-dimensional projections of cells seeded both on two-dimensional substrates or embedded within three-dimensional microenvironments, such as hydrogels and tissues.

Introduction

Cell morphology is commonly employed by clinicians and researchers in the study, diagnosis, prognosis, and treatment of human diseases. Fundamentally, cellular morphology represents the ensemble imprints of highly interactive molecular networks, that include metabolic, proteomic, epigenomic, and genomic components[1–6]. The coordinated orchestration of these inter-dependent cellular programs are critical to properly govern cellular behavior[4] and ultimately determine cellular responses to perturbations and stressors, mainly microenvironmental cues[7,8], biomechanical stimuli[9,10], and pharmacological treatments[11–13]. Advances in high-content imaging[6,14,15], image processing[16,17], and machine learning[18–21] have greatly improved the throughput and accuracy of cell morphological measurements and have bolstered its utility in digital pathology[22–25], biomarker identification[1,26], and phenotypic screens[12,27–29].

Cell morphology is traditionally quantified using a handful of geometric parameters[14,30], delineating the size (e.g. area, perimeter) and shape (e.g. shape factor, aspect ratio) of cells and their corresponding nuclei. These measures are often complemented by fluorescence readouts of protein expressions, together with intensity patterns and localization within cells. Measuring cell and nuclear sizes can be readily achieved using open-source software platforms, such as CellProfiler[31,32] and ImageJ/Fiji[33]. However, defining and quantifying cellular shapes are more complicated.

Classically, shape descriptors, such as shape factor ($4\pi A/P^2$, where A is the area of the object and P is perimeter), aspect ratio (long axis length/short axis length) and eccentricity (see **glossary**), all measure the deviation of a cell's shape from a circle. While these geometric parameters are geared towards biological simplicity and provide the ability to quickly and directly detect differences among tested cellular conditions, these parameters tend to insufficiently capture the true complexities of cell shapes[1].

To illustrate this, we described the morphologies of mouse embryonic fibroblasts (MEFs) using conventional shape features, including shape factor, aspect ratio, and solidity (see **glossary**). From this analysis, we observed that taking a subset of cells having highly similar values of these parameters, still resulted in a high degree of morphological variability among individual cells. Underscoring the notion that conventional cell morphology

parameters may be insufficient to capture cellular differences (**Figure 1**). Furthermore, mesenchymal cells on flat substrates or cells embedded within physiologically relevant 3D collagen gels, which often feature extensive dendritic protrusions and nuclear blebs[37,39–41], are similarly difficult to distinguish using these traditional parameters.

A popular approach to address this shortcoming consists of defining additional geometric and statistical descriptors of cells, some of which are based on the curvature and roughness of the cell and nuclear contours[14,30,42]. This has led to an expansion of morphological descriptors, with the premise that these additional descriptors would help to better define and differentiate cellular subtypes. While increasing the number of shape descriptors allows users to capture more complex cell morphologies, visualizing differences in cell morphology and assigning biological meaning for these additional morphology descriptors are challenging.

To address this challenge, we recently developed a cell morphology analysis software that provides improved visualization and quantitative analysis of complex shape morphologies. The software, which we named Visually Aided Morpho-Phenotyping Image Recognition (VAMPIRE), is highly automated and allows users to rapidly process large datasets of post-segmented images of cells and/or their corresponding nuclei.

Development of the protocol

VAMPIRE analysis was initially developed to better interpret morphological data that we acquired for a set of 11 pancreatic cancer cell lines using a custom high-throughput microscopy imaging system[1]. Our goal was to identify a potential morphological signature of metastasis in pancreatic ductal adenocarcinoma (PDAC). Among the samples used, five were collected from patient-derived primary tumors, four were obtained from liver metastases, and two were non-neoplastic pancreatic epithelial cell lines. For direct visual assessment of cell and nuclear shapes, we randomly selected subsets of individual cell contours (after alignment) and found no overt morphological differences between primary tumor cells and liver-metastasis cells, which was likely due to the irregularities of cell shapes.

To quantify cell shapes, we used commonly defined morphological features, such as cell area, shape factor, and aspect ratio. However, these features could not reflect the observed extent of cell shape variations, since even a small subset of cells displaying an

extremely narrow range of values of these conventional shape descriptors appeared radically different from each other.

To address this problem, we established and validated VAMPIRE analysis, which provides morphological information beyond classically defined geometric parameters[1,6,26]. VAMPIRE analysis is a visual aid that compares cell morphologies by first identifying representative shape modes (see **Glossary**) among all cell shapes present within a cell population. Then using these shape modes, VAMPIRE determines the abundance of cells classified within each shape mode per condition. VAMPIRE comprises four essential computational steps: I) the determination and registration of the coordinates of equally-spaced points along cell and nuclear contours to define morphological descriptors; II) the reduction of the number of morphological descriptors using principal component analysis (PCA); III) the identification of shape modes through unsupervised K-means clustering analysis, and IV) the determination of abundances and distributions of cells within each shape mode for all tested cell samples and conditions (**Figure 2**).

Following segmentation, the coordinates for points along the contour of each cell and its corresponding nuclei are aligned, scaled, and shifted to unify the sizes and reduce shape variations due to rotational variations and mirror effects. Briefly, the alignment of cell and nuclear shapes was done based on Procrustes analysis [37,43,44]. To represent the highly complex shapes of cells and nuclei, a sufficient number of equally spaced points along each contour (typically 50 points) (**Figure 2A**) is used to define high-dimensional "features". Then these coordinates along the boundaries of each cell and/or nuclei are subtracted by their mean value to shift the center of each cell and/or nuclei to the location (0,0). To normalize each contour and reduce the contributions from the cell and nuclear sizes, a characteristic length scale is determined for each cell and/or nuclei, based on the following equation:

$$R = \sqrt{\sum_{i=1}^{50} (x_i^2 + y_i^2)/50}$$

where R is the characteristic length scale, and x and y are the coordinates along the shape boundary/contour.

Using the value of R calculated for each cell and/or nuclei, shapes are then

normalized by dividing the contour coordinates for each shape by its corresponding *R*. To reduce shape variations that could arise due to rotational variations or mirror effects, each shape is aligned along its major axis length by applying a rotation matrix. Since cell and nuclear shapes are enclosed objects, each of the 50 points along the boundaries are iteratively assessed in both the clockwise and counterclockwise directions to ensure the most stable and comparable rotational confirmation among shapes[1] (**Figure 2B**).

Next, using the 50 points along the contours of each normalized shape as high dimensional features, principal component analysis (PCA) is then used to determine the eigenshape vectors (see **Glossary**). The eigenshape vector that accounts for 95% of the total variance is then used as a reduced set of descriptors for all cell and/or nuclear shapes[45–48] (**Figure 2C**). To empirically determine the representative shape modes for a given cell population, K-means clustering is performed using the reduced shape descriptors determined from the PCA[49] (**Figure 2D**). Among several classification methods tested, such as DBscan, OPTICS, Meanshift, and K-means, the K-means clustering algorithm was chosen for its fast calculation, robustness, and simplicity in setting the parameters.

Following this, each cell and/or nucleus is classified and binned into each cluster, which determines the distribution of shape modes per condition. To identify the representative shape for each shape mode for visualization purposes, the centroid locations of each cluster within the PCA-reduced features are then used to reconstruct the average morphology for each shape mode (**Figure 2E**). Lastly, using these representative shapes, together with the abundance of cells and/or nuclear within each shape mode, this analysis provides both a quantitative and visual handle for biological inferences on morphological data per condition. In addition, these shape mode distributions are used to compute the degree of morphological heterogeneity per condition based on the Shannon entropy (see **Glossary**).

In the previous study of pancreatic cancer cells[1] (see above), VAMPIRE analysis showed that metastasized cells present significantly lower heterogeneity than primary tumor cells based on the Shannon entropy. A lower heterogeneity was also found in a cohort of 10 breast cancer cell lines comparing metastatic to non-metastatic cancer cells[1]. Furthermore, deciphering the relative contributions to this heterogeneity, we identified

potential sources stemming from the cell cycle, cell-cell contacts, and heritable morphological variations (see **Glossary**).

In a separate study, the utility of the VAMPIRE analysis was further demonstrated by investigating how the morphologies of single-cell clones derived from a metastatic breast cancer cell line were associated with metastatic potential. We found that cell morphology is an emergent property of cancer cells, encoding information related to molecular determinants, and allowing the robust prediction of metastasis[6]. Lastly, we have used this approach to evaluate the morphological signature of healthy aging from skin dermal fibroblast cells[26]. We found that cellular age could be used to stratify individuals based on the cell morphology using a cohort of 32 samples of primary dermal fibroblasts collected from individuals between 2 and 96 years of age (see **Anticipated Results** and **Figure 7** for a subset of this re-analyzed data).

Since these previous studies[1,26], the core algorithms of VAMPIRE analysis remain unchanged. For this protocol, we have translated the original MATLAB code to Python, providing an open-source platform that is more amendable for distribution and implementation among various laboratories. In addition, we have optimized the performance and speed, and integrated the software into an easy-to-use graphic user interface (GUI), allowing users to input post-segmented images to generate a comprehensive panel of results that include plots, tables, and readouts of population heterogeneity.

Overview of the protocol

In this protocol, we established a python-based graphic user interface, "VAMPIRE GUI". We note that VAMPIRE GUI does not provide a segmentation tool; it analyzes cell and/or nuclear shapes that are already detected and segmented. The segmentation can be performed using software platforms such as ImageJ/Fiji[33] or CellProfiler[31], with easy integration of the segmentation results into VAMPIRE GUI. For simplicity, we have chosen to demonstrate how to perform VAMPIRE analysis using cell and nuclear segmentations generated using CellProfiler. However, please note the steps needed if other segmentation software platforms are used (see **Procedure**).

To help users explore the software and all its functionalities, we provide two small image datasets under "Example images" in **Supplementary Data** on GitHub repository

(https://github.com/kukionfr/VAMPIRE_open). See the directory of **Supplementary Data** in the **Supplementary Note** to locate example images and workflow in the **Overview of the protocol** section. Results from the VAMPIRE analysis using provided image datasets are also included in **Supplementary Note S2** and **Supplementary Data** under "Example output". Before applying VAMPIRE analysis to new image datasets, we recommend that users first perform VAMPIRE analysis using the image datasets provided and follow the detailed procedure provided in the **Procedure** section. We also illustrate the utility of VAMPIRE analysis by analyzing the morphology of mouse embryonic fibroblasts (MEFs) confined to adhesive micropatterns (akin to spatial restriction of cells in tissue) in the presence and absence of nuclear protein Lamin A/C, dermal fibroblasts derived from healthy individuals with increasing age, and for cells embedded within tissue sections (see **Anticipated results**).

The overall protocol is performed using four main steps: image segmentation from fluorescence/bright-field images of cells and nuclei, formatting segmentation data before importing into the VAMPIRE GUI, generating a VAMPIRE model from a training set of images, and applying the VAMPIRE model to the training set or a new image set. The procedure starts with the segmentation of fluorescence or bright-field images of cells to generate binary images of segmented cells (**Step 1**). This segmentation is executed outside of the VAMPIRE software using a segmentation tool of the user's choice.

To import segmented cells into VAMPIRE, the segmented images need to be organized in a designated format for use in the VAMPIRE GUI (**Step 2**). The segmented images must be grayscale images, with non-zero integer values representing the detected cell areas, and zero integer values for the background (non-cell areas). For instance, within an image, object 1 has pixel values of 1, object 2 has pixel values of 2, etc. This required format is a standard output in most segmentation software. Once segmented images are properly imported into the VAMPIRE GUI, it reads the images to obtain the coordinates along the curvilinear boundaries of the cell and/or nuclear contours. In addition, a few classic morphological parameters are computed for each object, including surface area, perimeter, major and minor axis length, circularity, and aspect ratio (see **Supplementary Table 1** for list of parameters generated).

Once the dataset to be analyzed by VAMPIRE is segmented and properly organized, the user decides the set of images to be used to train a VAMPIRE model by specifying the

image folder locations in a comma-separated values (CSV) file (**Step 3**). We refer to these specified images as "training set" hereafter. An example CSV file of this list, *"segmented image sets to build model.csv"*, can be found in **Supplementary Table 2**. The resulting VAMPIRE model built based on the specified training set will be saved within a designated local folder. (**Step 4-10**). Following this training step, the model can then be applied to either the same image set used to train the model or to a new image set by specifying the location in a new CSV file (**Step 11-13**). Ideally, users will apply the model to the same image set that was used to train. However, there are instances when it is appropriate to apply the VAMPIRE model to an entirely new dataset. For instance, a) if the datasets are unbalanced between experimental replicates or conditions, the user can balance the dataset by selecting a subset of datasets from certain replicates or condition in building the model; b) if the datasets grow to a point that it takes too long to build a new model with every run, a user can save time in building a new model by selecting a subset of datasets; c) if a user wants to validate the model or directly compare different conditions using the same shape modes. In so doing, the user can build a model on one experimental replicate, or similar cell types/conditions and apply it to another data set. Beyond these three examples, we intend to offer more flexible applications by allowing users to select specific datasets in building and applying the model.

It is important to note that these cases are only valid if the dataset used for training is expansive and similar enough to represent the newly acquired data, as it influences the appropriate classification of cells within each shape mode. To quantify this, users should use the 'distance from cluster center values, to determine how well cells were classified within each shape mode (see **Limitations**).

The output of the VAMPIRE model includes a plot showing the frequency distribution of each shape mode per condition, the CSV files that contain the shape mode for each cell and/or nuclei. (**Step 13**). Specifically, data for each cell includes the "xy" coordinates of cell centroids within the image, the area, circularity, aspect ratio, and assigned shape mode index (IDX), as well as the goodness of the shape mode classification for each cell that we refer to as "distance from cluster center" (see **Glossary**). This datasheet can be directly linked to the morphological features generated by CellProfiler, which makes VAMPIRE and CellProfiler analyses complementary in this regard. This seamless integration will allow users to further compare shape modes with other morphological features, and associate to other cell features

such as cell cycle state, protein expression, etc. Example datasheets showing the results from the analysis using both platforms are provided in **Supplementary Table 1 and 3**, labeled "*CellProfiler datasheet.csv*" and *"VAMPIRE datasheet.csv*".

Applications of VAMPIRE

We have previously demonstrated the utility of VAMPIRE in three key studies, (a) the morphological changes displayed by human pancreatic cancer cells as they spread from the primary tumor to the liver[1], and (b) the ability of single-cell morphologies to encode metastatic potential in breast cancer, and (c) the morphological changes of dermal fibroblasts derived from healthy individuals during ageing[26].

In general, VAMPIRE can be applied to any set of segmented images of cells or nuclei to detect and analyze changes in their morphology across multiple conditions and cell-culture systems. For instance, VAMPIRE can be applied to the study cell morphologies in response to a wide range of physiochemical changes, i.e. molecular characteristics[2,3,40,41] (e.g. cell cycle state, genetic and epigenetic status), microenvironmental and biomechanical perturbations[9,50,51] or disease states[1,6,26]. VAMPIRE analysis is also suitable for applications in phenotypic or drug screening [11,12,15]. Changes in cell morphology are often used in high-throughput biochemical discovery screens[52]; however, the large volume of data that is typically generated in such screens makes it difficult to visually inspect cell responses. Here, VAMPIRE provides users with the ability to rapidly classify phenotypically distinct cellular conditions in large amounts of data to identify drug-induced changes in the abundance and distributions of shape modes.

VAMPIRE analysis can also be applied to the cellular images derived beyond standard 2D cell culture models. We have recently demonstrated the utility of VAMPIRE analysis for cells embedded in 3D collagen matrices[1]. In that study, we obtained the 2D contours of cells from the z-projected images. VAMPIRE analysis showed that shape modes for cells in 3D cultures were distinctly more protrusive than the same cells in more traditional 2D cultures[1]. In addition to cell-culture systems in three-dimensional matrices, VAMPIRE analysis is applicable to study changes in cell and nuclear shapes in cells embedded within tissue sections (see **Figure 8**). A growing number of studies have shown that nuclear shape can encode prognostic information for patients in different types of cancers[53,54]. Segmented

nuclei within tissue sections can be imported directly into the VAMPIRE workflow, for instance, to assess changes in nuclear morphology that associate with tumor progression, drug responses, and patient outcomes.

Limitations of VAMPIRE

A key assumption of VAMPIRE analysis is that the shapes of segmented cells and nuclei faithfully represent the original cell and nuclear shapes. The accuracy of this segmentation, using for instance CellProfiler, relies on the user properly optimizing the image processing pipeline, choosing appropriate noise-reduction filters, and using suitable thresholding parameters. If the segmentation is not accurate, the shape modes generated using VAMPIRE will not be representative of the actual shapes of cells and nuclei. To address this potential issue, the user should evaluate the accuracy of segmentation before running VAMPIRE. This can be done via visual inspection by overlaying segmented cell contours onto the original image to gauge deviations. If the deviation between the segmented contours and the original images is substantial, the results from VAMPIRE analysis will not be reliable. Furthermore, VAMPIRE in its current version is designed to work on 2D projections of cells (x,y) and is not amendable to the analysis of 3D cells (x,y,z).

A challenge for any cell-morphological tool is the analysis and classification of highly complex cell shapes, such as cells with highly protrusive morphologies. Although VAMPIRE can compute a vast number of features from the coordinates of points along the shape boundaries to examine the complexity of cell shapes, the use of a reduced number of coordinates (i.e. 50 points) together with the dimensional reduction from the PCA can lead to shape modes with limited spatial resolution. In this case, users can either increase the number of coordinate points (which will also increase computing time) or use more suitable morphological analyses that directly quantify cell protrusions[9] or takes better account of cell protrusions[30]. Since users have the option to perform VAMPIRE analysis on cells and/or their corresponding nuclei to generate results for both, VAMPIRE analysis needs to be run separately on cell contours and nuclear contours. This allows users to specify different parameters (i.e. number of shape modes) to accurately describe both cell and nuclear shapes since cell shapes tend to be more complex than nuclear shapes.

To allow users to evaluate the goodness of the shape mode classification, we have provided the ability to gauge the distance between the computationally assigned shape modes and the actual cell shapes within the given data set. This metric is called "distance from cluster center (see **Glossary**). It is provided as part of the standard output data provided in **Supplementary Table 1**, "*VAMPIRE datasheet.csv*". If this distance is large, the VAMPIRE model has failed, and the model should be re-assessed. In addition, this depends on the parameters used in the VAMPIRE model, which can be improved by increasing the number of shape modes, or by eliminating 'outlier' cells, (see **Experimental design**).

Another limitation is that the shape modes determined by VAMPIRE are only as good as the dataset with which the model is trained. This means that to obtain the best results the training set should expansive enough and inclusive of cell types and conditions of interest. Because VAMPIRE uses a data-driven approach to identify dominant cell and/or nuclear shapes, rare shape populations may not be well classified, especially if the training data set is small. However, to gain insights into rare or less frequent shapes, the number of shape modes can be increased and optimized to suit. Lastly, if the new dataset (applying) includes a subpopulation of cell shapes that is non-existent in the dataset used to train, this would also result in misclassification of cells, and a higher fit error of cells classified within the pre-defined shape mode.

Comparison with other methods

Core algorithms of VAMPIRE include the PCA analysis on aligned outlines of cells that contains high-dimensional information to reduce it to lower-dimensional PCA components, retaining most of cell shape information and variation within a given dataset for further analysis. The eigenshape vectors from PCA have been readily used to quantitatively explore the new insights between cell morphology and physiology[45].

PCA is one of the data-driven dimensional reduction approaches. Other methods like Independent Component Analysis (ICA)[55] and more recently the Shape Component Analysis (SCA)[56] have been proposed to analyze cell shapes. As opposed to the data-driven approaches, deterministic approaches such as Fourier shape descriptors have also been used to decompose the shape outlines to represent the cell shapes with fewer shapes[57].

While the result of PCA is based on the variation of a given dataset, that of ICA is based on how much the generated components are independent to each other. Since cell shape data are often highly heterogenous, we chose to use PCA. To obtain subtypes of cells (shape mode), we applied the unsupervised clustering machine learning method to the reduced set of eigenshape vectors from PCA. Among several classification methods tested, such as DBscan, OPTICS, Meanshift, and K-means, the K-means clustering algorithm was chosen for its fast calculation, robustness, and simplicity in setting the parameters.

Similar to the outline-based shape analysis, the binary image of the shape can also be decomposed by PCA or ICA to vectors with lower-dimension. Also, Zernike polynomial, a deterministic method, has also been to decompose the binary image of shapes to represent the shape as components of a set of Zernike polynomials. It has been demonstrated that the outline-based analysis with PCA effectively represent and reconstruct the cell shapes with sufficient biologically relevant details using least number of shape vectors. Also, its computational process is relatively efficient in comparison to other methods such as Fourier and Zernike decompositions[46].

Recent advances in image modeling with neural networks made it feasible to derive the lower-dimensional representation of cell shapes while containing detailed cell shape. Unsupervised learning approach such autoencoder[58] and generative adversarial networks [59]has been recently extended to analyze the morphology of cells[60,61]. A recent study examined various autoencoder algorithms performance in the representation of 2D cell shapes, and their results show the outline-based PCA in generally perform similarly in lower dimensions and better in higher dimensions representation of cell than outline-based autoencoder shape analysis. Though Image-based autoencoder shape representation can slightly outperform the outline-based PCA in terms of accuracy, the outline PCA is computationally cheaper and faster than autoencoder. Given the field of deep learning evolve rapidly, there could be a strong potential that is significantly better approach to representing cell shapes can be derived from neural network methods.

Directly characterizing the cell shapes with a set of human-interpretable descriptive features has been the most commonly used approach to study the morphology of cells. Different types of shape features have been proposed from the simple ones such as shape factors, curvature and roughness of the cell and nuclear contours[14,30,42] to the more complex

features such as scale invariance feature transform (SIFT)[62] and speed up robust feature (SURF)[63]. This type of analysis is based on discriminative methods that try to capture just enough information about shapes to be able to distinguish the investigating biological states [17]. Two commonly used tools for these type of cell shape analysis are the CellProfiler [31]and MorpholibJ[64], a plugin for ImageJ[33]. These tools extract an extensive list of features, such as shape factor, eccentricity, and Zernike number. As mentioned previously, too many shape descriptors can limit the ease of biological interpretation and visualization of morphological data, especially with abstract features. For example, the CellProfiler provides a set of ~1500 morphological features to describe the morphology of cells, including parameters that describe size, shape, intensity, and texture[14]. While the pure magnitude of the features assessed (~1500) increases the likelihood of identifying differences among cell populations, this expanse of parameters could limit biological interpretation if the features are abstract (e.g. Zernike number, angular 2nd moment). In sum, each user should decide the appropriate software solution for their morphology quantification based on the questions at hand.

Experimental design - Selection of parameters for VAMPIRE analysis

Within the VAMPIRE interface, a key input parameter for establishing the model is the number of shape modes. We encourage the user to tune this parameter to obtain optimal results. Here, we briefly present the underlying basis for the selection of the number of shape modes. During the dimensional reduction steps, we implement K-means clustering to relate individual cell to the centroid of each cluster (shape mode), where the distance from the cluster centroid is stored as the "distance from centroid" (see **Glossary**). This K-means clustering stratifies cells on the principle of minimizing a parameter known as the inertia. This inertia is calculated as the sum of the squared distance between the cluster centroid and each data point within the cluster (**Figure 4A**). Inertia can be thought of as the metric that defines how internally coherent clusters are, with the optimal inertia value being zero.

Fundamentally, increasing the number of clusters reduce the inertia and improves cluster coherence. To illustrate the effect of the number of clusters on the inertia, we plotted the number of clusters as a function of the inertia for cells cultured on adhesive micropatterns (**Figure 4B**). We observed an elbow-shaped decay function, at which point there was only a minimal benefit to increasing the number of clusters

Control experiments

Examining cells of pre-defined shapes is the most straightforward way to validate VAMPIRE analysis. Using adhesive micropatterning techniques, the user evaluates the morphologies of cells confined to pre-defined adhesive shapes (see **Anticipated results**). As a result, cells cultured on circular and triangular adhesive micropatterns should exhibit shape modes that are predominantly circular and triangular, respectively.

Materials

Equipment

- A computer with at least 8GB of RAM running Microsoft Windows 10 (64 bit)

Software

- VAMPIRE executable software
- CSV editor (e.g. Microsoft Excel, Numbers)
- Choice of a standard segmentation tool:
    - CellProfiler 3.1.9 software (https://cellprofiler.org/releases/)
    - ImageJ/FIJI (https://imagej.net/Fiji/Downloads)
    - MATLAB (https://www.mathworks.com/downloads)

Procedure

Segment images of cells or nuclei ● TIMING 10-60 min

1| Segment the fluorescence/brightfield images to identify the boundaries of cells and/or nuclei. The VAMPIRE GUI does not segment cells. Users should accomplish this task with software, including, but not limited to, CellProfiler, ImageJ, or MATLAB. More information on how to use these segmentation tools can be found on their official websites

- ImageJ: https://imagej.net/Segmentation;
- MATLAB: https://www.mathworks.com/help/images/detecting-a-cell-using-image-segmentation.html;
- CellProfiler: https://cellprofiler.org/tutorials.

To demonstrate the VAMPIRE analysis procedure, we provide sample images of fluorescently tagged cells in the **Supplementary Data** under the "Example images" folder and its corresponding results through VAMPIRE analysis procedure. Two sample sets, MEF_LMNA-- and MEF_wildtype, are provided and correspond to mouse embryonic fibroblast cells having wildtype expression of Lamin A or Lamin A knockout, respectively (**Figure 5A**). Throughout this **Procedure**, refer to the directory of supplementary data in **Supplementary Note S1** to locate example data and results. We have provided segmented example images using CellProfiler (**Figure 5A**), as well as a sample CellProfiler segmentation pipeline in **Supplementary Data**. Note that the example workflow is designed using CellProfiler version 3.1.9, and it may not work with more recent versions of CellProfiler.

? TROUBLESHOOTING

2| Convert the segmented image data to the required format that is compatible with VAMPIRE analysis, if needed. To prepare images for VAMPIRE analysis, images should be stared as binary tiff files, where the area of each cell must have a non-zero integer value. Segmented images for the same condition or those having multiple fluorescence channels should be placed in the same folder. To properly store images, the segmented images must have filenames that distinguish objects by channel (i.e. *xy001c1.tif* and *xy001c2.tif*). A sample format of segmented images is provided in the **Supplementary Data** for reference.

? TROUBLESHOOTING

Build shape-analysis VAMPIRE model ● TIMING 3-10 min

3| Generate a CSV file to specify the location of the segmented image sets for use in constructing a VAMPIRE model. In this CSV file, the first row contains column headings specifying the information to be entered. Each column specifies information about the specific segmented images. From the second row, each column should be filled with information of a specific segmented image set with the following order:

i. "set ID": row index number. "set ID" and "condition name" will be part of the VAMPIRE output filename (i.e. *Shape mode distribution_1_wildtype.png*).

ii. "condition name": description of an image set.

iii. "set location": the location/path of the folder containing segmented images

iv. "tag": a string of text. Only segmented images in the set location with filenames containing the tag will be identified and analyzed. For example, if "tag" is set as "c1", for an image set location containing segmented images from multiple channels (i.e. *xy001c1.tif, xy001c2.tif, xy002c1.tif, xy002c2.tif*) only image filenames containing "c1" (i.e. *xy001c1.tif* and *xy002c1.tif*) will be analyzed.

v. "note": any information about the image sets needed for the user's record. This information is not used in the VAMPIRE analysis.

An example CSV file named *"Segmented image sets to build model.csv"* can be found in **Supplementary Table 2**. Users can download and directly modify the example CSV files using Excel or other CSV editors. To use the example segmented images provided in the **Supplementary Data** for the following analysis, the user may need to update the set location column in the example CSV file with the actual location of the example segmented images.

4| Download VAMPIRE stand-alone software named "vampire.exe" from GitHub (https://github.com/kukionfr/VAMPIRE_open/releases). Launch VAMPIRE Graphic User Interface (GUI) by opening the VAMPIRE.exe file.

Note: the current version of VAMPIRE GUI is only available for Windows 10 users. Source codes are available on the GitHub and PyPI (https://pypi.org/project/vampireanalysis). Python users can download the source code using pip installer with the following command: "pip install vampireanalysis". These repositories will be continually updated and maintained.

5| Locate the CSV file generated in **step 3** to build VAMPIRE model in the "Build Model" section of the VAMPIRE GUI. Click "Load CSV". This will open a popup window for the user to select the CSV file.

6| Specify the number of coordinates to extract from the cell contours in Build Model section of VAMPIRE GUI under the "number of coordinates" box. The default value is fifty. A higher number of coordinates will better represent the object boundary at the expense of analysis speed. A lower number of coordinates may not capture the details of the object boundary and the result of analysis may under-represent the actual cell morphology.

7| Determine the number of shape modes in the "Build Model" section of the VAMPIRE GUI under the "number of shape modes" box. The default value is ten. To optimize this number, refer to the ***Selection of parameters for VAMPIRE analysis*** section.

8| Specify where the output model should be saved. This information can be entered in the "Build Model" section of VAMPIRE GUI under the "Model output folder" box.

9| Name the model in the "Build Model" section of VAMPIRE GUI under the "Model name" box. This name will be used to generate a pickle file that contains model parameters.

10| Click "Build Model" in VAMPIRE GUI to generate a VAMPIRE model based on the specified parameter values provided in **steps 6 and 7**. Once the model is generated, it will be saved in the output folder specified in **step 8**. Within this new folder, VAMPIRE model data will be saved into a subfolder "[*model name*]" that contains:

- A VAMPIRE model file that is named "[*model name*].pickle".
- A subfolder named "[*model name*] figures" that contains:
  - The overlay of 20 randomly selected raw shapes classified into each shape mode named *"registered objects.png".*
  - The dendrogram showing the level of correlation between shape modes named *"shape mode dendrogram.png".*

Example output files of this step are provided in the **Supplementary Data**, under "Example output". These files are generated from the example segmented images provided in **step 2,** using the default values of parameters from **steps 6 and 7**.

? TROUBLESHOOTING

Analyze cell shapes with VAMPIRE model ● TIMING 1-10 min

11| Repeat **step 3** to specify the sets of segmented images to apply the VAMPIRE model. If you need to prepare new sets of segmented images, repeat **steps 1 and 2**. The format of the CSV file remains the same. Once the user generates the CSV file, go back to the VAMPIRE GUI. In the "Apply Model" section of the VAMPIRE GUI, click "load CSV". This will open a popup window for the user to select the CSV file.

12| Specify the previously built model to analyze the segmented images. Click the "load model" button to choose the pickle file generated in **step 10**. Refer to **Supplementary Note S1** to locate the pickle file.

13| Perform the VAMPIRE analysis on the specified images by clicking the "Apply Model" in VAMPIRE GUI. When this process is finished, a new folder will be created named "Result

based on [*model name*]" within the VAMPIRE model folder. This new folder contains a collection of distributions showing the fractional abundance for cells within each shape mode, with the percent of cells within each shape mode denoted on the top of each bar (**Figure 5B, 6B, 7A**). Each distribution is saved with the naming convention: *"Shape mode distribution_[condition].png".* Clicking the "Apply Model" button also generates a VAMPIRE datasheet CSV file in each segmented image set folder. Each datasheet CSV contains:

- Filename: name of the segmented image file that contains the object
- ImageID: ID number of the segmented image file
- ObjectID: ID number of the object within the segmented image file
- X and Y: location of the object's center of mass within the segmented image
- Area: area of the object
- Perimeter: length of object's circumference
- Lengths of the major axis and minor axis
- Circularity: shape factor calculated by $\frac{4\pi A}{P^2}$. Its value varies from 0 to 1. The circularity of a perfect circle is 1.
- Aspect ratio: it is calculated by major axis length divided by minor axis length.
- Shape mode ID number: a number that represents the shape mode where each cell belongs to.
- Distance from cluster center: a metric to determine the goodness of the classification into shape modes defined as the distance between the cluster centroid and the selected object centroid.

Example output files for this step are provided in the **Supplementary Data**, under "Example output". These files are generated using the VAMPIRE model provided in **Supplementary Data** under the same folder "Example output". See the directory of **Supplementary Data** in **Supplementary Note S1** to locate the output files. A compiled example of shape parameters of the VAMPIRE datasheet is shown in the **Anticipated Results** (**Figure 6B**).

Troubleshooting

Troubleshooting guidance can be found in Table 2.

TABLE 2| Troubleshooting table.

Step	Problem	Possible reason	Solution
1	Cannot run pipeline: the pipeline did not identify any image sets	The user did not load any images in the "Images" module.	Drag and drop images into the "Images" module of CellProfiler.
1	Subfolder under CellProfiler output folder is named "None"	The metadata extraction rule is incorrect	Modify the extraction rule under the "Metadata" module in CellProfiler.
4	The MATPLOTLIBDATA environment variable was deprecated in Matplotlib 3.1 and will be removed in 3.3.	The executable file of VAMPIRE is created using a software that uses a variable that will be removed in the future.	Ignore this message since VAMPIRE is not affected by this warning.
10	IndexError: arrays used as indices must be an integer	Segmented images do not contain any cell or nucleus	Check if segmented images have a correct format as specified in step 2 and have at least one cell or nucleus.
10	RuntimeWarning: Mean of empty slice	The number of objects is less than the number of clusters	Provide images with a greater number of cells than the number of clusters.
10	Permission denied	CSV file is open while the analysis is running	Close all CSV files open and repeat step 10.

Timing

The timing information below is estimated based on the analysis of 10,000 cells using an i7-8700k Intel CPU with 5.0 GHz clock speed on Windows 10 pro OS. This time corresponds to the time it takes an experienced VAMPIRE user to perform analysis. More time may be required when setting up or using VAMPIRE for the first time.

Step 1-2, Segment images of cells or nuclei, 10-60 mins
Steps 3-10, Build shape-analysis VAMPIRE model, 3-10 mins
Steps 11-13, Analyze cell shapes with VAMPIRE model, 1-10 mins
Total, steps 1-13, complete VAMPIRE analysis, 14-80 mins

BOX 1 | GLOSSARY

Eigenshape vectors—Mathematical descriptors used to describe cell shapes based on the principal component analysis (PCA) of cellular shape features. Once determined, a linear combination of Eigenshape vectors are used to reconstruct the original shape of each cell.

Shape modes—mathematical descriptors of cell and nuclear shapes based on clustering analysis of user-generated eigenshape vectors. Once these shape modes are identified, the abundance of cells within each shape mode is assessed and the entropy to determine the extent of heterogeneity can be computed.

Shannon entropy—a mathematical description used to quantify the degree of diversity within a population of cells based on the number of shape modes and the abundance of cells within each shape mode. It is given by the general equation:

$$S = -\sum p_i ln(p_i)$$

S is the Shannon entropy and p_i is the occurrence of cells in each shape mode.

Cellular heterogeneity—a property that describes the extent of cell-to-cell variations within a cell population.

Eccentricity—a measure of how similar a cell shape is to a circle or an ellipse, calculated as the ratio of the distance between the foci of the ellipse and its major axis length.

Solidity—Ratio of cell area to convex hull area of the cell, (convex hull area is the area of the smallest convex polygon that encloses the region).

Curvature—defined as the degree of deviation from a straight line. It is often calculated

based on the change in radius of the smallest circle rolling along the boundary of the enclosed shape.

Roughness—a measure calculated based on the change in the length of a vector that is centered at the geometric centroid of an enclosed object as it rotates along with each boundary point.

Area—the number of pixels of a given size that can fit within the enclosed region, since the size of each pixel is known, the area of cells and/or nuclei can be converted into various scales, including square microns (μm^2).

Distance from cluster center— A distance matrix is computed from pairwise euclidean distance between the 50 equidistant points along the contour of each cell and those comprising shape mode. The smallest value in this matrix denoted the 'distance from cluster center'.

Principal component analysis— Abbreviated as PCA, is a mathematical technique for reducing the dimensionality of large datasets, increasing interpretability but at the same time minimizing information loss by finding new uncorrelated variables, principal components, from possibly correlated variables.

Heritable morphological variations—cell-to-cell variations that are persistent along with many cell generations, or is propagated along cell divisions.

Anticipated results

To demonstrate the utility of VAMPIRE, we examined the shapes of mouse embryonic fibroblasts (MEFs) in response to different surface topographies. These cells are either wild type (MEF LMNA +/+) or deficient in lamin A/C (MEF LMNA-/-). Cells were seeded onto three different two-dimensional substrates: 1. circular or 2. triangular shaped fibronectin coated islands, surrounded by poly-ethylene glycol (PEG) passivated regions, and 3. Uniform fibronectin coated surfaces. Cells were incubated overnight on each substrate then fixed and stained with DAPI and Alexa Fluor 488 Phalloidin, highlighting nuclear DNA and F-actin fibers respectively. Cells and their corresponding nuclei were segmented using CellProfiler, then the contours were analyzed using VAMPIRE with 10 shape modes and 50 contour points (**Figure 6A**).

We quantified the shape mode distribution for each of the probed conditions and examined whether cells on patterns exhibited associations with particular shape modes that resembled circles and triangles (**Figure 6B**). As expected, results showed that both LMNA +/+ and LMNA -/- cells seeded on un-patterned surfaces exhibited mixed shape profile i.e., similar abundance in all identified cellular shape modes, as opposed to the cells seeded on the patterned substrates. Cells seeded on circular patterns exhibited enrichment in the circular shape mode (mode 4) with an average abundance of 55% and 52% of the total cell populations for LMNA +/+ AND LMNA -/-, compared to 8.1% and 21% of those seeded on an unpatterned substrate. Cells seeded on triangular patterns were primarily classified into two triangular shape modes, the "sharp" (mode 1) and "blunted" vertex (mode 2) triangles, with decreased abundance in the remaining shape modes (mode 6-9) (**Figure 6B**).

Interestingly, LMNA -/- cells seeded on triangular patterns were classified as "blunt" (mode 2) three times more (abundance of 34%) than "shape" (mode 1) (abundance of 12%). We did not observe such a difference between the two shape modes in LMNA +/+ cells. This bias suggests that the deficiency in lamin A/C limits the ability of these cells to form acute angle vertices, potentially through defective nucleo-cytoskeletal connections[51,65]. Our results reveal that cells can respond morphologically differently to the same shape constrains and VAMPIRE analysis can visualize and quantify the subtle differences.

We computed the Shannon entropy for the cell populations and observed no significant differences between LMNA +/+ and LMNA -/- within the same micropattern (**Figure 6B**). However, looking across conditions, we observe a significant decrease in the population heterogeneity for both LMNA +/+ and LMNA -/- seeded on circular patterns, relative to cells seeded on unpatterned surfaces and triangular patterns. The aspect ratio of LMNA +/+ cells increased from 1.66 (no pattern) to 2.20 (triangle pattern), suggesting a more elongated shape for these cells. However, evaluating the shape factor in the same cells showed an increase from 0.34 (no pattern) to 0.51 (triangle pattern), suggesting rounder cell shapes on circular patterns. These seemingly contradictory results measured by shape factor and aspect ratio suggests that VAMPIRE analysis can provide direct visual insight to better monitor the transition of cell morphology than classical morphology parameters.

We also examined the association between cellular morphology and chronological ages of dermal fibroblasts derived from seven healthy individuals using VAMPIRE analysis[26].

While the morphology of mouse embryonic fibroblast was emphasized by artificial micropatterns, this example illustrates the sensitivity of VAMPIRE's to classify subtle, biologically meaningful morphology changes. Previously, we demonstrated that cell and nuclear morphologies of dermal fibroblasts encode key information about the biological age for healthy individuals[26]. Using ten shape modes, the VAMPIRE analysis shows a decrease in the frequency of cells having rounded morphologies shape modes, and an increase in cells having irregular nuclear morphologies with increasing age. This is measured by a negative age-correlations for shape modes 1 and 2 having rounded shapes, and positive age-correlations for irregular nuclear shape modes 3, 4, and 7 (**Figure 7A**). Correlation coefficients denote Pearson's correlation. We also note that computing standard shape parameters, including shape factor and aspect ratio, yielded very similar values for the cells in different shape modes, (SF: 0.77-0.83, and AR: 1.51-1.64), even for shape modes having opposite trends in age correlations (R: -0.6 and +0.6)—i.e. shape modes 1 and 3. Furthermore, circular shape modes 1 and 2 have very similar shape parameters (SF and AR) to ellipsoidal shape modes 9 and 10 (**Figure 7B**). Again, this demonstrates the utility of VAMPIRE analysis to visually and quantitatively identifying morphological changes that would otherwise go unnoticed using traditional morphological parameters.

Lastly, boasting the utility of VAMPIRE analysis beyond cultured cells, we have successfully implemented VAMPIRE analysis to analysis of tissue sections. Here we compare the morphologies of cells derived from the human epidermis and reticular dermis based on hematoxylin and eosin (H&E) stained tissue sections (**Figure 8A**). Note that we segmented nuclei within the tissue sections using a custom image analysis algorithm. To compare the morphology of cells in the epidermis and reticular dermis region, we built VAMPIRE model using nuclei segmented from the scanned image of an H&E stained skin tissue biopsy from a 79 years old donor. We observed that shape modes 1 through 3 were more elongated (i.e. less circular) relative to modes 4 through 10 (**Figure 8B**). As expected, VAMPIRE analysis was able to decipher differences between the two regions of the tissue section, with nearly 50% of dermal cells being classified as modes 1 through 3, as compared to only 6.4% for epidermal cells (**Figure 8B**).

Data availability

The datasets generated during and/or analyzed during the current study are available from an example dataset is deposited on GitHub: https://github.com/kukionfr/VAMPIRE_open/tree/master/Supplementary%20Data

Code availability

The source code is available on GitHub : https://github.com/kukionfr/VAMPIRE_open . The code can be accessed and used by readers without restriction.

References

1. Wu, P.-H. *et al.* Evolution of cellular morpho-phenotypes in cancer metastasis. *Sci. Rep.* **5**, 18437 (2016).
2. Chen, W.-C. *et al.* Functional interplay between the cell cycle and cell phenotypes. *Integr. Biol. (Camb).* **5**, 523–34 (2013).
3. Chambliss, A. B., Wu, P. H., Chen, W. C., Sun, S. X. & Wirtz, D. Simultaneously defining cell phenotypes, cell cycle, and chromatin modifications at single-cell resolution. *FASEB J.* **27**, 2667–2676 (2013).
4. Bakal, C., Aach, J., Church, G. & Perrimon, N. Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science (80-.).* **316**, 1753–1756 (2007).
5. Rohban, M. H. *et al.* Systematic morphological profiling of human gene and allele function via cell painting. *Elife* **6**, (2017).
6. Wu, P.-H. *et al.* Single-cell morphology encodes metastatic potential. *Sci. Adv.* (2020) doi:10.1126/sciadv.aaw6938.
7. Driscoll, M. K. *et al.* Robust and automated detection of subcellular morphological motifs in 3D microscopy images. *Nat. Methods* (2019) doi:10.1038/s41592-019-0539-z.
8. Yeung, T. *et al.* Effects of substrate stiffness on cell morphology, cytoskeletal structure, and adhesion. *Cell Motil. Cytoskeleton* (2005) doi:10.1002/cm.20041.
9. Guo, Q. *et al.* Modulation of keratocyte phenotype by collagen fibril nanoarchitecture in membranes for corneal repair. *Biomaterials* **34**, 9365–9372 (2013).

10. Sero, J. E. *et al.* Cell shape and the microenvironment regulate nuclear translocation of NF-κB in breast epithelial and tumor cells. *Mol. Syst. Biol.* **11**, 790 (2015).
11. Simm, J. *et al.* Repurposed High-Throughput Images Enable Biological Activity Prediction For Drug Discovery. *doi.org* 108399 (2017) doi:10.1101/108399.
12. Bray, M.-A. *et al.* A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *Gigascience* **6**, 1–5 (2017).
13. Wawer, M. J. *et al.* Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proc. Natl. Acad. Sci.* **111**, 10911–10916 (2014).
14. Bray, M. A. *et al.* Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774 (2016).
15. Beghin, A. *et al.* Localization-based super-resolution imaging meets high-content screening. *Nat. Methods* **14**, 1184–1190 (2017).
16. Meijering, E., Carpenter, A. E., Peng, H., Hamprecht, F. A. & Olivo-Marin, J. C. Imagining the future of bioimage analysis. *Nat. Biotechnol.* **34**, 1250–1255 (2016).
17. Ruan, X. & Murphy, R. F. Evaluation of methods for generative modeling of cell and nuclear shape. *Bioinformatics* (2019) doi:10.1093/bioinformatics/bty983.
18. Piccinini, F. *et al.* Advanced Cell Classifier: User-Friendly Machine-Learning-Based Software for Discovering Phenotypes in High-Content Imaging Data. *Cell Syst.* **4**, 651-655.e5 (2017).
19. Danuser, G. Computer vision in cell biology. *Cell* vol. 147 973–978 (2011).
20. Falk, T. *et al.* U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* **16**, 67–70 (2019).
21. Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Mining* (2017) doi:10.1186/s13040-017-0155-3.
22. Gabril, M. Y. & Yousef, G. M. Informatics for practicing anatomical pathologists: Marking a new era in pathology practice. *Modern Pathology* vol. 23 349–358 (2010).
23. Fuchs, T. J. & Buhmann, J. M. Computational pathology: Challenges and promises for tissue analysis. *Computerized Medical Imaging and Graphics* vol. 35 515–530 (2011).
24. Sarnecki, J. S. *et al.* A robust nonlinear tissue-component discrimination method for computational pathology. *Lab. Investig.* **96**, 450–458 (2016).

25. Beck, A. H. *et al.* Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. *Sci. Transl. Med.* **3**, 108ra113-108ra113 (2011).
26. Phillip, J. M. *et al.* Biophysical and biomolecular determination of cellular age in humans. *Nat. Biomed. Eng.* **1**, 0093 (2017).
27. Pegoraro, G. & Misteli, T. High-Throughput Imaging for the Discovery of Cellular Mechanisms of Disease. *Trends in Genetics* vol. 33 604–615 (2017).
28. Lang, P., Yeow, K., Nichols, A. & Scheer, A. Cellular imaging in drug discovery. *Nature Reviews Drug Discovery* vol. 5 343–356 (2006).
29. Loo, L. H., Wu, L. F. & Altschuler, S. J. Image-based multivariate profiling of drug responses from single cells. *Nat. Methods* **4**, 445–453 (2007).
30. Sailem, H. Z., Sero, J. E. & Bakal, C. Visualizing cellular imaging data using PhenoPlot. *Nat Commun* **6**, 1–6 (2015).
31. McQuin, C. *et al.* CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.* (2018) doi:10.1371/journal.pbio.2005970.
32. Carpenter, A. E. *et al.* CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* (2006) doi:10.1186/gb-2006-7-10-r100.
33. Schindelin, J. *et al.* Fiji: An open-source platform for biological-image analysis. *Nature Methods* (2012) doi:10.1038/nmeth.2019.
34. Jayatilaka, H. *et al.* EB1 and cytoplasmic dynein mediate protrusion dynamics for efficient 3-dimensional cell migration. *FASEB J.* (2018) doi:10.1096/fj.201700444RR.
35. Giri, A. *et al.* The Arp2/3 complex mediates multigeneration dendritic protrusions for efficient 3-dimensional cancer cell migration. *FASEB J.* **27**, 4089–4099 (2013).
36. Fraley, S. I. *et al.* A distinctive role for focal adhesion proteins in three-dimensional cell motility. *Nat. Cell Biol.* (2010) doi:10.1038/ncb2062.
37. Jayatilaka, H. *et al.* Synergistic IL-6 and IL-8 paracrine signalling pathway infers a strategy to inhibit tumour cell migration. *Nat. Commun.* **8**, (2017).
38. Jayatilaka, H. *et al.* Tumor cell density regulates matrix metalloproteinases for enhanced migration. *Oncotarget* **9**, 32556–32569 (2018).
39. Phillip, J. M., Aifuwa, I., Walston, J. & Wirtz, D. The Mechanobiology of Aging. *Annu. Rev. Biomed. Eng.* **17**, 113–141 (2015).
40. Kim, D.-H. *et al.* Volume regulation and shape bifurcation in the cell nucleus. *J. Cell Sci.*

129, 457–457 (2016).

41. Yu, Y. *et al.* Inhibition of Spleen Tyrosine Kinase Potentiates Paclitaxel-Induced Cytotoxicity in Ovarian Cancer Cells by Stabilizing Microtubules. *Cancer Cell* **28**, 82–96 (2015).
42. Driscoll, M. K. *et al.* Automated image analysis of nuclear shape: What can we learn from a prematurely aged cell? *Aging (Albany. NY).* **4**, 119–132 (2012).
43. Bookstein, F. L. Landmark methods for forms without landmarks: Morphometrics of group differences in outline shape. *Med. Image Anal.* (1997) doi:10.1016/S1361-8415(97)85012-8.
44. Dryden, I. L. & Mardia, K. V. *Statistical shape analysis, with applications in R: Second edition*. *Statistical Shape Analysis, with Applications in R: Second Edition* (2016). doi:10.1002/9781119072492.
45. Keren, K. *et al.* Mechanism of shape determination in motile cells. *Nature* (2008) doi:10.1038/nature06952.
46. Pincus, Z. & Theriot, J. A. Comparison of quantitative methods for cell-shape analysis. *J. Microsc.* **227**, 140–156 (2007).
47. MacLeod, N. Generalizing and extending the eigenshape method of shape space visualization and analysis. *Paleobiology* (1999) doi:10.1666/0094-8373-25.1.107.
48. Tsai, A. *et al.* A shape-based approach to the segmentation of medical imagery using level sets. *IEEE Trans. Med. Imaging* (2003) doi:10.1109/TMI.2002.808355.
49. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* (2011).
50. Kim, D. H. & Wirtz, D. Focal adhesion size uniquely predicts cell migration. *FASEB J.* **27**, 1351–1361 (2013).
51. Kim, J.-K. *et al.* Nuclear lamin A/C harnesses the perinuclear apical actin cables to protect nuclear morphology. *Nat. Commun.* **8**, 2123 (2017).
52. Zheng, W., Thorne, N. & McKew, J. C. Phenotypic screens as a renewed approach for drug discovery. *Drug Discovery Today* (2013) doi:10.1016/j.drudis.2013.07.001.
53. Kashyap, A., Jain, M., Shukla, S. & Andley, M. Role of nuclear morphometry in breast cancer and its correlation with cytomorphological grading of breast cancer: A study of 64 cases. *J. Cytol.* (2018) doi:10.4103/JOC.JOC_237_16.

54. Seethala, R. R. *et al.* Noninvasive follicular thyroid neoplasm with papillary-like nuclear features: A review for pathologists. *Modern Pathology* (2018) doi:10.1038/modpathol.2017.130.
55. Lee, T.-W. & Lee, T.-W. Independent Component Analysis. in *Independent Component Analysis* 27–66 (Springer US, 1998). doi:10.1007/978-1-4757-2851-4_2.
56. Lee, H. C., Liao, T., Zhang, Y. J. & Yang, G. Shape component analysis: Structure-preserving dimension reduction on biological shape spaces. *Bioinformatics* (2016) doi:10.1093/bioinformatics/btv648.
57. Burger, W. & Burge, M. J. Fourier Shape Descriptors. in 169–227 (Springer, London, 2013). doi:10.1007/978-1-84882-919-0_6.
58. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science (80-.).* (2006) doi:10.1126/science.1127647.
59. Goodfellow, I. J. *et al. Generative Adversarial Nets*. http://www.github.com/goodfeli/adversarial.
60. Osokin, A., Chessel, A., Salas, R. E. C. & Vaggi, F. GANs for Biological Image Synthesis. *Proc. IEEE Int. Conf. Comput. Vis.* **2017-October**, 2252–2261 (2017).
61. Johnson, G. R., Donovan-Maiye, R. M. & Maleckar, M. M. Generative Modeling with Conditional Autoencoders: Building an Integrated Cell. (2017).
62. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**, 91–110 (2004).
63. Bay, H., Ess, A., Tuytelaars, T. & Van Gool, L. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* (2008) doi:10.1016/j.cviu.2007.09.014.
64. Legland, D., Arganda-Carreras, I. & Andrey, P. MorphoLibJ: Integrated library and plugins for mathematical morphology with ImageJ. *Bioinformatics* (2016) doi:10.1093/bioinformatics/btw413.
65. Kim, D. H. & Wirtz, D. Cytoskeletal tension induces the polarized architecture of the nucleus. *Biomaterials* **48**, 161–172 (2015).

Acknowledgement

This work was supported in part by the National Institutes of Health Grants U54CA143868 (DW), R01CA174388 (DW), and U01AG060903 (DW, JMP, PHW)

Author Contributions

JMP and PHW designed and conducted experiments; PHW, JMP, DW and WC conceived analysis and workflow of VAMPIRE; PHW developed the original VAMPIRE software; KSH converted the VAMPIRE software from MATLAB to Python; KSH developed the graphical user interface of VAMPIRE; KSH and JMP analyzed and plotted data; PHW and DW supervised the study; JMP, DW, KSH and PHW wrote and edited the protocol; DW, JMP, and PHW secured funding.

Competing interests

The authors declare no competing interests.

FIGURE CAPTIONS

Figure 1. Cells confined to narrow ranges of traditional morphological parameters still exhibit highly variable shapes. Scatter plot showing the distributions of 37,750 mouse embryonic fibroblast cells confined to a three-dimensional axis of aspect ratio, shape factor, and solidity. The subset of 10 cells highlighted in red display substantial morphological heterogeneity, despite highly similar values of aspect ratio, circularity, and solidity.

Figure 2. Overview of VAMPIRE analysis, from the extraction of contour coordinates to the automatic generation of shape modes A. The contour of a single cell described by 50 equidistant points along its contour. **B.** Unaligned (left) shapes of a set of cells are pooled, normalized by size, and aligned (right). **C**. Eigenshape vectors (i.e., principal components or PCs) are obtained from a principal component analysis (PCA) of the contour coordinates of aligned cells. **D**. Reconstructed cell shape from a reduced number of eigenshape vectors. The reduced number of eigenshape vectors was defaulted at the number of vectors that comprise 95% of the shape variations among all assessed cells. **E**. Representative cellular shape modes are obtained by applying a K-means clustering method to a set of cell morphology data described by the reduced number of Eigenshape vectors.

Figure 3. Overview of VAMPIRE implementation with the VAMPIRE GUI. A. The VAMPIRE Graphic User Interface (GUI). **B**. Flow diagram illustrating key steps in the implementation of VAMPIRE analysis with VAMPIRE GUI. Images of cells are first segmented into binary images that highlight the cellular region and/or nuclear region. In VAMPIRE GUI top section (highlighted in red) allows users to specify analysis parameters and the location of segmented images to be used to create a VAMPIRE analysis model. Once the VAMPIRE analysis model is established, the user can specify the sets of segmented images to be analyzed using the previously established model (highlighted in blue).

Figure 4. **Determinants of cluster coherence in the shape mode distributions. A.** Schematic illustrating the concept of inertia in K-means clustering. The inertia is measured by total squared distances of all data points to the centroids of their corresponding subtype. The lower the inertia value indicates better segregation of clusters indicating more inter-cluster coherence. **B**. The inertia in principle decays with an increasing number of clusters. The corresponding cluster number at the elbow point where the inertia decay rate starts to drop is the suggested cluster number to use in VAMPIRE for K-means clustering. The example inertia profile is calculated based on 17,093 MEF cells. The error bars indicate the standard deviation from five separate runs using VAMPIRE analysis.

Figure 5. **VAMPIRE analysis of *LMNA*+/+ and *LMNA*-/- mouse embryonic fibroblasts. A.** Images of phalloidin-stained (top) wild-type (LMNA+/+, left) and lamin-deficient (LMNA-/-, right) mouse embryonic fibroblasts. Segmentation is obtained using CellProfiler. B. Bar plots showing the distribution of cell shape modes from the VAMPIRE analysis of the MEFs. Numbers above the bars represent the abundances [%] of cells in each shape mode.

Figure 6. VAMPIRE analysis of mouse embryonic fibroblasts seeded on adhesive micro-patterned surfaces. A. Fluorescence microscopy images of wild-type (LMNA+/+) and lamin-deficient (LMNA-/-) mouse embryonic fibroblasts cultured on circular (top row) and triangular (middle row) adhesive fibronectin-coated micropatterns. Control cells (bottom row) are placed on the fibronectin-coated glass. Cells were fixed and stained for F-actin (red) and nuclear DNA (blue). Segmented fluorescence images (right). On the left are

the raw images of cells and their nuclei with the segmented contours highlighted in yellow; on the left are the same cells color-coded according to the shape mode to which they belong. Inserts are magnified views of cells. The identified shape modes are located on the right of the panel. B. The table on the left shows the frequency of cells classified within each shape mode for LMNA+/+ and LMNA-/- cells cultured on circular or triangular micropatterns (top and middle rows) and unpatterned surfaces (bottom row). The table on the right displays the values for traditional morphological parameters, including average area, shape factor (SF), and aspect ratio (AR) of cells, as well as the number of cells analyzed (#), lamin A/C status and the Shannon entropy of the cells. These results indicate that traditional morphological parameters insufficiently discriminate between the nuclear morphological responses of LMNA+/+ and LMNA-/- on different adhesive micropatterns (right table). In contrast, the differential morphological response of these cells is readily revealed when measured via shape mode distributions (left color-coded table). The reported values for each condition are the average abundance of cells based on two replicates of the same condition.

Figure 7. VAMPIRE analysis of human dermal fibroblasts from donors of different ages. A. Distributions of nuclear shape modes for dermal fibroblasts from age 3 to 96. Each row shows the distribution of shape modes for each donor the number of nuclei assessed are: #=643, 420, 407, 531, 373, 575, 637, respectively. The sample numbers of nuclei for each cell line are from two distinct replicates. Cells from younger donors populate the rounder shape mode (mode 1 and 2), while cells from older donors have nuclei classified that populate the irregular shape modes (mode 3, 4, and 7). B. Table showing the Pearson's correlation (R), shape factor (SF), and aspect ratio (AR) of each nuclear shape mode. R is the age correlation based on the abundance of nuclei in a specific shape mode. SF and AR are calculated as the mean of all nuclei classified in each shape mode across all ages.

Figure 8. Analysis of nuclear shape in H&E stained tissue sections with VAMPIRE. A. Images of a skin tissue section stained with hematoxylin and eosin (H&E) and obtained from the cancer genome atlas (TCGA case ID: TCGA-EE-A20I). Nuclei in the epidermis and the reticular dermis regions were segmented and analyzed with VAMPIRE. **B.** Bar graphs show the distribution of nuclei shape modes, comparing epidermal cells (N=1579) and dermal

cells (N=498) using VAMPIRE analysis. Numbers above the bars represent the abundances [%] of nuclei in each shape mode. Results also show a lower Shannon entropy in cells derived from the reticular dermis (S=2.1) relative to cells from the epidermis (S=2.25), indicating lower heterogeneity in the reticular dermis.