**A robust method to quantify cell morphological heterogeneity**

Jude M. Phillip1, 5,\*, Kyu-Sang Han1,\*, Wei-Chiang Chen1, Denis Wirtz1-4, +, Pei-Hsun Wu1, +

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](mailto:pwu@jhu.edu) and D.W. [wirtz@jhu.edu](mailto: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, and is commonly employed 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 a cell population remains a challenge. Here we present a protocol and software for the analysis of cell morphology data using the VAMPIRE algorithm. This algorithm enables cell profiling through the 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 relate cell shapes with cellular subtypes, which are defined by 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, taking less than 10 min to analyze 10,000 cells.

**INTRODUCTION**

Cell morphology is commonly employed by clinicians and researchers in the study, diagnosis, prognosis, and treatment of human diseases. Fundamentally, cellular morphology represent the ensemble imprints of highly interactive molecular networks, including metabolic, proteomic, epigenomic and genomic components1–5. The coordinated orchestration of these inter-dependent cellular programs are critical to properly govern cellular behavior4 and ultimately determine cell responses to perturbations and stressors, such as microenvironmental cues, biomechanical stimuli6,7, and pharmacological treatments8–10. Advances in high-content imaging11,12, image processing11,13,14, and machine learning15–17 have greatly improved the throughput and accuracy of cell morphological measurements and have bolstered its utility in digital pathology18–21, biomarker identification1,22, and phenotypic screens9,23–25.

Cell morphology is traditionally quantified using a handful of geometric parameters11,26, 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 measurements of protein expression based on fluorescent labels and their intensity patterns and localization within the cell. Measuring cell and nuclear sizes can be readily achieved using open-source segmentation algorithms, such as CellProfiler27 and ImageJ/Fiji28,29.

However, defining and quantifying cellular shapes are more complicated. Classically, shape descriptors, such as shape factor (4πA/P2, where A is the surface area of the morphology and P its 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 cell conditions, these parameters tend to insufficiently capture the complexities of cell shapes1.

To illustrate this issue, we described the morphologies of mouse embryonic fibroblasts (MEFs) using conventional shape features, including shape factor, solidity, and aspect ratio. Taking a subset of cells which have highly similar values of these parameters, we still observed a high degree of morphological variability among individual cells, suggesting that conventional cell morphology parameters may be insufficient in capturing cellular differences (**Figure 1**).Furthermore, mesenchymal cells on flat substrates, or cells embedded inside more physiological 3D collagen gels with extensive dendritic protrusions and nuclear blebs35–38 are similarly difficult to distinguish using these traditional parameters.

A popular approach to address this shortcoming consists in defining additional geometric descriptors of cell morphology, some of which are based on the curvature and roughness of the cell and nuclear contours11,26. 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 allow users to capture more complex cell morphologies, effective visualization of 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 to provide improved visualization and quantitative analysis of complex morphological changes in cells. The software, which we named Visually Aided Morpho-Phenotyping Image Recognition (VAMPIRE), is highly automated and allows to process large datasets rapidly.

**Development of the VAMPIRE 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. 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 metastatic (LM; four lines), and two were non-neoplastic pancreatic epithelial cell lines. For direct visual assessment of cell and nuclear shapes, we randomly selected subsets of individual cells traces after alignment and found no overt morphological differences between primary tumor cells and liver-metastatsis cells, due partly to the irregularity of cell shapes.

To measure cell shapes, we examined commonly used morphological features, such as spreading area, shape factor, and aspect ratio. 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 parameters1,22. VAMPIRE analysis is also a visual aid that compares cell morphologies by first identifiying representative shape modes (see **Glossary**) among all cell shapes present within cell populations, then determines the abundance of cells classified within each shape mode. VAMPIRE comprises four essential computational steps (**Figure 2**): I) the determination and registration of the coordinates of equally-spaced points along nuclear and cellular contours to define morphological descriptors; II) the reduction in the number of morphological descriptors using principal component analysis (PCA); III) the identification of shape modes through unsupervised clustering analysis, and IV) the analysis of shape mode distributions of all tested cell samples.

To represent highly complex shapes of cells and nuclei, a sufficient number of equally spaced points along their boundaries or contours (typically 50, **Figure 2A**) are used to define high-dimensional “features”. These coordinates of contours are normalized to unify cell sizes and aligned to eliminate any effects of rotational variations among cells (**Figure 2B**). After the cell contours are normalized and registered, eigen shape vectors from the PCA (see **glossary**) that comprise 95% of the observed variance are computed and used as reduced features for cell and/or nuclei shapes (**Figure 2C**). Next, representative shape modes are empirically identified from the populations of cells by applying K-means clustering algorithm39 to the reduced shape features determined from PCA (**Figure 2D-E**), then each cell and/or nucleus is binned into a shape mode. The centroid locations of each cluster in the PCA-reduced features from K-means clustering analysis can be directly used to reconstruct and visualize the morphology of each shape mode. The shape modes are then used to classify the shapes of individual cells and occurrences of cells in different shape modes per condition provides both quantitative and visual insights of cell morphology. Among several classification methods such as DBscan, OPTICS, Meanshift, and K-means, we chose to use K-means clustering algorithm for VAMPIRE analysis because of its fast calculation and simplicity in setting the parameters.

In the study of pancreatic cancer cells (see above), VAMPIRE analysis showed that metastasized cells present significantly lower heterogeneity than primary tumor cells. We also found a similar morphological signature of metastasis for a cohort of 10 breast cancer cell lines­1. We further deciphered the relative contributions to heterogeneity from cell cycle, cell-cell contacts, cell stochasticity and heritable morphological variations. In a separate study, the utility of VAMPIRE analysis was further demonstrated by investigating the morphological signature of healthy aging using skin dermal fibroblast cells22. We found that cellular age could be used to stratify individuals on the basis of cell morphology using a cohort of 32 samples of primary dermal fibroblasts collected from individuals between 2 and 96 years of age.

**Overview of the protocol**

In this protocol, for the accessibility and easiness to perform the VAMPIRE shape analysis, we established a python-based graphic user interface (GUI), VAMPIRE GUI. We note that VAMPIRE GUI does not provide a segmentation tool; it analyzes cell and nuclear shapes that are already detected and segmented. The segmentation can be performed using software such as ImageJ/FiJi 28,29 and CellProfiler27 with easy integration of the segmentation results into VAMPIRE GUI.

To help the user explore the software and all its functionalities, we provide two small image datasets in the **Supplementary files** under“example fluorescence images”**.** See the directory structure of supplementary files in **Supplementary information S1** to locate any provided example images and workflow in this **Overview of the protocol** section.The expected results of VAMPIRE analysis using provided image datasets are also included in **Supplemental information** **S4** and **Supplementary files** under “expected results”. Before applying VAMPIRE analysis to their own data, users should first install and run VAMPIRE analysis using provided image datasets following the detailed procedure provided in the **Procedure** section.  We also illustrate the power 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, and for dermal fibroblasts derived from healthy individuals of increasing age (see **Anticipated results**).

The overall protocol is composed of five main parts: installation of VAMPIRE, segmentation of images, formatting segmentation data, generating a VAMPIRE model, and applying the VAMPIRE model. After installation of VAMPIRE (**Step 1-3**)**,** the procedure continues with the segmentation of fluorescent images of cells and corresponding nuclei to generate binary greyscale images (**Step 4**). These segmented images can be obtained using standard segmentation tools, including but not limited to CellProfiler and ImageJ. In the example workflow/data, we use CellProfiler to segment the fluorescent images of cells and nuclei.

To import segmented cells and nuclei to VAMPIRE, the segmentation data needs to be organized in a designated format for the use in the VAMPIRE GUI (**Step 5**). This required format is standard and usually requires only a few changes from routine segmentation outputs. If CellProfiler (version 3.0.0 or above) is used to segment cells, we have provided a function in the VAMPIRE GUI to help organize the segmented image outputs in the compatiable format. This step is recommended since CellProfiler outputs the contours of individual cells and nuclei as individual images, which can reduce the amount of data stored on the local hard-drive. Step 4A allows the user to combine these separate images into a single 16-bit greyscale image. Once binary images are imported into the VAMPIRE workflow, it converts each image to a greyscale image where the contours are recorded together with the coordinates of coutour points, and a subset of other morphological parameters.

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 them in a comma separated values (CSV) file (**Step 6**). Note that we refer 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 files**. The VAMPIRE model built based on the training set will be saved in a designated local folder. (**Step 7-11**). The model can then be applied to a new set of segmented images by specifying them in a new CSV file (**Step 12-14**). We note that VAMPIRE analysis can also be done by training the model and applying the same model to the training set.

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 information for individual cells and nuclei. (Steps 14). Specifically, data for each cell and nucleus include: the “xy” coordinates of cell and nuclear centroids within the image, the spread 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 “contour fit” (see **Glossary**). The datasheet can be directly linked to the features generated by CellProfiler, which makes VAMPIRE and CellProfiler analyses complementary. Example datasheets showing the results of analysis using both platforms are provided in **Supplementary files**, labeled “cellprofiler output.csv” and “vampire output.csv”.

For all the procedures, approximate timing is indicated. This time corresponds to the time it takes for an experienced user. More time may be required when using VAMPIRE for the first time.

**Applications of VAMPIRE**

We have previously demonstrated the utility of VAMPIRE with two studies, (a) the morphological changes displayed by human pancreatic cancer cells as they spread from the primary tumor to the liver1, and (b) the morphological changes that dermal fibroblasts derived from individuals undergo during healthy aging22.

In general, VAMPIRE can be applied to any set of segmented images of cells and nuclei to detect and analyze changes in their morphology across multiple conditions in cell-culture model systems. For instance, VAMPIRE can be applied to the study of cell morphology changes in response to changes in cell cycle state and genetic and epigenetic status1–3,6,22,36–38. VAMPIRE analysis is similarly applicable to other types of perturbations, including mechanical perturbations40,41. VAMPIRE analysis is also suitable for drug screening 8,9,12. Changes in cell morphology are often used in high-throughput biochemical and discovery screens42. The large volume of data that is typically generated in such screens makes it difficult to visually inspect cell responses. Here, the application of VAMPIRE provides users with the ability to rapidly classify phenotypically distinct cellular conditions in large amount of data and identify drug-induced changes in the frequency distributions of shape modes.

VAMPIRE analysis can 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 matrices1. In that study, we obtained the 2D contours of cells from the z-projected images of these cells. VAMPIRE analysis showed that shape modes for cells in 3D cultures were distinctly more protrusive than the same cells in more traditional 2D cultures1. VAMPIRE analysis can also be applied to study changes in nuclear shapes in tissue sections. A growing number of studies have shown that nuclear shape can encode prognostic information for patients in different types of cancers43,44. Segmented nuclei of cells in tissue sections can be imported directly into the VAMPIRE worflow to assess changes in nuclear morphology associated with tumor progression and patients outcomes1.

**Limitations of VAMPIRE**

An 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, for instance using 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 shape 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.

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 outlines to examine the complexity of cell shapes, the use of a reduced number of coordinates along with the PCA can still lead to detected shape modes of limited spatial resolution. In this case, a morphological analysis that provides for the direct quantification of cell protrusions6 or takes better account of cell protrusions26 is more suitable. The distance between the computationally assigned shape modes and actual cell shapes provides a direct way to examine the fitness of the VAMPIRE model applied to any given image dataset. If this distance is large, the VAMPIRE model has failed. To further examine the fitness of VAMPIRE analysis, we have included a measure called “contour fit” (see **Glossary**), which delineates the goodness of the match between shape modes and individual cell/nuclear contours as part of the output data provided to users. Note that the fitness score depends on the parameters used in VAMPIRE analysis and can be improved by increasing the number of shape modes or eliminating “outlier” cells, as described in details below in the **Experimental design** section.

**Comparison with other methods**

Two commonly used tools for cell shape analysis are CellProfiler’s measurement modules and MorpholibJ45, a plugin for ImageJ28 These tools segment images and 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. Here, VAMPIRE presents three major advantages compared to conventional approaches to assess and visualize cell morphology 11,26,29: 1) VAMPIRE provides an *in silico* visualization scheme that makes it easy to compare and learn biologically meaningful information from complex morphology data1,22. 2) VAMPIRE bypasses the feature-extraction step to avoid using overly abstract morphological parameters. 3) VAMPIRE provides a direct means to compute the degree of cellular heterogeneity in cell populations across tested conditions using the Shannon entropy46, based on the abundance of cells within each of the identified shape mode:

Here, *S* is the Shannon entropy and *pi* is the occurrence of each shape mode.

***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 each individual cell to the centroid of each cluster, where the distance from the cluster centroid is defined as the contour fit (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 improve 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 4GB of RAM running Microsoft Windows 7 or later (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 and/or nuclei** ● **TIMING 10-60 min**

**1|** Segment the fluorescence images to identify the boundaries of cells and/or nuclei. The VAMPIRE GUI does not segment cells. User should accomplish this task with software, including, but not limited to, ImageJ, MATLAB, and CellProfiler. 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](https://cellprofiler.org/tutorials/).

To better demonstrate the VAMPIRE analysis procedure, we provide sample images of fluorescently tagged cells and nuclei in the **Supplementary files** under “Example images” folder. Throughout this **Procedure** section, refer to the directory of supplementary files in **Supplementary information S1** to locate example data and results. We have provided segmented example images using CellProfiler, as well as a sample CellProfiler segmentation pipeline in **Supplementary files**. 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. The segmented images require either 8-bit or 16-bit binary images and image pixels with non-zero integer value representing the detected cells area. (**Figure 5A-B**). If there are more than one type of objects (i.e. cell and nucleus) to be segmented, the segmented images from the different channels can be placed in the same folders but the user should ensure that the filename of segmented images contain distinguishable tags for different types of objects. Users can also refer to the format of example segmented images provided in the **Supplementary files**. The segmented images are already in compatible format with VAMPIRE analysis, and they are located under the folder named “Example segmented image”. This folder contains two subfolders AG04059 and AG04054 which corresponds to the two different fibroblast lines mentioned above.

**? troubleshooting**

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

**3|** Generate a CSV file to specify which segmented image sets to use to construct a VAMPIRE model. In this CSV file, the first row contains column labels. Each column specifies information about the specific segmented images. An example CSV file content can be found in **Supplementary information S2**. From the second row, each column should be filled with information of a specific segmented image set with the following order:

1. “condition” : the specific experimental condition of a sample.
2. “setID” : the name of the folder containing segmented images of a sample.
3. “set location” : the location/path of the folder containing segmented images
4. “note” : any information about the sample for your own record. This is not used in the VAMPIRE analysis.
5. “ch1” : distinguishable tag in the image filename that was created in step 5.
6. “ch2” : distinguishable tag in the image filename for an additional channel

We highly recommend that users download the example CSV files named “images to build model.csv” in the **Supplementary files** and modify the contents using Excel or other CSV editors to generate their own CSV file. To use the example segmented images provided in the **Supplementary files** for the further analysis, the user simply need to update the set location column in the example CSV file with the actual location of the example segmented images on one’s computer.

**4|** Launch VAMPIRE Graphic User Interface (GUI) by openingVAMPIRE.exe file from **Supplementary files**.

**5|** Locate the CSV file generated in **step 6** 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|** Determine the number of shape modes in the “Build Model” section of the VAMPIRE GUI under “number of shape modes” box. The default value is ten. To optimize this number, refer to the ***Selection of parameters for VAMPIRE analysis*** section in the **Introduction**.

**7|** Specify the number of coordinates to extract from the cell/nucleus contours in Build Model section of VAMPIRE GUI under “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 not fully properly represent the actual cell/nuclear morphology.

**8|** Specify the output model name in the “Build Model” section of VAMPIRE GUI under the “Name of the model” box.

**9|** Click “Build Model*”* in VAMPIRE GUI to generate a VAMPIRE model based on the specified parameter values provided in steps 7-10. Once the model is generated, it will be saved to a new folder at the same location as the CSV file generated in step 7. The folder name will be the model name specified in VAMPIRE GUI in step 10. Within this new folder, VAMPIRE model data will be saved into two subfolders “Model for ch1” and “Model for ch2”. Each folder contains:

* A VAMPIRE model file which is named as “[model name]\_[channel tag].pickle”.
* The overlay of randomly selected 20 raw shapes falling 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 is provided in the **Supplementary files**, under “Example model”. These files are generated from the example segmented images provided in **step 5,** using the default values of parameters from **step 8 and 9**.

**? troubleshooting**

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

**10|** Repeat step 6 to specify the sets of segmented images to be analyzed through VAMPIRE. If you need to prepare new sets of segmented images, repeat step 4 to do so. 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.

**11|** Specify the previously built model to analyze the segmented images. Click “load model” button to choose the folder created in step 11. This folder is named by the user in step 10. This folder is equivalent to the “Example model” folder in the **Supplementary files**.

**12|** Perform the VAMPIRE analysis on the specified images by clicking apply “Apply Model*”* in VAMPIRE GUI. When this process is finished, two new folders will be created named “Result for ch1” and “Result for ch2” inside the VAMPIRE model folder. These new folder contain output files for nuclei and cells, respectively. Each folder contains a collection of distributions showing the fractional abundance for cells or nuclei within each shape mode, with the percent of cells or nuclei within each shape mode denoted on the top of the bars. Each distribution is saved with the naming convention: “Shape mode distribution\_ch[#]\_[*condition*].png” (**Supplementary information S4 A**). Clicking the “Apply Model*”* button also generates a registry CSV file in each segmented image set folder. Each registry CSV contains (**Supplementary information S4 B)**:

* 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 major axis and minor axis
* Circularity: shape factor calculated by . 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: number that represents the shape mode where each cell/nucleus belongs to.
* Contour fit: distance between the cluster center and this object

Example output files of this step is provided in the **Supplementary files**, under “Example model”. These files are generated using the example model provided in **step 11**. See the directory of **Supplementary files** in **Supplementary information S3** to locate the output files.

**TROUBLESHOOTING**

table 2| Troubleshooting table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **Problem** | **Possible reason** | **Solution** |
| 1 | Cannot run pipeline: the pipeline did not identify any image sets | 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 “Metadata” module in CellProfiler. |
|  |  |  |  |
| 2 | IndexError: arrays used as indices must be an integer | Segmented images does not contain any cell or nucleus | Check if segmented images have correct format as specified in step 2, and have at least one cell or nucleus. |
| 9 | 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. |
| 9 | Permission denied | CSV file is open while the analysis is running | Close all CSV files open and repeat step 9. |
|  |  |  |  |

**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.

Step 1-3, Install VAMPIRE Graphic User Interface (GUI), 2-5 mins

Step 4, Segment cells and/or nuclei, 10 mins

Step 5, Create a list of images to build the shape-analysis model, 1-3 mins

Steps 6-9, Build shape-analysis model in VAMPIRE, 1-5 mins

Steps 10-13, Application of the model to analyze shapes across conditions, 1-10 mins

Total, steps 1-4, complete VAMPIRE analysis, 15-33 mins

**BOX 1 | GLOSSARY**

**Eigen shapes 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 eigen shapes reconstructs the original shape of each cell.

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

**Shannon entropy**—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* is the Shannon entropy and *pi* is the occurrence of cells in each shape mode.

**Cellular heterogeneity**—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 geometric centroid and the focus point by the distance between the geometric centroid and the vertex.

**Cell contour-fit**— euclidean distance between the 50 equidistant points along the contour of an individual cell/nucleus and those comprising the shape mode.

**PCA**— Abbrevation for principle component analysis. 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.

**ANTICIPATED RESULTS**

To demonstrate the utility of VAMPIRE, we examine 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 glass surfaces: 1. circular, 2. triangular shape fibronectin coated islands, surrounded by poly-ethylene glycol (PEG) passivated regions, and 3. surface without any island nor PEG. Cells were incubated overnight, then fixed and stained with DAPI and Alexa Fluor 488 Phalloidin, highlighting the 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 all probed conditions and examined whether cells on patterns exhibited associations for particular shape modes that resembled circles and triangles (**Figure 6B**). As expected, results showed that both MEF LMNA +/+ and LMNA -/- cells seeded on un-patterned surfaces exhibited a more mixed shape profile i.e., similar aboundance in all identified cellular shape modes, as opposed to the cells seed on the pattened substrates. Cells seeded on circular patterns exhibited an enrichment in the circular shape mode (mode 4) with an average abundance of 55% and 52% of the total cell populations respectively, compared to 8.1% and 21% of those seeded on unpatterned substrate. Cells seeded on triangular patterns the two triangular shape modes, the “sharp” (mode 1) and “blunted” vertex (mode 2) triangles, while decreasing the cell aboundance in irregular cell shapes (mode 6-9) (**Figure 6B**). Interestingly, LMNA -/- cells seeded on triangular patterns were classified as “blunt” (mode 2) three times more with abundance of 34% than “shape” (mode 1) with abundance of 12%. We did not observe such difference between two shape mode abundance 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 connections41,47. 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 the cell shape is more elongated; however, their shape factors increased from 0.34 (no pattern) to 0.51 (triangle pattern), suggesting the cell shape is more round. This self-contradicting result 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 parmaters.

We also examined the association between cellular morphology and chronological ages of dermal fibroblasts with VAMPIRE analysis model from a panel of seven healthy individuals22. Previously, we demonstrated that cell and nuclear morphology of dermal fibroblasts encode key information about the biological age for healthy individuals22. 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 non-rounded morphologies with increasing age. This is measured by a negative age-correlations for shape modes 1 and 2 with rounded shape, and positive age-correlations for non-rounded shape modes 3, 4 and 7 (**Figure 7A**). 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**). VAMPIRE analysis, however, is capable of visually identifying cell morphological changes that otherwise would have been unnoticed.

Boasting the utility beyond cultured cells, we have successfully implemented VAMPIRE analysis based on imaged from in-vitro studies. In this example we analyzed two areas of tissue, namely dermis and epidermisusing hematoxylin and eosin (H&E) stained tissue section (**Figure 8A**). Note that we segmented the H&E stained skin tissue section using custom image analysis algorithm. ~~VAMPIRE analysis GUI does not offer segmentation tool (~~**~~Figure 8B~~**~~)~~. To compare the morphology of cells in dermis and epidermis region, we built VAMPIRE model using nuclei of cells segmented from the scanned image of H&E stained skin tissue biopsy from an 79 years old donor. . Based on the results, we observed that shape modes 1 through 4 are more elongated (i.e. less circular) than modes 5 through 10 (**Figure 8C**). As expected, VAMPIRE analysis shows that nearly 50% of dermal cells are classified as modes 1 through 4, as compared to only 6.4% for epidermal cells (**Figure 8C**).

**ACKNOWLEDGEMENTS.** 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 FINANCIAL INTERESTS.** The authors declare no competing financial interests.

**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. Guo, Q. *et al.* Modulation of keratocyte phenotype by collagen fibril nanoarchitecture in membranes for corneal repair. *Biomaterials* **34**, 9365–9372 (2013).

7. 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).

8. Simm, J. *et al.* Repurposed High-Throughput Images Enable Biological Activity Prediction For Drug Discovery. *doi.org* 108399 (2017). doi:10.1101/108399

9. 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).

10. 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).

11. 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).

12. Beghin, A. *et al.* Localization-based super-resolution imaging meets high-content screening. *Nat. Methods* **14**, 1184–1190 (2017).

13. Carpenter, A. E. Extracting rich information from images. *Methods Mol. Biol.* **486**, 193–211 (2009).

14. 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).

15. 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).

16. Danuser, G. Computer vision in cell biology. *Cell* **147**, 973–978 (2011).

17. Falk, T. *et al.* U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* **16**, 67–70 (2019).

18. Gabril, M. Y. & Yousef, G. M. Informatics for practicing anatomical pathologists: Marking a new era in pathology practice. *Modern Pathology* **23**, 349–358 (2010).

19. Fuchs, T. J. & Buhmann, J. M. Computational pathology: Challenges and promises for tissue analysis. *Computerized Medical Imaging and Graphics* **35**, 515–530 (2011).

20. Sarnecki, J. S. *et al.* A robust nonlinear tissue-component discrimination method for computational pathology. *Lab. Investig.* **96**, 450–458 (2016).

21. Beck, A. H. *et al.* Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. *Sci. Transl. Med.* **3**, 108ra113-108ra113 (2011).

22. Phillip, J. M. *et al.* Biophysical and biomolecular determination of cellular age in humans. *Nat. Biomed. Eng.* **1**, 0093 (2017).

23. Pegoraro, G. & Misteli, T. High-Throughput Imaging for the Discovery of Cellular Mechanisms of Disease. *Trends in Genetics* **33**, 604–615 (2017).

24. Lang, P., Yeow, K., Nichols, A. & Scheer, A. Cellular imaging in drug discovery. *Nature Reviews Drug Discovery* **5**, 343–356 (2006).

25. 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).

26. Sailem, H. Z., Sero, J. E. & Bakal, C. Visualizing cellular imaging data using PhenoPlot. *Nat Commun* **6**, 1–6 (2015).

27. McQuin, C. *et al.* CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.* (2018). doi:10.1371/journal.pbio.2005970

28. Schindelin, J. *et al.* Fiji: An open-source platform for biological-image analysis. *Nature Methods* (2012). doi:10.1038/nmeth.2019

29. Hodneland, E., Kögel, T., Frei, D. M., Gerdes, H. H. & Lundervold, A. CellSegm - a MATLAB toolbox for high-throughput 3D cell segmentation. *Source Code Biol. Med.* **8**, (2013).

30. 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

31. 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).

32. 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

33. Jayatilaka, H. *et al.* Synergistic IL-6 and IL-8 paracrine signalling pathway infers a strategy to inhibit tumour cell migration. *Nat. Commun.* (2017). doi:10.1038/ncomms15584

34. Jayatilaka, H. *et al.* Tumor cell density regulates matrix metalloproteinases for enhanced migration. *Oncotarget* **9**, 32556–32569 (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**, (2017).

36. Phillip, J. M., Aifuwa, I., Walston, J. & Wirtz, D. The Mechanobiology of Aging. *Annu. Rev. Biomed. Eng.* **17**, 113–141 (2015).

37. Kim, D.-H. *et al.* Volume regulation and shape bifurcation in the cell nucleus. *J. Cell Sci.* **129**, 457–457 (2016).

38. 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).

39. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* (2011).

40. Kim, D. H. & Wirtz, D. Focal adhesion size uniquely predicts cell migration. *FASEB J.* **27**, 1351–1361 (2013).

41. Kim, J.-K. *et al.* Nuclear lamin A/C harnesses the perinuclear apical actin cables to protect nuclear morphology. *Nat. Commun.* **8**, 2123 (2017).

42. 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

43. 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

44. 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

45. Legland, D., Arganda-Carreras, I. & Andrey, P. MorphoLibJ: Integrated library and plugins for mathematical morphology with ImageJ. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw413

46. Altschuler, S. J. & Wu, L. F. Cellular Heterogeneity: Do Differences Make a Difference? *Cell* **141**, 559–563 (2010).

47. Kim, D. H. & Wirtz, D. Cytoskeletal tension induces the polarized architecture of the nucleus. *Biomaterials* **48**, 161–172 (2015).

**FIGURE CAPTIONS**

**Figure 1. Cells confined to narrow ranges of traditional morphological parameters still exhibit highly variable shapes.** Each point represents an individual mouse embryonic fibroblast cell (MEF) seeded on a collagen-coated substrate. Cells were fixed and stained for F-actin and nuclear DNA to delineate cell and nuclear boundaries, respectively. Stained samples were mounted onto an epifluorescence microscope and imaged at 10x magnification. The subset of 10 cells highlighted in red were selected from the pool of 37,750 cells within 1% of each other based on their aspect ratio, circularity and solidity. Despite the fact that these cells feature near-identical values for these traditional morphological parameters, they still show large morphological heterogeneity.

**Figure 2. VAMPIRE workflow, from extraction of contour coordinates to the automatic generation of shape modes. A.** The contour of a SUM149 human breast cancer cell described by 50 equidistant points along its contour. **B.** Unaligned (left) shapes of a subset of SUM149 breast cancer cells are pooled, normalized by size, and aligned (right). **C**. Eigen-shape vectors are constructed from a Principal Component Analysis (PCA) of cells. Eigen-shape vector of first principal component (PC) follows the longest dimension, thus explains the largest variance. Second PC fits the error produced by the first PC, and so forth. **D**. Reconstructed eigen-shape vectors of cells. Here, the number of principle components of cells was set to 20, and reconstructed from shape vectors that comprise 95% of the shape variations across all assessed cells. **E**. Dendrogram of representative cellular shape modes.

**Figure 3. Key steps of VAMPIRE analysis and VAMPIRE Graphic User Interface (GUI). A.** Flow diagram illustrating steps involved in image preprocessing and the generation of shape mode distributions. Step 1 (green) combines individual segmented images into VAMPIRE compatible labeled grayscale image. The VAMPIRE GUI does not provide this step. Users could use CellProfiler to accomplish step 1 and generate a datasheet of individual segmented objects using CellProfiler. This datasheet can be directly compared to the one generated from VAMPIRE analysis. Step 2 (red) highlights the steps involved in building the VAMPIRE model, which includes the selection of the image training set, the number of shape modes, and the number of coordinate-points along the cell/nuclear contours to define their shapes, and the name of the model. The Experimental design section of this protocol explains how to set these parameters. Step 3 (blue) highlights the steps involved in applying the model to other image sets. This involves selecting the image sets and the model to be used for analysis. **B**. VAMPIRE Graphic User Interface (GUI) with key steps delineated in the flowchart (red and 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 a measure of cluster coherence. **B**. The inertia decays in an elbow shape for an increasing number of clusters. Data used to generate the elbow decay plots were computed based on images of mouse embryonic fibroblasts (MEFs) confined to circular and triangular micropatterns, and unconfined MEFs. Here, the inertia does not decrease significantly beyond the cluster number of 10. **D**. The effect of eliminating cells/nuclei based on their distance from the cluster centroid. Pearson correlation between original shape mode distribution and modified shape mode distributions. Cells and nuclei were eliminated from the original shape mode distribution based on their distance from the nuclei, decreasing the Pearson correlation coefficient. It reaches a plateau as we remove the outliers. Error bars indicate the standard error of the mean.

**Figure 5**. **Binarization of raw images and generation of shape mode distributions for aging dermal fibroblasts. A.** Greyscale and binarized images of the H33342-stained nuclei (left) and phalloidin-stained F-Actin networks (right) of dermal fibroblasts derived from a 29-year (AG04054) and a 96-year old volunteers (AG04059). **B**. Distribution of cell and nuclei shape modes comparing dermal fibroblasts from a 29-year (AG04054) and a 96-year old volunteers (AG04059).

**Figure 6. VAMPIRE analysis of mouse embryonic fibroblasts seeded on adhesive micro-patterned surfaces. A.** Fluorescence microscope images of wild-type (MEF++) and lamin-deficient (MEF--) mouse embryonic fibroblasts cultured on circular (top row) and triangular (middle row) adhesive fibronectin-coated micropatterns. Control cells (bottom row) are placed on 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**. Shape mode distributions for MEF++ and MEF-- 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, and aspect ratio of cells, as well as the number of cells analyzed (N), lamin A/C status and the Shannon entropy of the cells. These results show that traditional morphological parameters cannot discriminate between the nuclear morphological responses of MEF++ and MEF-- 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).

**Figure 7. VAMPIRE analysis of human dermal fibroblasts from donors of different ages. A.** Distribution of nuclear shape modes for dermal fibroblasts from age 3 to 96. Cells from younger donors populate the rounder shape mode (mode 1 and 2), while cells from older donors have nuclei classified that populate the non-rounded shape modes (mode 3, 4, and 7). **B.** Shape factor (SF) and aspect ratio (AR) of each nuclear shape modes. SF and AR in red indicate shape modes that have negative correlations with chronological age, while those in blue indicate positive correlations.

**Figure 8. VAMPIRE analysis of epidermis and dermis from H&E stained skin tissue section.**

**A.** Scanned image of Hematoxylin and Eosin (H&E) stained skin tissue section. Skin tissue section was prepared from a biopsy of a 79 years old white male patient diagnosed with melanoma (TCGA case ID: EE-A20I). Dermis and epidermis are highlighted green and blue, respectively. **B.** Two different regions of H&E stained skin tissue: dermis and epidermis have been segmented to detect nucleus boundary. **C.** Distribution of nuclei shape modes comparing epidermal cells (N=1579) and dermal cells (N=498). Numbers above the bars represent the abundances [%] of nuclei in each shape modes.