Integrative Biology

Morphology-based Prediction of Cancer Cell Migration Using Artificial Neural Network and Random Decision Forest

Journal:	*Integrative Biology*
Manuscript ID	IB-ART-06-2018-000106.R1
Article Type:	Paper
Date Submitted by the Author:	14-Oct-2018
Complete List of Authors:	Zhang, Zhixiong; University of Michigan, Electrical and Computer Engineering Chen, Lili; University of Michigan, Electrical and Computer Engineering Humphries, Brock; Center for Molecular Imaging, Department of Radiology, University of Michigan Brien, Riley; University of Michigan, Electrical and Computer Engineering Wicha, Max; Comprehensive Cancer Center, University of Michigan; Forbes Institute for Cancer Discovery, University of Michigan Luker, Kathryn; University of Michigan, Radiology Luker, Gary; University of Michigan, Depeartment of Radiology; Comprehensive Cancer Center, University of Michigan; Department of Biomedical Engineering, University of Michigan; Department of Microbiology and Immunology, University of Michigan Chen, Yu-Chih; University of Michigan, Electrical and Computer Engineering; Comprehensive Cancer Center, University of Michigan; Forbes Institute for Cancer Discovery, University of Michigan Yoon, Euisik; University of Michigan, EECS; Department of Biomedical Engineering, University of Michigan

Integrative Biology

Insight Box

Cancer cells with enhanced motility and invasiveness migrate away from primary tumor site and initiate metastatic process. Identifying key aspects for cell migration is crucial for understanding and ultimately overcoming metastasis. Considerable efforts have focused on discovering markers for epithelial-to-mesenchymal transition (EMT), in which epithelial cells acquire migratory and invasive phenotypes to promote metastasis, yet marker-based approaches are limited by inconsistences among patients, cancer types, and partial EMT states. The recent developments of computer vision and deep learning provide a potent alternative to define cell properties based on morphology. In this work, we present a comprehensive morphological analysis using deep learning methods to establish the correlation between cellular morphology and migration behavior.

Integrative Biology

ARTICLE

Morphology-based Prediction of Cancer Cell Migration Using Artificial Neural Network and Random Decision Forest

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Zhixiong Zhang[†1], Lili Chen[†1], Brock Humphries[2], Riley Brien[1], Max S. Wicha[3, 4], Kathryn E. Luker[2], Gary D. Luker[2, 3, 5, 6], Yu-Chih Chen[*1, 3, 4], and Euisik Yoon[*1, 6]

Metastasis is the cause of death in most patients of breast cancer and other solid malignancies. Identification of cancer cells with highly migratory capability to metastasize relies on markers for epithelial-to-mesenchymal transition (EMT), a process elevating cell migration and metastasis. Marker-based approaches are limited by inconsistences among patients, types of cancer, and partial EMT states. Alternatively, we analyzed cancer cell migration behavior using computer vision. Using microfluidic single-cell migration chip and high-content imaging, we extracted morphological features and recorded migratory direction and speed of breast cancer cells. By applying Random Decision Forest (RDF) and Artificial Neural Network (ANN), we achieved over 99% accuracy for cell movement direction prediction and 91% for speed prediction. Unprecedentedly, we identified highly motile cells and non-motile cells based on microscope images and machine learning model, and pinpointed and validated morphological features determining cell migration, including not only known features related to cell polarization but also novel ones that can drive future mechanistic studies. Predicting cell movement by computer vision and machine learning establishes a ground-breaking approach to analyze cell migration and metastasis.

Introduction

Metastasis is the leading cause of mortality in patients with breast cancer, being responsible for over 40,000 deaths per year in the US. Despite advances in early detection and treatment, once metastases develop, breast cancer is incurable[1, 2]. Cancer cells with enhanced motility and invasiveness migrate away from the primary tumor site and initiate the metastatic process[1]. Therefore, identifying key aspects for cell migration is crucial for understanding and ultimately overcoming metastasis. Currently, considerable efforts have focused on elucidating mechanisms that govern epithelial-to-mesenchymal-transition (EMT), a developmental program in which epithelial cells acquire migratory and invasive phenotypes to promote metastasis. In recent decades, various EMT biomarkers including membrane proteins (e.g. E-CAD, N-CAD), cytoskeletal markers (e.g. Vimentin, Cytokeratins), transcriptional factors (e.g. Snail, Slug, ZEB1, ZEB2, Twist) were developed3-5. However, these and other markers for defining EMT underscore problems of marker-based approaches across multiple cancers: 1) cancer cells undergo differing extents of partial EMT; 2) multiple sets of markers have been used to define EMT even within a single type of cancer; 3) markers are inconsistent across different malignancies[3]. Inconsistencies of existing EMT markers highlight the need for new approaches to identify highly migratory cells[4, 5].

Not only does the recent development of Artificial Intelligence (AI) and computer vision provide a potent alternative to define cell properties based on morphology, but also use of fluorescent probes and reporters to label proteins, protein activity, and organelles has advanced our ability to study mitochondria. Mitochondrial morphology correlates with metabolic state, drug response, and cell viability, providing potential insights into overall status and function of cells[6-8]. Advances in computer technology now allow high-content images of mitochondria to be processed by the computer vision program[9,10]. After training on data sets, the computer vision software can autonomously interpret meanings of images and classify cells based on imaging features. Various algorithms such as Random Decision Forests[11] (RDFs construct decision trees in training and make decisions based on voting of trees) and Artificial Neural Networks (ANNs build a group of nodes interconnected with weighted linkage in training and

1 Department of Electrical Engineering and Computer Science, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109-2122;
2 Center for Molecular Imaging, Department of Radiology, University of Michigan, 109 Zina Pitcher Place, Ann Arbor, MI 48109-2200, USA;
3 Comprehensive Cancer Center, University of Michigan, 1500 E. Medical Center Drive, Ann Arbor, MI 48109, USA;
4 Forbes Institute for Cancer Discovery, University of Michigan, 2800 Plymouth Rd., Ann Arbor, MI 48109, USA;
5 Department of Microbiology and Immunology, University of Michigan, 109 Zina Pitcher Place, Ann Arbor, MI 48109-2200, USA;
6 Department of Biomedical Engineering, University of Michigan, 2200 Bonisteel, Blvd. Ann Arbor, MI 48109-2099, USA
† These authors contributed equally to this work.
* Corresponding author
Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See

classify things accordingly)[12] were developed. However, people so far have only analysed single imaging features using small numbers of cells to investigate correlations between the distribution of mitochondria and cell movement[13]. Cutting-edge computer vision techniques were not used to fully explore the potency of morphological features in determining cell migration direction and speed.

In addition to imaging analysis capability, an effective cell monitoring scheme is also critical to the success of comprehensive cell morphological analysis. Microfluidic technology has emerged as a state-of-the-art approach for cell biology because of precise manipulation of single cells and high potential in scaling[14-16]. As compared to tracking cells randomly seeded in a dish, cells in a microfluidic chip are precisely positioned and easily tracked in a high-throughput manner. Thus, the migration distance of individual cells can be accurately measured to correlate with its morphology. More importantly, chemoattractant gradients can be generated on-chip to model chemotaxis in cancer metastasis. Hence, we applied the high-throughput cell migration chip we have previously developed for this study[17].

In this work, we present a comprehensive morphological analysis using cutting-edge computer vision methods including random decision forests and artificial neural networks to establish the correlation between cellular morphological features and cell movement direction and speed. We first collected 1,358 cellular and mitochondrial images and then trained and optimized the machine learning model. Using the built model, we successfully predicted the migration direction for more than 99% of cells and picked out highly-motile cells (top 10% fast-moving cells) and non-motile cells (top 10% slow-moving cells) with 91% accuracy. Based on the prediction, we identified critical morphological markers determining cell movement direction and speed. To validate the importance of markers we found, we impaired cell movement using commonly used chemotherapeutics as well as sorted highly migratory cells from the bulk population for comparison. Both experiments validated the importance of identified morphological features in determining cell movement. The presented work represents a new method to predict and understand the cell migration process, which will advance studies of mechanisms driving cell migration.

Materials and methods

Microfluidic Migration Chip Design and Fabrication

The migration devices were fabricated from a single layer of PDMS (Polydimethlysiloxane, Sylgard 184, Dow Corning), which was fabricated on a silicon substrate by standard soft lithography, and a glass slide. Two masks were used to fabricate the multiple heights for main channel (40 μm height) and the migration channel (5 μm height). One device contains 900 migration channels (450 channels in one side), and the migration channel is 30 μm in width, 5 μm height, and 1 mm in length. The PDMS layer was bonded to the glass slide after activated by oxygen plasma treatment (80 Watts, 60 seconds) to form a complete fluidic channel (Fig. S1(a)). The microfluidic chips were sanitized by UV radiation prior to use to ensure aseptic conditions. Before cell loading, a collagen (Collagen Type 1, 354236, BD Biosciences) solution (1.45 mL Collagen, 0.1 mL acetic acid in 50 mL DI water) was flowed

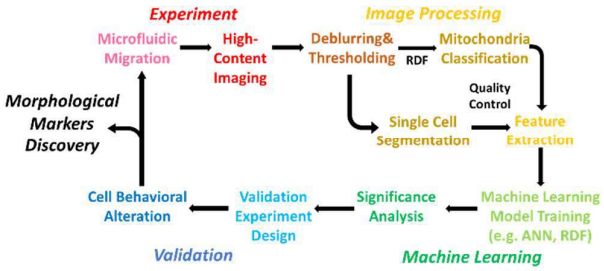

Fig. 1. Workflow of critical morphological features discovery in cell migration, which includes microfluidic migration chip experiments, high-content imaging, image processing, machine learning modelling, and control experiment validation.

through the device for ten hours to coat collagen on the substrate to enhance cell adhesion. Devices were then rinsed with PBS (Gibco 10082) for five minutes to remove the residual collagen solution. Culture media was used to rinse devices before cell loading.

Cell Culture

SUM159 cells were authenticated by short tandem repeat profiling performed by the University of Michigan DNA Sequencing Resource. The cells were cultured in the F-12 based media (Ham's F12, Gibco 11765) supplemented with 1% penicillin/streptomycin (Gibco 15070), 5 μg/mL (2.5 mg/500 mL) insulin (Sigma I6634), 1 μg/ mL (0.5 mg/500 mL) hydrocortisone (Sigma H4001), and 5% FBS (Gibco 10082). We maintained all cells in a humidified incubator with 5% CO2.

Transduction of Stable Cell Line

We transduced cells with a lentiviral vector expressing cytochrome C oxidase 8 (COX8) fused to GFP to visualize mitochondria (plasmid from Addgene)[28]. To generate a lentiviral vector for constitutive expression of mCherry, we amplified mCherry from pmCherry-N1 (Takara) with PCR primers 5'-ATGCTCTAGAGCAGAGCTGGTTTAG-3' and 5-ATGTGGTATGGCTGATTATGATC-3'. We cloned the PCR product into pLVX puro (Takara) cut with XbaI (New England Biolabs) and then confirmed the final construct by DNA sequencing. We prepared lentiviruses in 293T cells as described previously[29].

Cell Migration Assay

Cells were harvested from culture plates with 0.05% Trypsin/EDTA (Gibco 25200) and centrifuged at 1000 rpm for 5 minutes. Then, the cells were re-suspended in culture media to a concentration of 3 x 10^5 cells/mL. One hundred microliters (100 μL) of this cell suspension was pipetted into the lower

inlets. After 3 minutes, cell solution in the left and right inlets was replaced with 50 µL of serum culture media, and 40 µL serum culture media was applied to the central inlet for both high and low sides. After 30 minutes, media in all inlets were emptied out and replaced by 200 µL serum-free culture media (for the high-left and high-right inlets), and 200 µL serum culture media (for the high-central inlet) to induce chemotactic migration. Then, the entire chip was put into a cell culture incubator for 5 hours to prepare for image acquisition.

Image Acquisition

The microfluidic chips were imaged using an inverted microscope (Nikon). The fluorescent images were taken with a 40X objectives (NA = 0.75, WD = 0.66 mm, resolution = 0.17 µm) and a charge-coupled device (CCD) camera (Coolsnap HQ2, Photometrics). A FITC (for GFP mitochondria)/TRITC (for mCherry cell) filter set was used for the fluorescent imaging of mitochondria of SUM159 cells. The fluorescent imaging was performed using an exposure time of 200 ms, minimizing the phototoxic effect on cells. The microfluidic cell chamber was scanned with a motorized stage (ProScan II, Prior Scientific). Before each scanning, the stage was levelled to ensure the image remained in the focus throughout the whole imaging area.

Image Processing Program

A customized MATLAB program was used to process raw images under FITC channel collected from the microscope. Each pixel had a value ranged from 0 to 255 indicating its brightness. Auto-fluorescence of PDMS and signal caused by camera dark current were removed from raw images by setting one brightness value (from 0 to 255) which had the greatest number of pixels as threshold. Then, a 10-pixel-large spike noise filter was used to remove tiny dazzling dots from pre-processed images. After that, erosion followed by a dilation were applied to these images which was fulfilled by "imopen" in MATLAB. Before converting these images to black-white images, a 5-pixels by 5-pixel Wiener filter was used for image deblurring. Also, a 15-pixel-large area filter was implemented to remove small dots caused by image deblurring. Processed images often contained more than one cell. Our customized MATLAB program aimed to cut a bounding box which contained the leftmost cell only. For each processed image, nucleus center was labelled manually with the help of MATLAB (refer to Nucleus Center Validation). With this nucleus, the corresponding single cell under TRITC channel could be retrieved by choosing the connected domain which contained the point of nucleus.

Definition of Cell Speed

Images for cells at the same location were taken with a roughly 10-minute period. Movement for a single cell was defined as the movement of center of mass of images under FITC channel. In order to allow the cell speed to be more accurate, imaging time was recorded for each image. With this information, cell speed could be calculated by dividing movement of a single cell by difference of imaging time.

Image Quality Control and Image Selections

Image selection took three issues into account: image quality, chemo-attraction correctness, and two cells in the same channel. Imaging out-of-focus or very thick cells could lead to the problem of low image quality. These low-quality images could not provide us with enough information about mitochondria. To ensure image quality, image quality control was introduced and fulfilled by mitochondrial classification. In our experiment, mitochondria were classified into three different classes based on their shape: fiber, intermediate, and dot. The out-of-focus images (fiber ratio lower than 11%) were discarded to guarantee good image quality for following analysis (Fig. S2). Chemo-attraction correctness was defined as cells moving towards the central channel due to chemo-attraction. Chemo-attraction correctness prevailed in all SUM159 cells. Around 10% of cells might move in the opposite direction. In our experiment, we only took normal chemotactic cells into consideration. Images with two cells in the same channel were also discarded because the interaction between two cells could affect their individual movement significantly.

Feature Extraction and Data Pre-processing

With the information of processed images from FITC/TRITC channel, we extracted 61 features for all 1358 single cells in our database. Definitions for these features were included in the Supplementary Information and feature extraction was done with the help of MATLAB R2017a. For direction prediction, these 61 features were taken with their original value. For speed prediction, features with name starting like "TopDownxxx", CenterShift, FiberUpDownRatio, MaxWidth, MaxWidthSum, TotalAreaRatio, TotalPerimeterRatio, TotalPerimeterHalfRatio, HeadAverageWidth, HeadAverageWidthRatio, RedShift, RedFoot, RedGreenDist, and TotalAreaHalfRatio were reconsidered to make cell moving direction no longer important (e.g. CenterShift values -1 (for up-moving cells) and 1 (for down-moving cells) were different for direction prediction but the same for speed prediction. Therefore both values should be taken as 1) in speed prediction.

After data normalization, whitening transformation was applied for feature decorrelation, which makes the covariance matrix of feature space to an identity matrix. By deploying Zero-phase Component (ZCA) whitening method with a ZCA constant of 0.0001, the correlation coefficient between most of the features are reduced to below 0.1 (Fig. S3). Wrapper method feature reduction was implemented to minimize the influence of unrelated or redundant features. These 61 features were first normalized (zero mean and unit variance), and then took turns to be all zeros for one feature. For each arrangement, an average error rate was calculated from 50 predictions using our Artificial Neural Network (ANN). In all 61

error rates, the feature with the lowest error rate was deleted. Features were deleted one by one until a new deletion would visibly increase the prediction error rate.

Parameters for Random Decision Forest and Artificial Neural Network

Bootstrapped-aggregated decision trees were constructed based on subsets of the training data set and this could reduce the variance significantly. We took the mode of all outputs from trees for classification and took the average for regression1. In MATLAB, we used the function "TreeBagger" to simulate the growth of random decision forest. For cell direction prediction and cell speed prediction, 500 trees were grown to let the error rate become stable. Artificial neural network was also chosen for cell direction and speed predictions because of its nonlinear characteristic. Every single feature in our raw data were set to zero average and unit variance before being inputted into our model. We chose a two-hidden-layer pattern net as our model and by going through all possible combinations of hidden node numbers for two hidden layers, the best result was achieved when the first hidden layer had 21 nodes and the second had 7 nodes. In our model, 70% of our data were used for training, 20% for validation and 10% for test. With the aim of distinguishing highly-motile cells against non-motile cells, we randomly pick out 134 cells (10% of the total dataset) from top 10% fast-moving cells and top 10% slow-moving cells as test set, while the rest of the cells were split to training set and validation set. Hyperbolic tangent sigmoid transfer function ("tansig" in MATLAB) was used as the activation function for two hidden layers and linear transfer function ("purelin" in MATLAB) was used as the activation function for the output layer. Scaled conjugate gradient backpropagation ("trainscg" in MATLAB) was used as our training function with the intention to reduce the mean square error.

Drug Treatment Experiment

Doxorubicin hydrochloride (CAS 25316-40-9) was obtained from Cayman Chemical (Cat. No. 15007), and was dissolved in SUM159 cell culture media to a final concentration of 0.5 μM (sub-IC50 concentration). To allow cancer cells some time to respond to the drug, we performed an on-chip drug pre-treatment before introducing the chemotactic gradient. Following the cell loading protocol described previously, 200 μL of the doxorubicin solution with no serum was loaded to the migration device 30 minutes after cell loading. After 6 hours of pre-treatment, all inlets and outlets were emptied out. Chemoattraction was introduced by filling the high-left and high-right inlets with 200 μL doxorubicin solution in serum-free culture media, while filling the central inlets with 200 μL doxorubicin solution in serum culture media. Then, the entire chip was put into a cell culture incubator for 5 hours to prepare for image acquisition.

Isolation of Highly-Migratory Cells

For recovery of highly-migratory cells from our microfluidics device, 24 hours after cell loading onto migration chip, we pipetted 200 μL of PBS to the inlets to wash the microfluidic channels for 5 minutes. After removing PBS (Gibco 10010) in the inlets, we pipetted 200 μL of trypsin (Gibco 25200) to the inlets to trypsinize cells and place the device in incubator for 5 minutes. We then collected trypsinized cells in the center outlet. The collected cells were cultured in dish for 2 weeks. After validation cell migration speed to confirm elevated migration using microfluidic migration assay, the cell population was used for this study as highly migratory cells.

Nucleus Center Validation

To quantify the error in manual nucleus center labelling, which helps in dividing each single cell into two halves, we compared our manual labelling location with the result given by an automatic nucleus center locating MATLAB program (Fig. S4). This nucleus center locating program calculated the center of mass based on a fluorescent image of cell nucleus. To visualize the cell nucleus, SUM159 cells were treated with 4',6-diamidino-2-phenylindole, dihydrochloride (DAPI) (Invitrogen, D-21490) in a 1:1000 dilution in PBS for 15 minutes, followed by PBS washing for 3 times. Microscope images were acquired immediately after washing. As a result of randomly picked 48 cells, the mean distance error between manual labelling and computing labelling was around 6 pixels (0.95 μm).

Results

Morphological Prediction of Cell Migration Pipeline

In order to discover and validate morphological features that contribute to migration, we developed a workflow including 4 steps: experiment, image processing, machine learning, and validation (Fig. 1). Firstly, the microfluidic migration chip provided single-cell resolution for cellular morphological analysis. After 6 hours of incubation time, fluorescence microscopy was used to obtain 40X high-content images of both the mitochondria and cell profile. After implementing image pre-processing procedures Based on pre-processed images, we implemented single-cell segmentation and a Random Decision Forest (RDF) classifier for mitochondrial classification, in which all the pieces of mitochondria in each single cell were sorted into three categories: fiber, intermediate, and dots6, with important morphology features including major axis, area, and aspect ratio. The distribution of mitochondrial types was then applied as one of the 61 extracted cellular morphological features (Supplementary information). Using these features from both cellular morphology and mitochondria profiles as inputs, we trained two machine learning models to predict cell migration direction and cell migration speed. The Random Decision Forest (RDF) and Artificial Neural Network (ANN) achieved 99.6% and 91.0% accuracy, respectively. More importantly, we

were also able to pinpoint significant morphological features critical to migration behavior based on the predictor importance analysis from the RDF classifier. By performing control experiments, we validated that our discovered morphological markers are highly correlated with cancer cell migration.

Imaging Processing and Mitochondrial Classification

As the "powerhouse" of eukaryotic cells[18], mitochondria are important in energy demanding behaviors. This includes cell migration, which requires cellular polarization, reorganization of actin filaments, and recruitment of structural and signaling components[19]. Studies have shown that perturbations to mitochondria dynamics (i.e. fragmentation and fusion) may affect cell development, cell cycle or cell signaling[20]. However, those studies were mostly carried out by means of subjective observation or qualitative explanation, instead of objective, quantitative analysis. To investigate mitochondria in migration more in-depth, we segmented each mitochondrion after a series of image pre-processing steps, including background noise removal, contrast enhancement, deblurring, and histogram-based auto-thresholding (Fig. 2(a-f)). We trained a RDF classifier to automatically categorize each mitochondrion into three types: dots, which represent fragmented mitochondria; fiber, which includes interconnected networks and elongated mitochondrial fibers; or intermediate, which defines mitochondria whose length is in between dots and fibers. By using 1000 manually labelled mitochondria as training

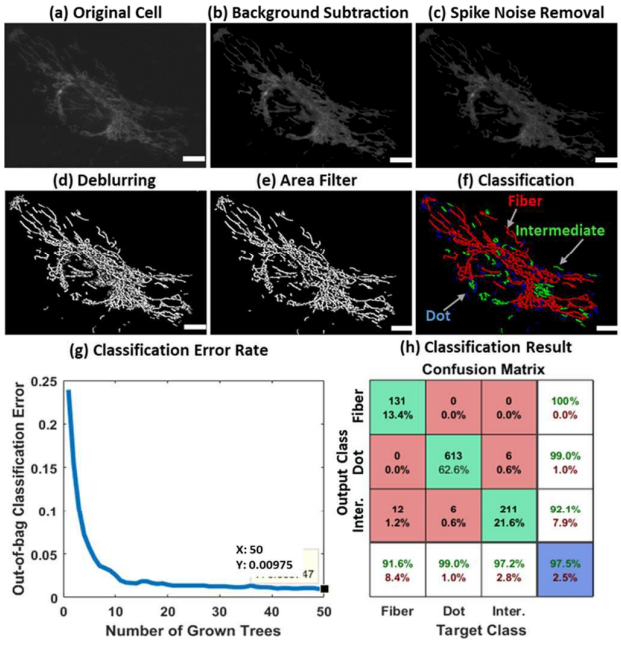

Fig. 2. Image processing flow for the original microscope image (a-f) and mitochondrial classification (g, h). (a) The original image of a SUM159 cell was taken with a 40x objective lens. (b) Background of the original image was removed to have a clear view of the cell. (c) Spike noise was removed from the image. (d) A 5-pixel by 5-pixel Wiener filter was applied for deblurring purposes. (e) A 15-pixel large area filter was applied for sharpening and thresholding. (f) All mitochondria of a cell were classified into three classes: fiber, intermediate and dot based on a Random Decision Forest (RDF) classifier. (scale bar: 10 µm). (g-h) Results of mitochondrial classification. (g) Out-of-bag classification error decreased with the increased number of grown trees using the RDF model. (h) Classification results shown by confusion matrix stated that the correction rate was above 97%, with no misclassification between fiber and dot class.

sets and 200 as test sets, we achieved 97.5% overall accuracy in mitochondrial classification (Fig. 2(g, h)). As an average of the result, 61.1% of the total area of mitochondria are classified as fiber, 20.5% as intermediate, while 18.4% as dots. This mitochondrial class distribution is skewed to 71.1% for fiber, 13.2% for intermediate, and 15.7% for dot, when exposed to a chemotherapeutic drug, and to 49.1% for fiber, 28.3% for intermediate, and 22.6% for dot when cells are classified as highly migratory from our microfluidic device.

Cell Migration Direction Prediction

In order to accurately quantify the moving speed of individual cancer cells, we loaded the cell suspension into a microfluidic migration chip, which consisted of 2 × 450 individual narrow migration channels divided into an upper half and lower half. A total of three inlets and three outlets are deployed on the chip as the loading interface (Fig. S1). Cancer cells were loaded to the inlets on both upper and lower sides and allowed to migrate towards the center with serum as the chemoattractant. The migration channels were designed to be 1000 µm × 30 µm × 5 µm (L × W × H). The dimension of the cross-section is small enough that only single cells could be positioned in each channel, while the migrating direction was also confined to the orientation of the migration channels. The study of migration direction, which involves the process of cellular reaction to mechanical/chemical cue, cytoskeleton polarization, and signalling dynamics, provides important clues to our understanding of underlying mechanisms of metastasis[21,22]. Therefore, we developed a machine learning model to predict and explain cell migration merely based on the morphological features of cancer cells.

Random Decision Forest (RDF) is a commonly used classifier, as well as a regression tool, which constructs a binary decision tree by asking a sequence of simple questions to inputs and assign a label to each condition[23]. To overcome an overfitting issue that is caused by using an over-complicated tree structure, a bootstrap-aggregated decision tree technique is often adopted for better model performance because it decreases the number of variables of the model and combines the results of multiple decision trees[24]. We first implemented RDF for cell migration direction classification. Due to confinement of the migration channels, all the cancer cells were only allowed to move bidirectionally and were

labeled according to the movement of the center of mass measurement by computer program. With the randomly scrambled cell image inputs, we achieved more than 99% accuracy in prediction of cell migration direction (Fig. 3(a)). We also performed a feature importance analysis by summing the estimates of all weak learners in the bagged decision trees. Based on the importance analysis, we validated that features reflecting cellular polarization are essential in deciding the migration direction (Fig, 3(b) and Fig. 4). These features include: TotalAreaRatio, which defines the ratio of the areas between the upper and lower half of the cell, CenterShift, which defines of the deviation of the center of mass

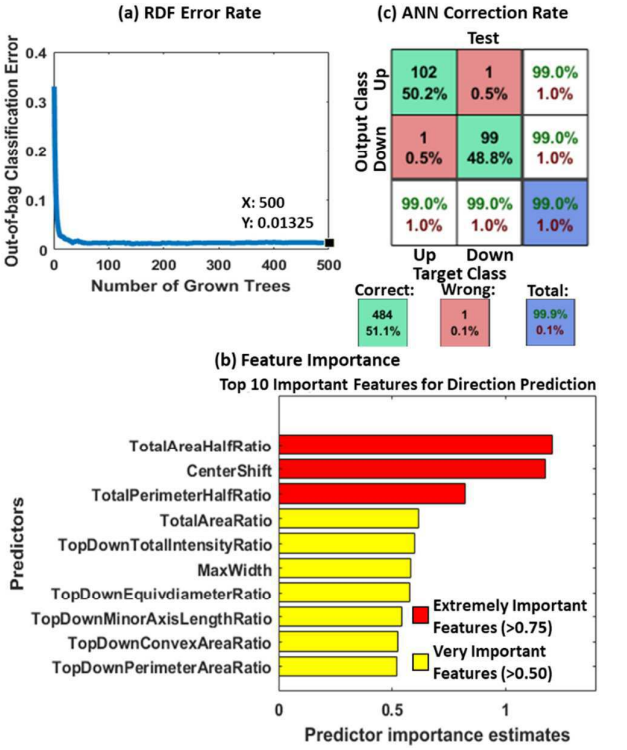

Fig. 3. Results and important features for cell migration direction prediction. (a) Out-of-bag error rate for cell migration direction prediction using RDF. With the increase in number of grown trees, the error rate reduces to less than 1%. (b) Confusion matrices for testing datasets. Accuracy for cell migration direction prediction is above 99%. (c) Top 10 important features for cell migration direction prediction.

from the graphic center, as well as TotalPerimeterRatio, which defines the ratio of the perimeters between upper and lower half of the cell. Although most of the critical morphological features are about cell polarization, each feature conveys its own unique information. For example, CenterShift indicates that the nucleus is more likely to appear in the rear portion of the cell, with a protrusion stretched to the front; TotalPerimeterRatio represents not only the effect from nucleus center shift, but also the border length of the cell frontier (i.e. cells with filopodia-like protrusions tend to have larger perimeters).

Although RDF is straightforward to handle and advantageous in interpreting feature importance, it sometimes will only give suboptimal solutions due to the nature of greedy growing algorithm, as well as unstable to even slight perturbations of the training data. Artificial Neural Network (ANN) is a nonlinear model for universal function approximation. With enough data sets, ANN is more likely to provide better prediction power. Therefore, we further explored ANN for predicting cell migration direction. Based on a database of 1,358 single-cell images collected using the presented method, we trained a four-layer-ANN model which achieves an overall accuracy of 99.6% combining training, validation, and test data (Fig. 3(c)).

Cell Motility (Migration Speed) Prediction

In addition to direction, we further explored the capability of our machine learning model in predicting motility or migration speed. This will provide insights for the discovery of critical markers determining cancer metastasis. Using the same workflow as described previously, we first applied RDF to pinpoint the important features affecting cell migration speed (Fig. 5(a, b)), and further enhanced the prediction power using

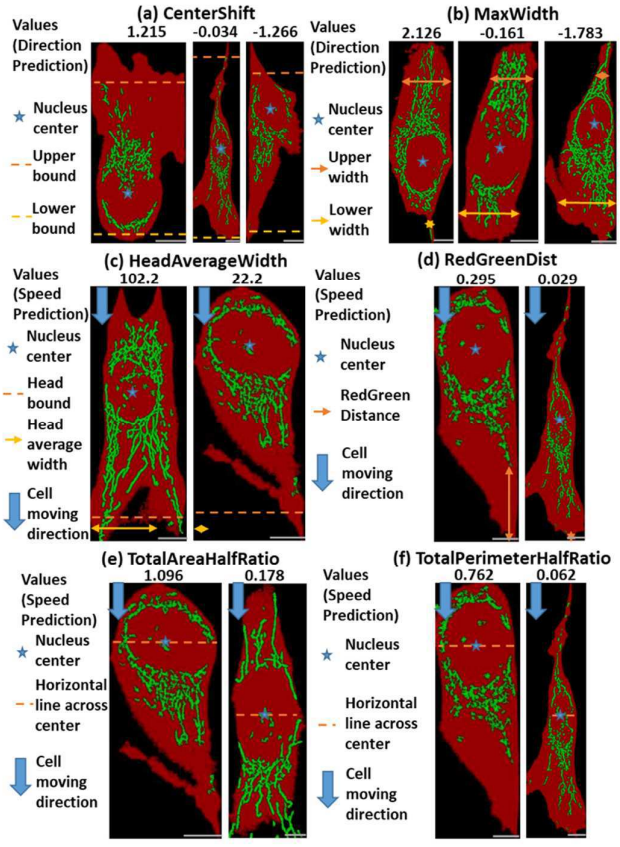

Fig. 4. Selected important features for cell migration direction/speed prediction. (a) For direction prediction, up-moving cells tend to have positive CenterShift values and down-moving ones tend to have negative CenterShift values. For speed prediction, large CenterShift values typically indicated fast-moving cells. (b) For direction prediction,

up-moving cells often had positive MaxWidth values and down-moving ones often had negative MaxWidth values. For speed prediction, large MaxWidth values typically indicated fast-moving cells. (c) For speed prediction, large HeadAverageWidth values typically indicated fast-moving cells. (d) For speed prediction, large RedGreenDist values typically indicated fast-moving cells. (e) For direction prediction, up-moving cells often had positive TotalAreaHalfRatio values and down-moving ones often had negative TotalAreaHalfRatio values. For speed prediction, large TotalAreaHalfRatio values typically indicated fast-moving cells. (f) For direction prediction, up-moving cells often had positive TotalPerimeterHalfRatio values and down-moving ones often had negative TotalPerimeterHalfRatio values. For speed prediction, large TotalPerimeterHalfRatio values typically indicated fast-moving cells. (scale bar: 10 μm).

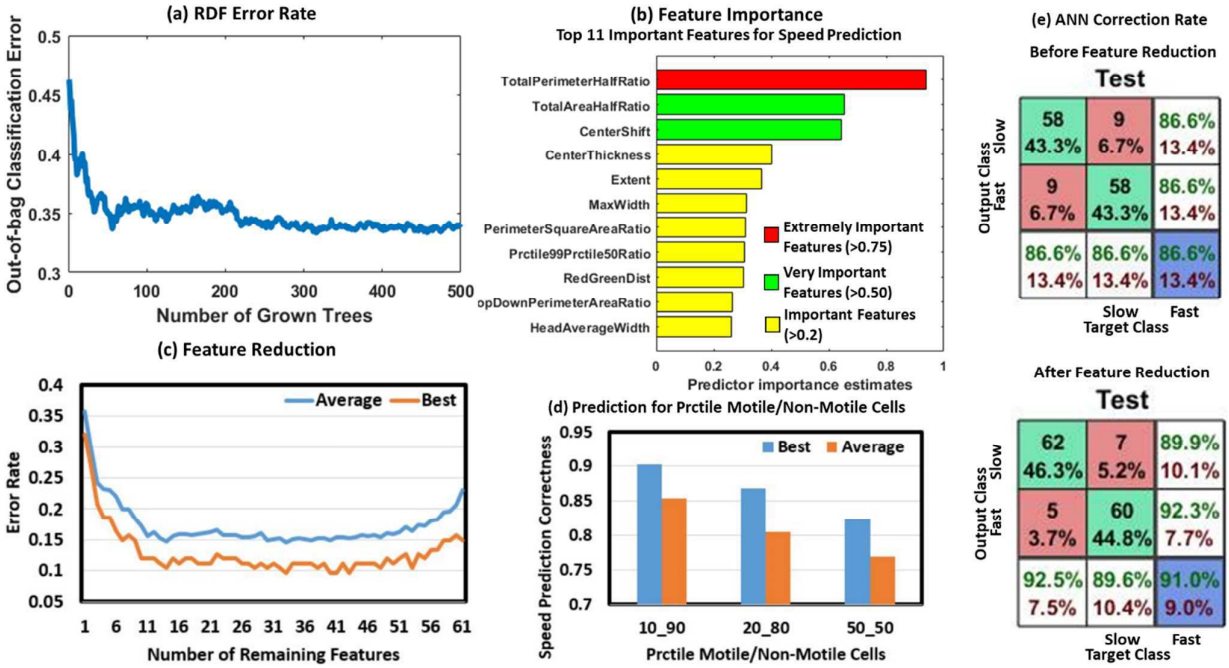

Fig. 5. Cell migration speed prediction. (a) Out-of-bag error rate for cell migration speed prediction using RDF. With the increase in number of grown trees, the error rate reduces to 33%. (b) Top 10 important features for cell migration speed prediction. (c) The error rate (average of 50 individual runs) can be further reduced to about 18% by eliminating 15 less relevant features and then the error rate remains stable until having 24 residual features. After that, the error rate goes up rapidly with the reduction of features. The features reduced at the last stage are critical features. (d) Correction rate for cell migration speed prediction before feature reduction was 86.6%. (e) Speed regression from our neural network model is correlated with cell actual speed. (pixels/s)

ANN. Previous studies suggest that cell migration is correlated with mitochondria distribution within a cell[6]. With the help of our machine learning model, we discovered that many other morphological features can also predict cell migration. Similar to cell migration direction, we found that cellular polarization-related features are still critical in determining speed. Unexpectedly, our model also found that other features also provide interesting insights into cell migration. For example, RedGreenDist, which is defined as the distance from the front of a cell to the first mitochondrion, normalized with the total length of the cell, is positively correlated with migration speed. This reveals that the mitochondria network does not necessarily have to be located at the leading edge of a cell to affect cell migration. Furthermore, CenterThickness, which is measured by taking the ratio of the average intensity over a small area of the nucleus region to the median intensity of the whole cell, suggests that the larger the difference is between the center area and cell edge, the more likely a cell moves faster. Evidence suggests that a fiber-like or fused mitochondrial network structure is favorable for supplying energy for cell migration, whereas dotted mitochondria, or mitochondrial fission, has been reported as an indicator of extracellular stress or cell apoptosis. However, in our study, we did not find strong correlation between mitochondria morphology with migration behaviors.

In the next step, we also implemented a neural network for migration speed prediction including classification of fast/slow moving cells and regression of moving speed. Targeting on selecting the high-migratory cells, which has been reported with significantly greater tumor formation and metastasis capabilities in mouse models, we set top 10% migration speed as a labelling threshold for fast-moving cells. Similarly, bottom

10% migration speed was used to label slow-moving cells, so that balanced inputs for both classes was obtained. Based on the extracted 61 morphology features, our 4-layer neural network classifier achieves 86.6% prediction accuracy at the best case (77.1% as average). To improve computational efficiency as well as avoid overfitting, we performed the wrapper method feature selection using the neural network as a performance evaluation model. We took an average of 50 individual runs on each of the leave-one-out subset of features, and picked the subsets that achieve the best accuracy on test data. As shown in Fig. 5(c), the NN classifier obtained an increase in accuracy when the number of features was reduced from 61 to around 33 with the best case reaching 91.0% accuracy (85.3% as average) (Fig. 5(d)). The further reduction of features will lead to a dramatic increase in error rate due to the loss of significant information. This also suggests that the longer one feature remains in the feature reduction process, the more it contributes to speed prediction. The remaining last 5 features were TotalAreaHalfRatio, TotalAreaRatio, CenterShift, TotalPerimeterHalfRatio, and MaxWidth, which matches well with the feature importance analysis in the random forest model. In addition, we also applied a neural network for regression to predict cell movement in a quantitative manner. This optimized model yields 0.0004 pixels/s (0.00006 μm/s) in normalized mean square error of migration speed (Fig. 5(e)).

Validation of Morphological Features by Altering Cell Migration Behaviors

Following our workflow, several morphological features were identified as critical markers of cellular migration. Due to the statistical nature of machine learning models, and its strong dependence on data inputs, our computational results could lead to a trivial or irreproducible discovery. Therefore, to validate the robustness and biological relevance of our model, we further designed control experiments with altered cell migration behavior and examined whether the critical morphological markers we found changed as the migration speed changed (Fig. 6). As a negative control, we inhibited the migration of SUM159 cells with doxorubicin, which has been widely used as treatment of metastatic breast cancer[25]. Doxorubicin treatment reduced the average migration speed by 25.3% as compared to cells treated with vehicle (control). As expected, we also observed a decrease in the average of some morphological markers, such as CenterShift, MaxWidth, and TotalAreaHalfRatio, which demonstrates a positive correlation with migration speed in our RDF prediction model. We also performed a positive control experiment by harvesting fast-moving cells (top 1% of the bulk SUM159 population), re-loading these selected cells to another of our migration devices, and observing their migratory behavior. We observed that highly-motile cells maintained their highly-migratory properties, and moved on average 34.0% faster than wild-type SUM159 cells. As a result, we also observed significantly higher values in those positively correlated morphological markers. A combination of the two experiments with “slow runners” and “fast runners” confirmed that the important features we pinpointed can be reliably used as morphological markers for cancer cell migration.

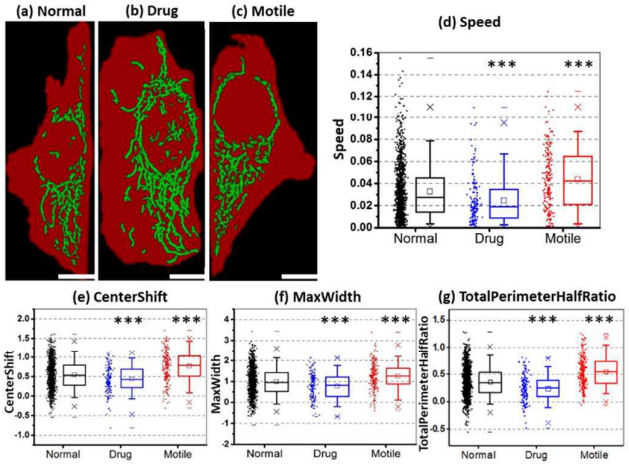

Fig. 6. Typical images for normal/drug-pretreated/motile cells and medium value differences of selected important features for cell migration direction/speed prediction. (a-c) Typical images for normal/drug-pretreated/motile cells. (a) Normal cells was not too fat and CenterShift was not so obvious. (b) Drug-pretreated cells were fat, flat and their mitochondria were more filamentous. (c) CenterShift values for motile cells were often quite large and their mitochondria were more fragmented. (scale bar: 10 μm). (d) Median speed for motile cells is faster than normal ones while median speed for drug-pretreated cells is slower than normal ones. (e-g) Medium value differences of selected important features for cell migration direction/speed prediction. The unit is pixel per second. (e) Medium CenterShift value for drug cells was below the one for normal cells while medium CenterShift value for motile cells was above the one for normal cells. (f) Medium MaxWidth value for drug cells was below the one for normal cells while medium MaxWidth value for motile cells was above the one for normal cells. (g) Medium TotalPerimeterHalfRatio value for drug cells was below the one for normal cells while medium TotalAreaHalfRatio value for motile cells was above the one for normal cells. “***” means significance level is smaller than 0.001.

Conclusions

Due to limitations of conventional marker-based approaches to identify motile cells, we aimed to establish a direct link between morphological features and cell migration. We focused on mitochondrial morphology because studies have shown that mitochondria influence cell migration. Although mitochondrial fragmentation has been reported to be associated with migratory behavior in different breast cancer cell lines[26], we found it has a weak correlation with the fast-moving and slow-moving SUM159 triple negative breast cancer cells. Furthermore, another study suggests a link

between mitochondria distribution and cell migration, yet the prediction power (53.4% accuracy) is too low to be reliable[13]. To improve upon this, our method extracted 61 morphological features of both mitochondria and the whole cell and correlated these features with migration at an accuracy of 72.0% max, 51.6% min, 54.7% mean, and 53.9% median. Although the accuracy is improved, this result suggests that the mechanisms underlying cell migration are complex and highlights limitations of conventional hypothesis-driven studies using only one parameter.

To address limitations of using single features, we applied cutting-edge Random Decision Forest (RDF) and Artificial Neural Network (ANN) models for prediction. To generate a large database for training models, we used our single-cell microfluidic migration chip[17] to track hundreds of cells on a chip. Using a database of 1,358 SUM159 cancer cells, we determined that the comprehensive computer vision method is significantly better than the conventional single feature-based prediction. To optimize the RDF model, we swept the number of trees and found that 500 trees are enough for prediction. Optimization of ANN was more complicated, as we had to remove redundant and irrelevant features as well as determine the numbers of layers and hidden nodes. We found that removing around 28 features and building a geometry using a 4-layer neural network (2 layers with hidden nodes, 21 hidden nodes in the first layer and 7 hidden nodes in the second layer) achieves the highest prediction power. Using the ANN model, we achieved over 99% correct prediction for movement direction and 91% for speed, while the RDF model is slightly less accurate (67% for speed).

In addition to prediction, we used the RDF model and reduced the features in the ANN model to pinpoint top-ranked key features important for cell migration. Some of these key features are known to relate to cell polarization (such as CenterShift), but we also identified novel features (such as RedGreenDist and CenterThickness) that correlated with cell migration. The identification of novel features highlights the limitations of current methods, and potentially advances our understanding about mechanisms involved in cell migration. To validate that the identified features are indeed critical for cell migration, we performed migration experiments using pharmaceutically pre-treated cells (expected to have lower speed), and highly migratory cells from our microfluidic device (expected to have higher speed)[27]. When comparing these experimental cell populations with wild-type cancer cells, we found the same associations between morphological features (CenterShift, MaxWidth, and TotalAreaHalfRatio) and cell speed, further supporting the importance of our discovered features in cell movement.

In this study, we established a method to predict cell movement by morphological features using computer vision and machine learning, achieving unprecedented prediction power for cell movement. This unbiased method discovers both known and novel features critical for the cell migration process. The features identified here can aid in our understanding of cancer cell migration, and lead to new approaches for identifying metastatic cancer cells. Although the current study focuses on one breast cancer cell line on a 2D substrate, the strong prediction power of the morphological markers suggests broader applications for this method. In the future, this method can be used to explore other cancer types, cell movement in a 3D environment, and other cell behaviors, such as metabolism and cell-cell interaction.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported in part by the Department of Defense (W81XWH-12-1-0325) and in part by the National Institutes of Health (1R21CA17585701, 1R21CA19501601A1, U01CA210152, and R01CA196018). Y.-C. Chen acknowledges the support from Forbes Institute for Cancer Discovery. The Lurie Nanofabrication Facility of the University of Michigan (Ann Arbor, MI) are greatly appreciated for device fabrication.

Notes and references

1 D. Hanahan, RA. Weinberg. Hallmarks of cancer: the next generation. cell. 2011, **144**, 646-74.
2 Steeg, P. S. Tumor metastasis: mechanistic insights and clinical challenges. Nat. Med. 12, 895–904 (2006).
3 Zeisberg, M. & Neilson, E. G. Biomarkers for epithelial-mesenchymal transitions. J. Clin. Invest. 119, 1429–1437 (2009).
4 Chaw, S. et al. Epithelial to mesenchymal transition (EMT) biomarkers – E-cadherin, beta-catenin, APC and Vimentin – in oral squamous cell carcinogenesis and transformation. Oral Oncol. 48, 997–1006 (2012).
5 Liu, F., Gu, L. N., Shan, B. E., Geng, C. Z. & Sang, M. X. Biomarkers for EMT and MET in breast cancer: An update (Review). Oncol. Lett. 12, 4869-4876 (2016).
6 Giedt, R. J. et al. Computational imaging reveals mitochondrial morphology as a biomarker of cancer phenotype and drug response. Sci. Rep. 6, 32985; 10.1038/srep32985 (2016).
7 Westermann, B. Bioenergetic role of mitochondria fusion and fission. Biochim. Biophys. Acta. 1817, 1833-1838 (2012)
8 Peng, J.-Y. et al. Automatic Morphological Subtyping Reveals New Roles of Caspases in Mitochondrial Dynamics. PLoS Comput. Biol. 7, e1002212 (2011).
9 Danuser, G. Computer Vision in Cell Biology. Cell 147, 973–978 (2011).
10 Grys, B. T. et al. Machine learning and computer vision approaches for phenotypic profiling. J. Cell Biol. 216, 65–71 (2016).
11 Breiman, L. Random forests. Machine learning 45(1), 5-32 (2001).
12 Zell, A. Simulation neuronaler netze (Vol. 1). Bonn: Addison-Wesley (1994).
13 Desai, S. C. A. P., Bhatia, S. N., Toner, M. & Irimia, D. Mitochondrial Localization and the Persistent Migration of Epithelial Cancer cells. Biophys. J. 104, 2077–2088 (2013).
14 Whitesides, G. M. The origins and the future of microfluidics. Nature 442, 368–373 (2006).
15 Xia, Y. & Whitesides, G. M. Soft lithography. Annu. Rev. Mater. Sci. 28, 153-184, (1998)

16 El-Ali, J., Sorger, P. K. & Jensen, K. F. Cells on chips. Nature 442, 403–411 (2006).
17 Chen, Y.-C. et al. Single-cell Migration Chip for Chemotaxis-based Microfluidic Selection of Heterogeneous Cell Populations. Sci. Rep. 5, 9980; 10.1038/srep09980 (2015).
18 Lane, N. Mitochondrial disease: Powerhouse of disease. Nature 440, 600–602 (2006).
19 Ridley, A. J. Cell Migration: Integrating Signals from Front to Back. Science 302,1704–1709 (2003).
20 Detmer, S. A. & Chan, D. C. Functions and dysfunctions of mitochondrial dynamics. Nat. Rev. Mol. Cell Biol. 8, 870-879 (2007).
21 Sheetz, M. P., Felsenfeld, D. P. & Galbraith, C. G. Cell migration: regulation of force on extracellular-matrix-integrin complexes. Trends Cell Biol. 8, 51–54 (1998).
22 Gardel, M. L., Schneider, I. C., Aratyn-Schaus, Y. & Waterman, C. M. Mechanical integration of actin and adhesion dynamics in cell migration. Annu. Rev. Cell Dev. Biol. 26, 315-333 (2010).
23 Ho, T. K. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832-844 (1998).
24 Mellor, A., Haywood, A., Stone, C. & Jones, S. The Performance of Random Forests in an Operational Setting for Large Area Sclerophyll Forest Classification. Remote Sens. 5, 2838–2856 (2013).
25 Bandyopadhyay, A. et al. Doxorubicin in Combination with a Small TGFβ Inhibitor: A Potential Novel Therapy for Metastatic Breast Cancer in Mouse Models. PLoS One 5, e10365 (2010).
26 Tu, Y. et al. Abstract 467: Mitochondrial dynamics regulates migration and invasion of breast cancer cells. Cancer Res. 72, 467–467 (2012).
27 Chen, Y.-C. et al. Microfluidic high-throughput motility-based cell selection for enriching tumor initiating cells and discovering inhibition pathways of cancer migration. In The 20th International Conference on Miniaturized Systems for Chemistry and Life Sciences (MicroTAS ′16) 49-50, Dublin, Oct. (2016).
28 Kitay, B. M., Mccormack, R., Wang, Y., Tsoulfas, P. & Zhai, R. G. Mislocalization of neuronal mitochondria reveals regulation of Wallerian degeneration and NMNAT/WLDS-mediated axon protection independent of axonal mitochondria. Hum. Mol. Genet. 22, 1601–1614 (2013).
29 Smith, M. C. P. et al. CXCR4 Regulates Growth of Both Primary and Metastatic Breast Cancer. Cancer Res. 64, 8604–8612 (2004).

Integrative Biology

Table of Contents Entry

a) Textual highlights:

Cell migratory direction and speed are predicted based on morphological features using computer vision and machine learning algorithms.

b) Graphic highlights: