HHS Public Access

Author manuscript

Nat Protoc. Author manuscript; available in PMC 2021 June 01.

Published in final edited form as:

Nat Protoc. 2021 February ; 16(2): 754–774. doi:10.1038/s41596-020-00432-x.

A robust unsupervised machine-learning method to quantify the morphological heterogeneity of cells and nuclei

Jude M. Phillip[1,2,6], **Kyu-Sang Han**[1,6], **Wei-Chiang Chen**[1], **Denis Wirtz**[1,3,4,5], **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, MD, USA.

[2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA.

[3]Department of Pathology, Johns Hopkins School of Medicine, Baltimore, MD, USA.

[4]Department of Oncology, Johns Hopkins School of Medicine, Baltimore, MD, USA.

[5]Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD, USA.

[6]These authors contributed equally: Jude M. Phillip, Kyu-Sang Han.

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 from fluorescence or bright-field images using the VAMPIRE algorithm (https://github.com/kukionfr/VAMPIRE_open). 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 provides 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

Correspondence and requests for materials should be addressed to D.W. or P-H.W. wirtz@jhu.edu; pwu@jhu.edu.

Author contributions

J.M.P. and P.H.W. designed and conducted experiments; P.H.W., J.M.P., D.W. and W.C. conceived analysis and workflow of VAMPIRE; P.H.W. developed the original VAMPIRE software; K.S.H. converted the VAMPIRE software from MATLAB to Python; K.S.H. developed the graphical user interface of VAMPIRE; K.S.H. and J.M.P. analyzed and plotted data; P.H.W. and D.W. supervised the study; J.M.P., D.W., K.S.H. and P.H.W. wrote and edited the protocol; D.W., J.M.P., and P.H.W. secured funding.

Competing interests

The authors declare no competing interests.

Supplementary information is available for this paper at https://doi.org/10.1038/s41596-020-00432-x.

Peer review information *Nature Protocols* thanks the anonymous reviewers for their contribution to the peer review of this work.

to quantify the morphologies from 2D projections of cells seeded both on 2D substrates or embedded within 3D microenvironments, such as hydrogels and tissues. The complete analysis pipeline can be completed within 60 minutes for a dataset of ~20,000 cells/2,400 images.

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, which include metabolic, proteomic, epigenomic, and genomic components[1-6]. The coordinated orchestration of these interdependent cellular programs is critical to properly govern cellular behavior[4] and ultimately determines 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 the utility of 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[33,34]. However, defining and quantifying cellular shapes is 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 the perimeter), aspect ratio (long axis length/short axis length) and eccentricity (see Box 1 for a glossary of geometric and statistical descriptors), all measure the deviation of the shape of a cell from that of 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, they tend to insufficiently capture the true complexities of cell shapes[1]. To illustrate this, we describe here the morphologies of fluorescently labeled mouse embryonic fibroblasts (MEFs) using conventional shape features, including shape factor, aspect ratio, and solidity (see Box 1). From this analysis, we observe that taking a subset of cells having highly similar values of these parameters still results in a high degree of morphological variability among individual cells (Fig. 1). This example underscores the notion that conventional cell morphology parameters may be insufficient to capture cellular differences. Furthermore, mesenchymal cells on flat substrates or cells embedded within physiologically relevant 3D collagen gels, which often feature extensive dendritic protrusions and nuclear blebs[35-39], 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[14,30,40]. This has led to an expansion of morphological

descriptors (see Box 1), 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 is challenging.

To address this challenge, we recently developed cell morphology analysis software that provides improved visualization and quantitative analysis of complex shape morphologies[1,6,26]. 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 fluorescently labeled 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. This was most likely to be 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 Box 1) 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 stages: (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 (Fig. 2).

Following segmentation of the fluorescence or bright-field images, 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 is done based on Procrustes analysis[35,41,42]. To represent the highly complex shapes of cells and nuclei, a sufficient number of equally-spaced points along each contour (typically 50 points) (Fig. 2a) is used to define high-dimensional "features". Then, these coordinates along the boundaries of each cell and/or nucleus are subtracted by their mean value to shift the center of each cell and/or nucleus 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 nucleus, 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 nucleus, shape are then normalized by dividing the contour coordinates for each shape by the 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 conformation among shapes[1] (Fig. 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 Box 1). 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[43-46] (Fig. 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[47] (Fig. 2d). Among several classification methods tested, such as DBscan[48], OPTICS[49], Meanshift, and K-means, the K-means clustering algorithm was chosen for its fast calculation, robustness, and simplicity in setting the parameters.

Each cell and/or nucleus is then classified and sorted 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 (Fig. 2e). Lastly, using these representative shapes, together with the abundance of cells and/or nuclei 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 Box 1).

In our 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 ten

breast cancer cell lines comparing metastatic to nonmetastatic 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 Box 1).

In a separate study, we further demonstrated the utility of the VAMPIRE analysis by investigating how the morphologies of single-cell clones derived from a metastatic breast cancer cell line were associated with metastatic potential[6]. 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. 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 classify 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’ for a subset of this re-analyzed data).

In all the studies mentioned[1,6,26], the core algorithms of VAMPIRE analysis remain unchanged. However, 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 graphical 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 (https://github.com/kukionfr/VAMPIRE_open).

Overview of the procedure

The overall procedure is performed using four main stages: image segmentation from fluorescence/bright-field images of cells and nuclei (Step 1), formatting segmentation data before importing into the VAMPIRE GUI (Steps 2–3), generating a VAMPIRE model from a training set of images (Steps 4–10), and applying the VAMPIRE model to the training set or a new image set (Steps 11–13) (Fig. 3).

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 (see ‘Experimental design’ for more detail).

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 nonzero 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 *VAMPIRE datasheet c1.csv* in Supplementary Data 1 for list of parameters generated).

Selecting image sets in building and applying the model

Once the dataset 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). Hereafter, we refer to these specified images as the "training set". An example CSV file of this list, "*segmented image sets to build model.csv*", can be found in Supplementary Data 1. The resulting VAMPIRE model that is built, based on the specified training set, will be saved within a designated local folder. (Steps 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 (Steps 11–13). Ideally, users will apply the model to the same image set that was used for training. However, there are instances when it is appropriate to apply the VAMPIRE model to an entirely new dataset. For instance, (i) 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 conditions in building the model; (ii) 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; (iii) 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 valid only if the dataset used for training is expansive and similar enough (e.g., in cell type, dimensionality (2D/3D) or magnification) to represent the newly acquired data, as this 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 are classified within each shape mode (see 'Limitations of VAMPIRE' below).

The output of the VAMPIRE model includes a plot showing the frequency distribution of each shape mode per condition, and the CSV files that contain the shape mode for each cell and/or nucleus. (Step 13). Specifically, data for each cell includes the "*xy*" coordinates of cell centroid 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 Box 1). 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 allows users to further compare shape modes with other morphological features, and associate them with 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 Data 1, labeled "*CellProfiler datasheet c1.csv*" and "*VAMPIRE datasheet c1.csv*". See the directory of Supplementary Data 1 in the Supplementary Fig. 1 to locate example CSV files.

Applications of VAMPIRE

We have previously demonstrated the utility of VAMPIRE in three key studies, (i) the morphological changes displayed by human pancreatic cancer cells as they spread from the primary tumor to the liver[1], (ii) the ability of single-cell morphologies to encode metastatic potential in breast cancer[6], and (iii) the morphological changes of dermal fibroblasts derived from healthy individuals during aging[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 study cell morphologies in response to a wide range of physiochemical changes, i.e. molecular characteristics[2,3,38,39] (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 obtained 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 3D matrices, VAMPIRE analysis is applicable to study changes in cell and nuclear shapes in cells embedded within tissue sections (see ‘Anticipated results’). A growing number of studies have shown that nuclear shape can encode prognostic information for patients with different types of cancer[53,54]. Segmented nuclei within tissue sections can be imported directly into the VAMPIRE workflow to assess, for instance, changes in nuclear morphology that are associated 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 by 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 image stacks (*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 (i) increase the number of coordinate points (which will also increase computing time) or (ii) use more suitable morphological analyses that directly quantify cell protrusions[9] or take 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 Box 1). It is provided as part of the standard output data provided in “*VAMPIRE datasheet c1.csv*” under Supplementary Data 1. 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 be expansive enough and include cell types and conditions of interest. As 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 includes a subpopulation of cell shapes that is nonexistent in the dataset used for training, this would also result in misclassification of cells, and a large distance-from-cluster-center error for cells classified within each predefined shape mode.

Comparison with other methods

In this section, we briefly describe and compare other methods used to characterize the morphology of cells. The most common approach to quantify cell/nuclear shape morphology is to use scalar descriptors such as shape factor, curvature, and roughness[14,30,40]. This type of analysis is based on discriminative methods that try to capture just enough information to distinguish and investigate biological states[17]. Two commonly used tools for these types of cell shape analysis are CellProfiler[31] and MorpholibJ[55] (a plugin for ImageJ[33]). These tools extract an extensive list of features, such as shape factor, eccentricity, and Zernike polynomials. For instance, CellProfiler provides a set of ~1,500 morphological features to

describe the morphology of cells, including features that describe size, shape, intensity, and texture[14]. While the pure magnitude of the features assessed increases the likelihood of identifying differences among cell populations, this large number of descriptive features could limit the interpretation, visualization and integrative view of shape changes.

In many cases, methods to reduce the dimensions of cell shapes can be applied directly to binary images. Furthermore, these data-driven approaches and deterministic decomposition methods, such as Zernike polynomials and Fourier descriptors, are available to decompose the binary images of shapes and represent shapes with fewer dimensions. However, both methods are less effective in representing the cell shapes in lower dimension forms than PCA[44].

Discriminative methods of using scalar shape features are limited in their ability to describe cell shapes. Alternatively, methods that reduce the dimensionality of cell shapes can be used to reconstruct the cell shapes in a lower dimension for further clustering analysis. Particularly, principal component analysis (PCA) has been used to qualitatively and quantitatively identify novel insights in the relationship between cell morphology and physiology[43]. One key step of VAMPIRE is the reconstruction of cell shapes based on a lower-dimensional representation of cell shapes. This step involves the use of PCA on the aligned outlines of cell shapes, thereby retaining most of the cell shape information and variation within a given dataset. In VAMPIRE, the PCA step identifies the linear combination of shape vectors to regenerate the original cell shapes[56,57]. Nonlinear methods such as shape component analysis (SCA)[58] are also used in the field. SCA aims to preserve the distance (i.e., Euclidean) metrics between different shapes in a lower-dimensional shape space, to avoid potential distortion using non-Euclidean distances. However, in recent studies, SCA did not show significant improvement in reconstructing the cell shapes over PCA[17].

Recent advances in image modeling with neural networks have provided a way to derive lower-dimensional representations of cell shapes. Unsupervised learning approaches such autoencoder[59] and generative adversarial networks[60] have been extended to analyze the morphology of cells[61,62]. A recent study examining the performance of various autoencoder algorithms found that while outlined-based autoencoder methods perform similar to PCA in terms of shape representation accuracy at a lower dimension ($d = 7$), they underperform at a higher dimension ($d = 100$)[59]. Expectedly, the autoencoder methods take a lot more computational time to process compared to PCA. However, since the field of deep learning evolves rapidly, there is a strong potential for enhanced approaches representing cell shapes based on neural network methodologies.

Cell shapes are highly heterogeneous, even for cells within the same population. In VAMPIRE analysis, we utilize an unsupervised machine-learning clustering method in the reduced shape space from PCA to obtain subtypes of cells (shape modes). K-means clustering is an effective solution that works for various geometries of datasets with a simple input parameter (the number of clusters). However, K-means clustering could perform poorly on elongated clusters or irregular shapes of clusters[63]. Other clustering methods such as DBscan[48] and OPTICS[49] generate clusters based on the density of the data and better

handle complex geometry. However, the clustering results from these methods could be sensitive to clustering parameters.

To summarize, each user should decide the appropriate software solution for their morphology quantification based on the questions at hand.

Experimental design

Example image datasets—To help users explore the software and all its functionalities, we provide two small image datasets in the “Example images” folder of Supplementary Data 1. Note that users can also download Supplementary Data 1 from the GitHub repository (https://github.com/kukionfr/VAMPIRE_open/releases/download/v1.0/Supplementary.Data.zip). See the directory of Supplementary Data 1 in Supplementary Fig. 1 to locate example images and workflow. Results from the VAMPIRE analysis using provided image datasets are also included in Supplementary Fig. 2 and Supplementary Data 1 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 in the Procedure. In ‘Anticipated Results’, we also illustrate the utility of VAMPIRE analysis by analyzing the morphology of (i) 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, (ii) dermal fibroblasts derived from healthy individuals with increasing age, and (iii) for cells embedded within tissue sections.

Sample preparation and imaging—One of the most common ways to image the morphology of cells and nuclei is through fluorescent labeling of nuclear and cellular regions of cells using typical fixed and stain methods[1,64]. For 2D culture, the cell samples to be imaged should be first placed on an optical-compatible substrate such as a glass bottomed or transparent plastic dish or plates. Once desired experimental conditions of cell sample are achieved, cells should be fixed to preserve their structures. In general, we use paraformaldehyde (PFA) as a fixative to fix cell samples. However, alternative fixatives should be considered if subsequent dyes or stains to be used are not compatible with PFA. After fixation, the sample often needs a membrane permeabilization step, such as treatment with Triton X-100, to allow fluorescent probes to pass through the cell and nucleus membrane. We typically label the cell nuclei with H33342 and label F-actin with phalloidin to image cell cytoplasm. Other types of nuclear or cytosolic stain can also be used as long as they can provide clearly labeled nuclei or cell images. If there are cellular structures or molecular targets of interest, their corresponding fluorescent probes can be used together with cell or nuclei stain as long as these do not interfere with the signal of the nuclear or cellular stain. Imaging for VAMPIRE analysis is compatible with multicolor fluorescent images.

After cell samples are stained, a fluorescent microscopy system that can perform filter-based sequential multicolor imaging should be used to image the samples. We typically aim to acquire ~1,000 cells per condition replicate to ensure the subtype analysis in VAMPIRE can provide more statistically meaningful results. Hence, it is recommended to use a motorized

stage system on a microscope to allow for a more rigid and effective acquiring of images of samples at multiple points. Our typical acquisition routine is performed with a 10× objective lens scanning a 9 by 9 continuous field of view that covers approximately 6 mm by 6 mm regions. To acquire the fluorescent images, we maximize the power of excitation light and then minimize the exposure time to a level that produces a substantial but nonsaturated intensity signal relative to background intensity to improve the throughput of image acquisition. In our workflow, the cell sample to be imaged typically has a density of ~30 cell/mm^2. If imaging a sample with higher cell density, fewer scanning points may be considered to improve throughput while obtaining sufficient cell counts. A higher magnification objective lens can also be used (i.e., 20× or higher) for deriving better spatially resolved cell images. In this case, the number of scanning points may need to be increased to obtain sufficient cell numbers for VAMPIRE analysis.

For cells in a 3D culture system, such as cells embedded in collagen gel, a similar process to that described above can be applied to obtain the images for VAMPIRE analysis. Special consideration should be given to the diffusion of staining dyes and molecules within the 3D gel. Therefore, the duration of incubation steps will be lengthened, see published literature[65-67]. Also, in image acquisition, since cells are randomly distributed within the 3D space, single focal plane acquisition, as typically performed on 2D samples, is likely to capture images containing substantial cells that are not in focus. Thus, it is better to acquire multi-focal planes (i.e., multi *z*-steps) for each field of view. The *z*-projected images can produce cell images with more boundary resolved detail. For tissue section, samples are prepared based on standard fluorescence or immunohistochemistry(IHC) staining protocols for tissue sections[68,69]

Segmentation of cells and nuclei—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 (Step 1). However, note the additional steps may be needed if other segmentation software platforms are used.

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 cells to the centroid of each cluster (shape mode), where the distance from the cluster centroid is stored as the “distance from centroid” (see Box 1). This K-means clustering classifies 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 (Fig. 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 reduces 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 (Fig. 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[70], users can evaluate 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 8 GB of RAM running Microsoft Windows 10 (64 bit)

Software

- VAMPIRE executable software (https://github.com/kukionfr/VAMPIRE_open/releases/download/v1.0/vampire.exe).
- 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), or MATLAB (https://www.mathworks.com/downloads)
- Example dataset: Micropattern Data: https://github.com/kukionfr/Micropattern_MEF_LMNA_Image; Aging Data: https://github.com/kukionfr/Aging_human_dermal_fibroblast_nucleus. For a smaller example dataset, see Supplementary Data 1.

Procedure

▲ **CRITICAL** To demonstrate the VAMPIRE analysis procedure, we provide sample images of fluorescently tagged cells in Supplementary Data 1 under the "Example images" folder and the corresponding results from the VAMPIRE analysis procedure. Two sample sets—MEF_LMNA-- and MEF_wild type stained with Alexa Fluor 488 Phalloidin (Thermo Fisher Scientific)—are provided and correspond to mouse embryonic fibroblast cells having wild-type expression of Lamin A or Lamin A knockout, respectively (Fig. 5a). Throughout this Procedure, refer to the directory of Supplementary Data 1 in Supplementary Fig. 1 to locate example data and results.

Segment images of cells or nuclei ● Timing 10–60 min

▲ **CRITICAL** The segmentation procedure described in Steps 1 and 2 is designed specifically for CellProfiler (see https://cellprofiler.org/tutorials for more information). Alternatively, cells can also be segmented using ImageJ (https://imagej.net/Segmentation) or

MATLAB (https://www.mathworks.com/help/images/detecting-a-cell-using-image-segmentation.html).

1. Segment the fluorescence or bright-field images to identify the boundaries of cells and/or nuclei. The VAMPIRE GUI does not segment cells. Navigate to the CellProfiler website (https://cellprofiler.org/) to download the installer of version 3.1.9. After installing and launching the CellProfiler, the user should create and customize the pipeline for image segmentation following the instructions (https://cellprofiler.org/tutorials) that are best suited for their images. The pipeline that is customized for the provided example data is provided in Supplementary Data 1 (CellProfiler segmentation pipeline.cppipe). To use the provided pipeline, load the CellProfiler segmentation pipeline.cppipe in CellProfiler software under menu bar>File>Import>Pipeline from File. The provided pipeline consists of nine modules. Once the pipeline file is loaded, the pipeline will appear on the left panel of the CellProfiler main window. The user can use this pipeline to process the provided example fluorescence images, starting with loading the downloaded images by dragging and dropping to the “image module”. Once the pipeline (i.e. set of modules) is set up successfully, click the “Analyze image” button in the CellProfiler software to obtain segmented image of cells.

   ▲ **CRITICAL STEP** The current pipeline only supports a single channel of fluorescence. For multi-channel images see (https://cellprofiler.org/tutorials) for more information. We provide segmented example images using CellProfiler (Fig. 5a), as well as a sample CellProfiler segmentation pipeline in Supplementary Data 1. Note that the example workflow is designed using CellProfiler version 3.1.9, and may need to be adapted for compatibility with later versions of CellProfiler. The user modified pipeline file can be saved by navigating to File>Export>Pipeline from the menu bar.

   ? 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 stored as binary TIFF files, where the area of each cell must have a nonzero 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*).

   ▲ **CRITICAL STEP** A sample format of segmented images is provided in Supplementary Data 1 for reference.

Build shape-analysis VAMPIRE model (model training) ● 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:

- "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*).
- "condition name": description of an image set.
- "set location": the location/path of the folder containing segmented images
- "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.
- "note": any information about the image sets needed for the user's record. This information is not used in the VAMPIRE analysis.

  For more explanation in selecting the image sets for model training and application in Steps 3 and 11, please refer to Selecting image sets in building and applying the model section in 'Overview of the procedure'.

  ▲ **CRITICAL STEP** An example CSV file named "*Segmented image sets to build model.csv*" can be found in Supplementary Data 1. 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 1 for the following analysis, the user needs 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/download/v1.0/vampire.exe). Launch VAMPIRE Graphical User Interface (GUI) by opening the VAMPIRE.exe file.

▲ **CRITICAL STEP** The current version of VAMPIRE GUI is only available for Windows 10 users. Source codes are available on GitHub (https://github.com/kukionfr/VAMPIRE_open) and PyPI (https://pypi.org/project/vampireanalysis). These repositories will be continuously updated and maintained.

? TROUBLESHOOTING

5 Locate the CSV file generated in Step 3 to build a 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 the Build Model section of VAMPIRE GUI under the “number of coordinates” box. The default value is 50. 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 10. To optimize this number, refer to the ‘Selection of parameters for VAMPIRE analysis’ section in the ‘Experimental design’.

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 the 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 the 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, the 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*”.

▲ **CRITICAL STEP** Example output files of this step are provided in Supplementary Data 1, 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 (model application) ● Timing 1-10 min

11 Repeat Step 3 to specify the sets of segmented images to apply the VAMPIRE model to. 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 Fig. 1 to locate the pickle file.

13 Perform VAMPIRE analysis on the specified images by clicking the "Apply Model" in the 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 percentage of cells within each shape mode denoted on the top of each bar (Figs. 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 and minor axes
- 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: is calculated as the major axis length divided by the 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.

▲ **CRITICAL STEP** Example output files for this step are provided in Supplementary Data 1, under "Example output". These files are generated using the VAMPIRE model provided in Supplementary Data 1 under the same folder "Example output". See the directory of Supplementary Data 1 in Supplementary Fig. 1 to locate the output files. A compiled example of shape parameters of the VAMPIRE datasheet is shown in 'Anticipated results' (Fig. 6b).

Troubleshooting

Troubleshooting advicecan be found in Table 1.

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 using VAMPIRE for the first time.

Steps 1 and 2, segment images of cells or nuclei, 10–60 min

Steps 3–10, build shape-analysis VAMPIRE model, 3–10 min

Steps 11–13, analyze cell shapes with VAMPIRE model, 1–10 min

Total (Steps 1–13), complete VAMPIRE analysis, 14–80 min

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 2D substrates: 1. circular or 2. triangular shaped fibronectin-coated islands, surrounded by polyethylene glycol 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 (Fig. 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 (Fig. 6b). As expected, results showed that both LMNA$^{+/+}$ and LMNA$^{-/-}$ cells seeded on unpatterned surfaces exhibited mixed shape profiles 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 (modes 6–9) (Fig. 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[51,71]. Our results show

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 (Fig. 6b). However, looking across conditions, we observed 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, suggest that, compared with classical morphology parameters, VAMPIRE analysis can provide direct visual insight to better monitor the transition of cell morphology.

Using VAMPIRE analysis[26] we also examined the association between cellular morphology and chronological ages of dermal fibroblasts derived from seven healthy individuals. While the morphology of mouse embryonic fibroblast was emphasized by artificial micropatterns, this example illustrates the sensitivity of VAMPIRE 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, VAMPIRE analysis shows a decrease in the frequency of cells having rounded shape modes with rounded morphologies, and an increase in cells having irregular nuclear morphologies with increasing age. This is measured by 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 (Fig. 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 (Fig. 7b). Again, this demonstrates the utility of VAMPIRE analysis to visually and quantitatively identify morphological changes that would otherwise go unnoticed using traditional morphological parameters.

Lastly, applying the utility of VAMPIRE analysis beyond cultured cells, we have successfully implemented VAMPIRE analysis for the 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 (Fig. 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 a VAMPIRE model using nuclei segmented from the scanned image of an H&E stained skin tissue biopsy from a 79 year old donor. We observed that shape modes 1–3 were more elongated (i.e., less circular) relative to modes 4–10 (Fig. 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 1–3, compared to only 6.4% for epidermal cells (Fig. 8b).

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets generated and/or analyzed during the current study are available from GitHub: Micropattern Data (https://github.com/kukionfr/Micropattern_MEF_LMNA_Image) and Aging Data (https://github.com/kukionfr/Aging_human_dermal_fibroblast_nucleus). A smaller example dataset is provided as Supplementary Data 1 and is also deposited on GitHub: https://github.com/kukionfr/VAMPIRE_open/releases/download/v1.0/Supplementary.Data.zip.

Code availability

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

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported in part by National Institutes of Health grants U54CA143868 (D.W.), R01CA174388 (D.W.), P30AG021334 (P.H.W. and J.M.P.) and U01AG060903 (D.W., J.M.P. and P.H.W.).

References

1. Wu PH et al. Evolution of cellular morpho-phenotypes in cancer metastasis. Sci. Rep 5, 1–10 (2015).
2. Chen W-C et al. Functional interplay between the cell cycle and cell phenotypes. Integr. Biol 5, 523–34 (2013).
3. Chambliss AB, Wu PH, Chen WC, Sun SX & Wirtz D Simultaneously defining cell phenotypes, cell cycle, and chromatin modifications at single-cell resolution. FASEB J 27, 2667–2676 (2013). [PubMed: 23538711]
4. Bakal C, Aach J, Church G & Perrimon N Quantitative morphological signatures define local signaling networks regulating cell morphology. Science 316, 1753–1756 (2007). [PubMed: 17588932]
5. Rohban MH et al. Systematic morphological profiling of human gene and allele function via cell painting. eLife 6, e24060 (2017). [PubMed: 28315521]
6. Wu P-H et al. Single-cell morphology encodes metastatic potential. Sci. Adv 10.1126/sciadv.aaw6938 (2020).
7. Driscoll MK et al. Robust and automated detection of subcellular morphological motifs in 3D microscopy images. Nat. Methods 10.1038/s41592-019-0539-z (2019).
8. Yeung T et al. Effects of substrate stiffness on cell morphology, cytoskeletal structure, and adhesion. Cell Motil. Cytoskeleton 10.1002/cm.20041 (2005).

9. Guo Q et al. Modulation of keratocyte phenotype by collagen fibril nanoarchitecture in membranes for corneal repair. Biomaterials 34, 9365–9372 (2013). [PubMed: 24041426]
10. Sero JE 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). [PubMed: 26148352]
11. Simm J et al. Repurposing high-throughput image assays enables biological activity prediction for drug discovery. Cell Chem. Biol 10.1016/j.chembiol.2018.01.015 (2018).
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 MJ et al. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. Proc. Natl Acad. Sci. USA 111, 10911–10916 (2014). [PubMed: 25024206]
14. Bray MA et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat. Protoc 11, 1757–1774 (2016). [PubMed: 27560178]
15. Beghin A et al. Localization-based super-resolution imaging meets high-content screening. Nat. Methods 14, 1184–1190 (2017). [PubMed: 29083400]
16. Meijering E, Carpenter AE, Peng H, Hamprecht FA & Olivo-Marin JC Imagining the future of bioimage analysis. Nat. Biotechnol 34, 1250–1255 (2016). [PubMed: 27926723]
17. Ruan X & Murphy RF Evaluation of methods for generative modeling of cell and nuclear shape. Bioinformatics 10.1093/bioinformatics/bty983 (2019).
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). [PubMed: 28647475]
19. Danuser G Computer vision in cell biology. Cell 147, 973–978 (2011). [PubMed: 22118455]
20. Falk T et al. U-Net: deep learning for cell counting, detection, and morphometry. Nat. Methods 16, 67–70 (2019). [PubMed: 30559429]
21. Chicco D Ten quick tips for machine learning in computational biology. BioData Mining 10.1186/s13040-017-0155-3 (2017).
22. Gabril MY & Yousef GM Informatics for practicing anatomical pathologists: Marking a new era in pathology practice. Modern Pathol. 23, 349–358 (2010).
23. Fuchs TJ & Buhmann JM Computational pathology: challenges and promises for tissue analysis. Computer. Med. Imag. Graphics 35, 515–530 (2011).
24. Sarnecki JS et al. A robust nonlinear tissue-component discrimination method for computational pathology. Lab. Investig 96, 450–458 (2016). [PubMed: 26779829]
25. Beck AH et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci. Transl. Med 3, 108ra113–108ra113 (2011).
26. Phillip JM et al. Biophysical and biomolecular determination of cellular age in humans. Nat. Biomed. Eng 1, 0093 (2017). [PubMed: 31372309]
27. Pegoraro G & Misteli T High-throughput imaging for the discovery of cellular mechanisms of disease. Trends Genet. 33, 604–615 (2017). [PubMed: 28732598]
28. Lang P, Yeow K, Nichols A & Scheer A Cellular imaging in drug discovery. Nat. Rev. Drug Discov 5, 343–356 (2006). [PubMed: 16582878]
29. Loo LH, Wu LF & Altschuler SJ Image-based multivariate profiling of drug responses from single cells. Nat. Methods 4, 445–453 (2007). [PubMed: 17401369]
30. Sailem HZ, Sero JE & 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. 10.1371/journal.pbio.2005970 (2018).
32. Carpenter AE et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 10.1186/gb-2006-7-10-r100 (2016).
33. Schindelin J et al. Fiji: An open-source platform for biological-image analysis. Nat. Methods 10.1038/nmeth.2019 (2012).
34. Jayatilaka H et al. EB1 and cytoplasmic dynein mediate protrusion dynamics for efficient 3-dimensional cell migration. FASEB J. 10.1096/fj.201700444RR (2018).

35. Jayatilaka H et al. Synergistic IL-6 and IL-8 paracrine signalling pathway infers a strategy to inhibit tumour cell migration. Nat. Commun 8, 15584 (2017). [PubMed: 28548090]
36. Jayatilaka H et al. Tumor cell density regulates matrix metalloproteinases for enhanced migration. Oncotarget 9, 32556–32569 (2018). [PubMed: 30220965]
37. Phillip JM, Aifuwa I, Walston J & Wirtz D The mechanobiology of aging. Annu. Rev. Biomed. Eng 17, 113–141 (2015). [PubMed: 26643020]
38. Kim D-H et al. Volume regulation and shape bifurcation in the cell nucleus. J. Cell Sci 129, 457–457 (2016). [PubMed: 26773007]
39. 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). [PubMed: 26096845]
40. Driscoll MK et al. Automated image analysis of nuclear shape: What can we learn from a prematurely aged cell? Aging 4, 119–132 (2012). [PubMed: 22354768]
41. Bookstein FL Landmark methods for forms without landmarks: Morphometrics of group differences in outline shape. Med. Image Anal 10.1016/S1361-8415(97)85012-8 (1997).
42. Dryden IL & Mardia KV Statistical Shape Analysis, with Applications in R 2nd edn. 10.1002/9781119072492 (2016).
43. Keren K et al. Mechanism of shape determination in motile cells. Nature 453, 475–480 (2008). [PubMed: 18497816]
44. Pincus Z & Theriot JA Comparison of quantitative methods for cell-shape analysis. J. Microsc 227, 140–156 (2007). [PubMed: 17845709]
45. MacLeod N Generalizing and extending the eigenshape method of shape space visualization and analysis. Paleobiology 25, 107–138 (1999).
46. Tsai A et al. A shape-based approach to the segmentation of medical imagery using level sets. IEEE Trans. Med. Imaging 10.1109/TMI.2002.808355 (2003).
47. Pedregosa F et al. Scikit-learn: machine learning in Python. J. Mach. Learn. Res (2011).
48. Ester M, Kriegel H-P, Sander J & Xu X A Density-based algorithm for discovering clusters in large spatial databases with noise. in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (1996).
49. Ankerst M, Breunig MM, Kriegel HP & Sander J OPTICS: ordering points to identify the clustering structure. SIGMOD Rec. 28, 49–60 (1999).
50. Kim DH & Wirtz D Focal adhesion size uniquely predicts cell migration. FASEB J. 27, 1351–1361 (2013). [PubMed: 23254340]
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). [PubMed: 29242553]
52. Zheng W, Thorne N & McKew JC Phenotypic screens as a renewed approach for drug discovery. Drug Discov. Today 10.1016/j.drudis.2013.07.001 (2003).
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 10.4103/JOC.JOC_237_16 (2003).
54. Seethala RR et al. Noninvasive follicular thyroid neoplasm with papillary-like nuclear features: a review for pathologists. Mod. Pathol 10.1038/modpathol.2017.130 (2018).
55. Legland D, Arganda-Carreras I & Andrey P MorphoLibJ: integrated library and plugins for mathematical morphology with ImageJ. Bioinformatics 10.1093/bioinformatics/btw413 (2016).
56. Abdi H & Williams LJ Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2, 433–459 (2010).
57. Shlens J A tutorial on principal component analysis. Preprint at https://arxiv.org/abs/1404.1100 (2014).
58. Lee HC, Liao T, Zhang YJ & Yang G Shape component analysis: Structure-preserving dimension reduction on biological shape spaces. Bioinformatics 10.1093/bioinformatics/btv648 (2016).
59. Hinton GE & Salakhutdinov RR Reducing the dimensionality of data with neural networks. Science 10.1126/science.1127647 (2006).

60. Goodfellow IJ et al. Generative adversarial nets. GitHub http://www.github.com/goodfeli/adversarial.
61. Osokin A, Chessel A, Salas REC & Vaggi F GANs for biological image synthesis. Proc. IEEE Int. Conf. Comput. Vis 2017, 2252–2261 (2017).
62. Johnson GR, Donovan-Maiye RM & Maleckar MM Generative modeling with conditional autoencoders: building an integrated cell. Preprint at arXiv https://arxiv.org/abs/1705.00092 (2017).
63. Liberti L Distance geometry and data science. TOP 28, 271–339 (2020).
64. Donaldson JG Immunofluorescence staining. Curr. Protoc. Cell Biol 60, 4.3.1–4.3.6 (1998).
65. 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). [PubMed: 23796785]
66. Fraley SI et al. A distinctive role for focal adhesion proteins in three-dimensional cell motility. Nat. Cell Biol 10.1038/ncb2062 (2010).
67. Artym VV & Matsumoto K Imaging cells in three-dimensional collagen matrix. Curr. Protoc. Cell Biol 10.1002/0471143030.cb1018s48 (2010).
68. Fischer AH, Jacobson KA, Rose J & Zeller R Hematoxylin and eosin staining of tissueand cell sections. Cold Spring Harb. Protoc 3, pdb.prot4986 (2008).
69. Kim SW, Roh J & Park CS Immunohistochemistry for pathologists: Protocols, pitfalls, and tips. J. Pathol. Transl. Med 50, 411–418 (2016). [PubMed: 27809448]
70. Hale CM et al. SMRT analysis of MTOC and nuclear positioning reveals the role of EB1 and LIC1 in single-cell polarization. J. Cell Sci 124, 4267–4285 (2011). [PubMed: 22193958]
71. Kim DH & Wirtz D Cytoskeletal tension induces the polarized architecture of the nucleus. Biomaterials 48, 161–172 (2015). [PubMed: 25701041]
72. Hale CM et al. SMRT analysis of MTOC and nuclear positioning reveals the role of EB1 and LIC1 in single-cell polarization. J. Cell Sci 124, 4267–4285 (2011). [PubMed: 22193958]

Box 1 |

Glossary of geometric and statistical descriptors

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 calculated as the reciprocal of the radius of a circle fitted at each boundary point[40].

Roughness—defined as the variance 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 comprising 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—the Euclidean distance between the morphology parameters of a cell and the centroid of the cluster it belongs to. The morphology parameters of the cell are represented from the reduced number of the principal components from PCA.

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 many cell generations.

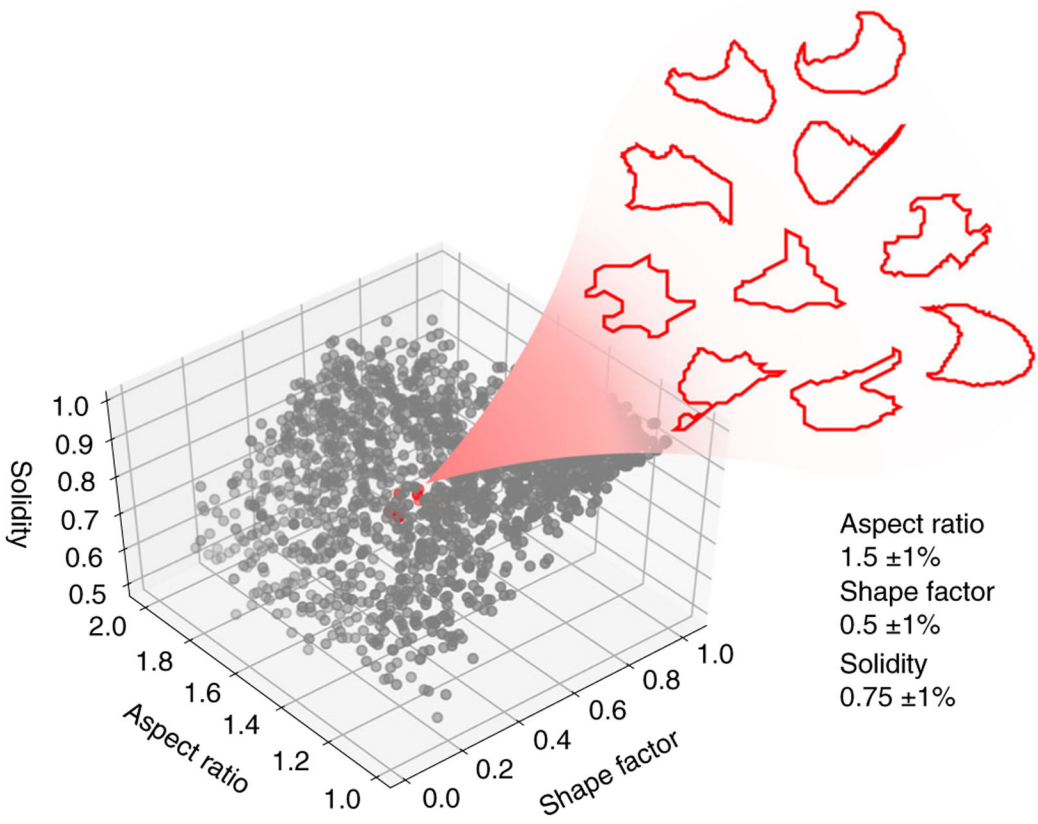

Fig. 1 |. Cells confined to narrow ranges of traditional morphological parameters still exhibit highly variable shapes.

Scatter plot showing the distributions of 37,750 mouse embryonic fibroblast cells confined to a 3D 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.

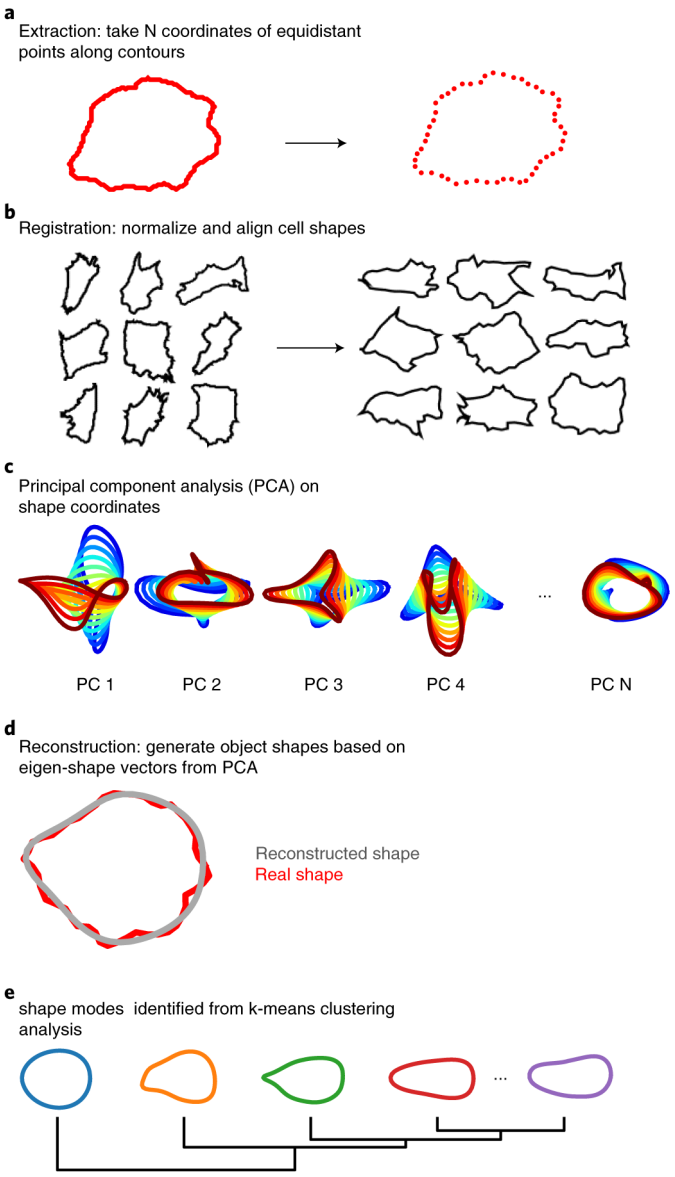

Fig. 2 |. Overview of VAMPIRE analysis, from the extraction of contour coordinates to the automatic generation of shape modes.

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 to 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.

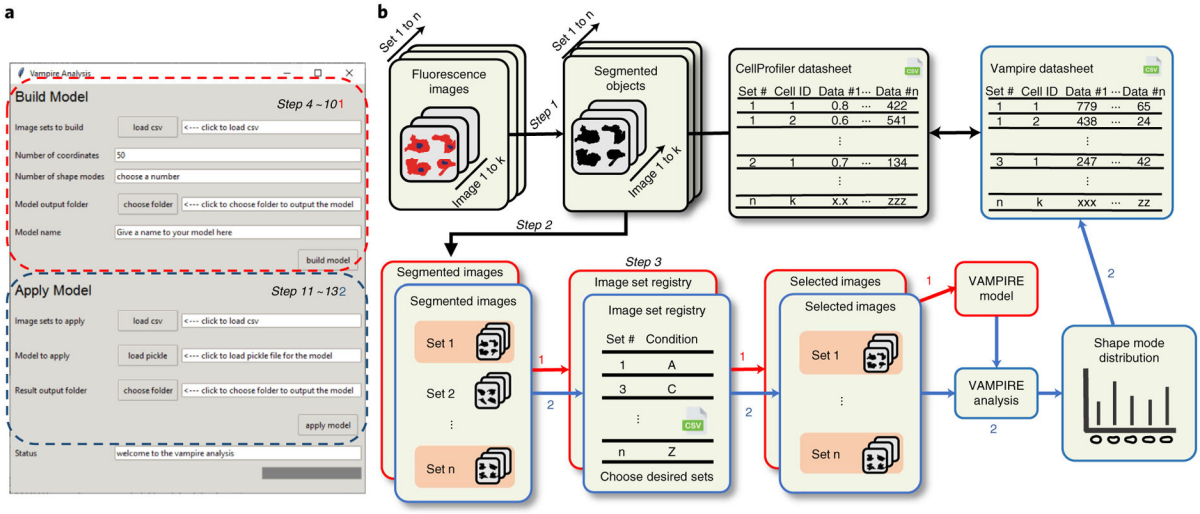

Fig. 3 |. Overview of VAMPIRE implementation with the VAMPIRE GUI.

a, The VAMPIRE Graphical 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. The 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).

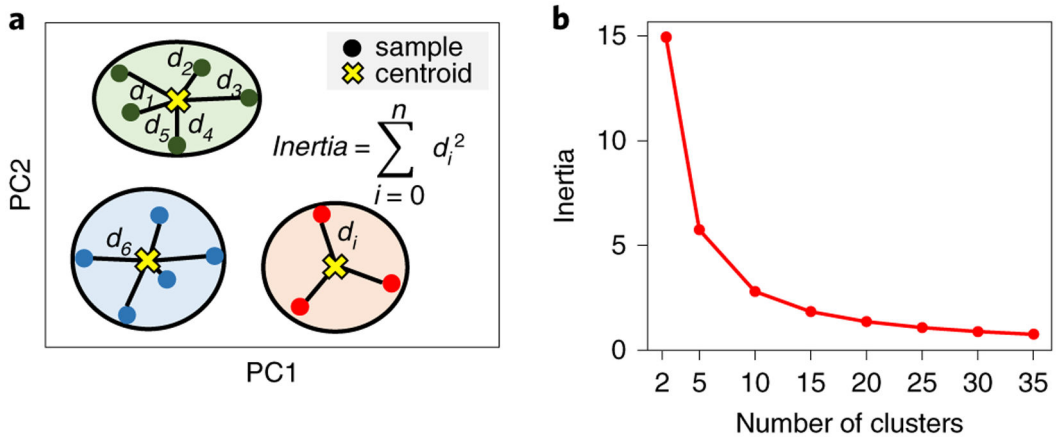

Fig. 4 |. Determinants of cluster coherence in the shape mode distributions.

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. A lower inertia value indicates better segregation of clusters indicating more intercluster 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 inertia value is calculated on ten separate runs of VAMPIRE analysis at each cluster number parameter value. In each run, the K-means clustering is by default repeated five times with different centroid seeds to find the initial seed that results in the lowest inertia value. The coefficient of variation of inertia between ten separate runs of VAMPIRE is less than 0.05% for this inertia profile, thus an error bar is not shown.

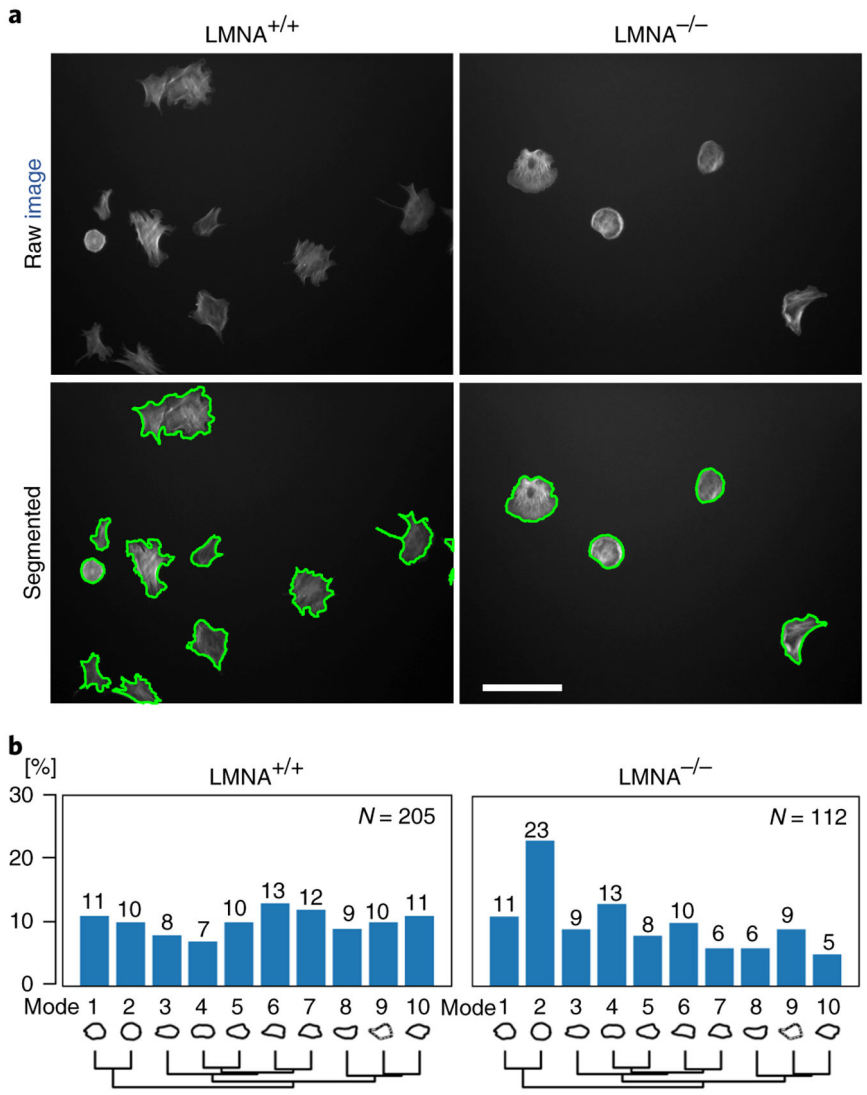

Fig. 5 |. VAMPIRE analysis of *LMNA*$^{+/+}$ and *LMNA*$^{-/-}$ mouse embryonic fibroblasts.
a, Images of phalloidin-stained (top) wild-type (LMNA$^{+/+}$, left) and lamin-deficient (LMNA$^{-/-}$, right) mouse embryonic fibroblasts. Segmentation is obtained using CellProfiler. Scale bar, 100 μm. **b**, 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.

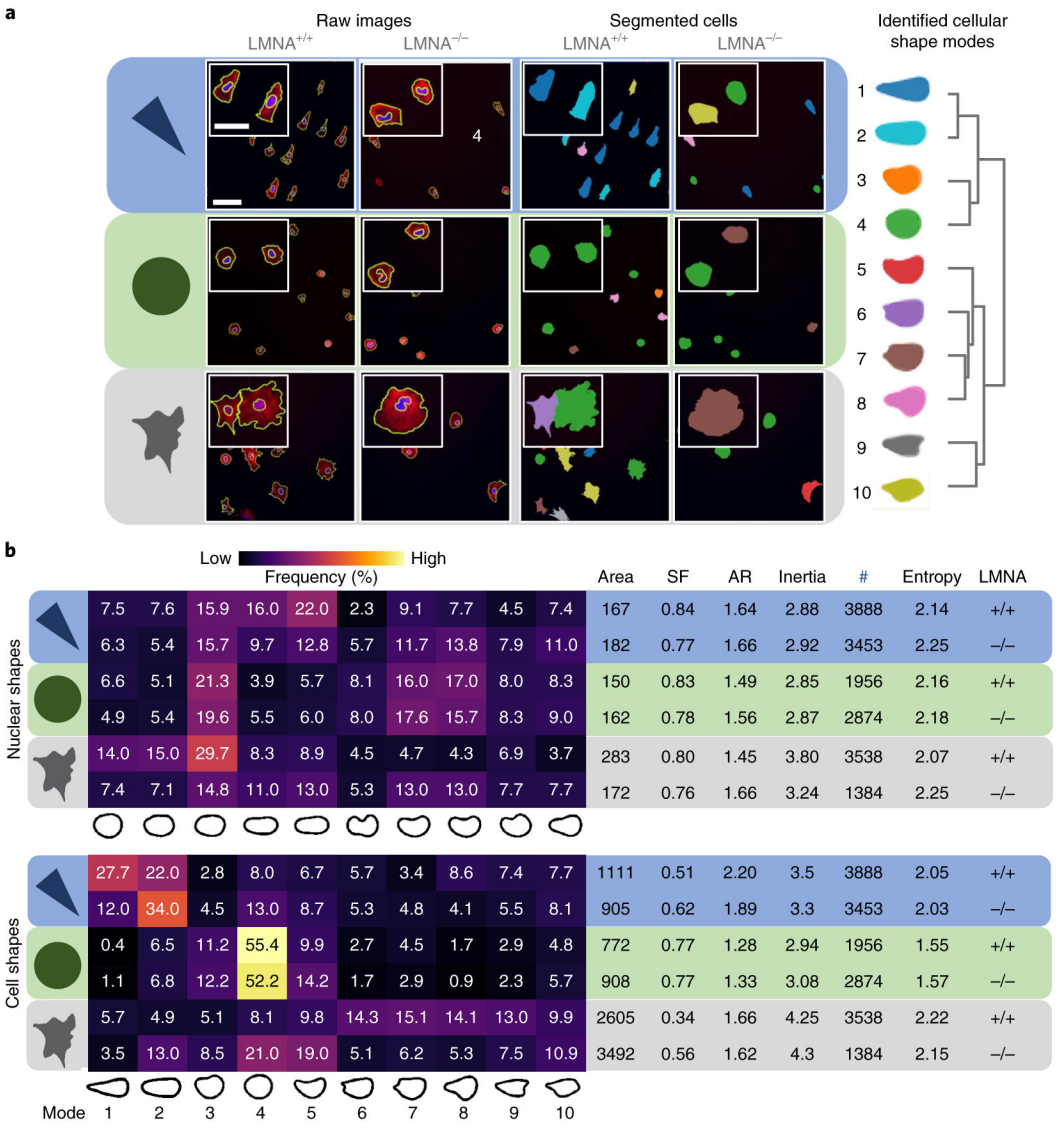

Fig. 6 |. VAMPIRE analysis of mouse embryonic fibroblasts seeded on adhesive micro-patterned surfaces.

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[72]. Control cells (bottom row) are placed on the fibronectin-coated glass. Cells were fixed and stained for F-actin using Alexa Fluor 488 Phalloidin (red) and nuclear DNA using DAPI (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 right are the same cells color coded according to the shape mode to which they belong. Scale bar, 100 μm. Inserts are magnified views of cells; scale bar, 50 μm. 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). By contrast, the differential morphological response of these cells is readily revealed when measured by 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.

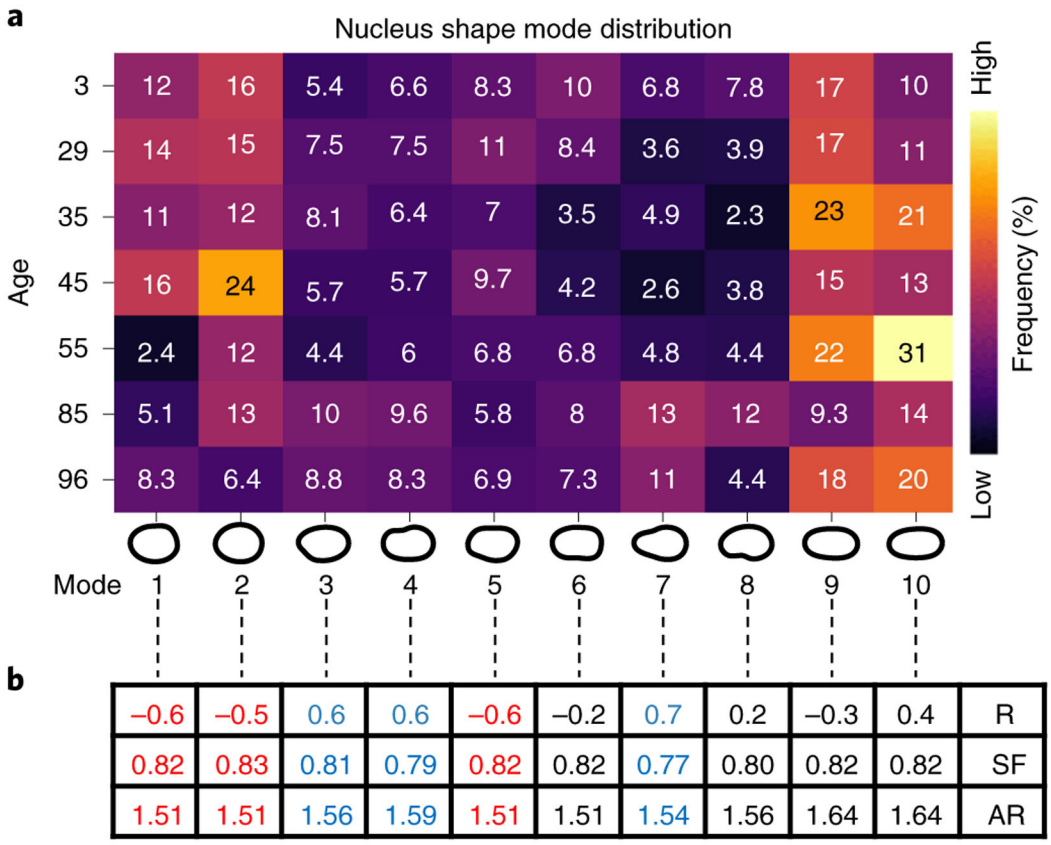

Fig. 7 | VAMPIRE analysis of human dermal fibroblasts from donors of different ages.
a, Distributions of nuclear shape modes for dermal fibroblasts from age 3 to 96. Each row shows the distribution of shape modes for each donor. The number of nuclei assessed are: # = 643, 420, 407, 531, 373, 575, 637, respectively. The sample numbers of nuclei for each cell line are from two distinct replicates. Cells from younger donors populate the rounder shape modes (modes 1 and 2), while cells from older donors have nuclei classified that populate the irregular shape modes (modes 3, 4, and 7). **b**, Table showing 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.

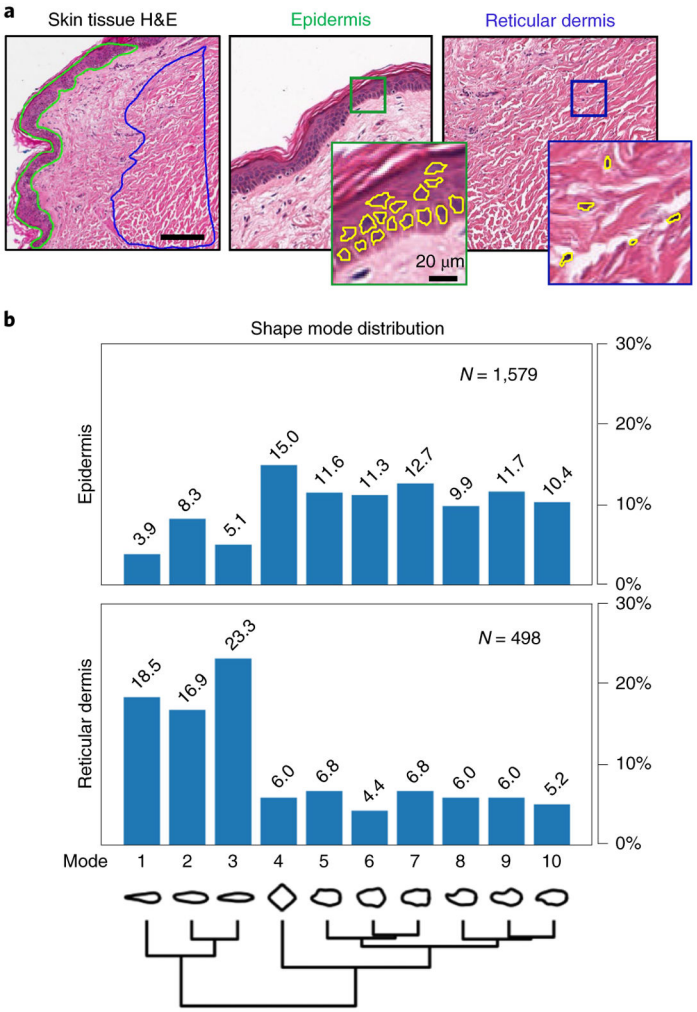

Fig. 8 |. Analysis of nuclear shape in H&E stained tissue sections with VAMPIRE.

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 = 1,579) 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.

Table 1 |

Troubleshooting table

Step	Problem	Possible reason	Solution
1	Cannot run the segmentation 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
	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	A warning is given that the MATPLOTLIBDATA environment variable is deprecated in Matplotlib 3.1 and will be removed in 3.3	The executable file of VAMPIRE is created using software that uses a variable that will be removed in the future	Ignore this message since VAMPIRE is not affected by this warning
10	The following warning appears: “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 that they have at least one cell or nucleus
	The following warning appears: “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
	The following warning appears: “Permission denied”	CSV file is open while the analysis is running	Close all open CSV files and repeat Step 10