Coarse Raman and optical diffraction tomographic imaging enable label-free phenotyping of isogenic breast cancer cells of varying metastatic potential

Santosh Kumar Paidi[1,*], Vaani Shah[2,*], Piyush Raj[1], Kristine Glunde[3,4], Rishikesh Pandey[5,6], Ishan Barman[1,3,7]

[1]Department of Mechanical Engineering, Johns Hopkins University, Baltimore, MD, 21218

[2]Fischell Department of Bioengineering, University of Maryland, College Park, MD, 20742

[3]The Russell H. Morgan Department of Radiology and Radiological Science, The Johns Hopkins University School of Medicine, Baltimore, MD, 21205

[4]The Sidney Kimmel Comprehensive Cancer Center, The Johns Hopkins University School of Medicine, Baltimore, MD, 21287

[5]CytoVeris Inc, Farmington, CT, 06032

[6]Department of Biomedical Engineering, University of Connecticut, Storrs, CT, 06269

[7]Department of Oncology, Johns Hopkins University, Baltimore, MD, 21287

*Both authors contributed equally to this work.

Corresponding author: Ishan Barman, ibarman@jhu.edu

The authors disclose no potential conflicts of interest.

Abstract

Identification of the metastatic potential represents one of the most important tasks for molecular imaging of cancer. While molecular imaging of metastases has witnessed substantial progress as an area of clinical inquiry, determining precisely what differentiates the metastatic phenotype has proven to be more elusive underscoring the need to marry emerging imaging techniques with tumor biology. In this study, we utilize both the morphological and molecular information provided by 3D optical diffraction tomography and Raman spectroscopy, respectively, to propose a label-free route for optical phenotyping of cancer cells at single-cell resolution. By using an isogenic panel of cell lines derived from MDA-MB-231 breast cancer cells that vary in their metastatic potential, we show that 3D refractive index tomograms can capture subtle morphological differences among the parental, circulating tumor cells, and lung metastatic cells. By leveraging the molecular specificity of Raman spectroscopy, we demonstrate that coarse Raman microscopy is capable of rapidly mapping a sufficient number of cells for training a random forest classifier that can accurately predict the metastatic potential of cells at a single-cell level. We also leverage multivariate curve resolution – alternating least squares decomposition of the spectral dataset to demarcate spectra from cytoplasm and nucleus, and test the feasibility of identifying metastatic phenotypes using the spectra only from the cytoplasmic and nuclear regions. Overall, our study provides a rationale for employing coarse Raman mapping to substantially reduce measurement time thereby enabling the acquisition of reasonably large training datasets that hold the key for label-free single-cell analysis and, consequently, for differentiation of indolent from aggressive phenotypes.

Keywords: Raman spectroscopy; optical diffraction tomography; breast cancer; metastasis; random forests; single-cell phenotyping

Introduction

Timely assessment of risk is critical for the detection and treatment of metastatic disease, which remains the main reason for cancer-related mortality [1]. The current clinical standard for assessment of metastatic risk relies on pathologic examination of sentinel lymph nodes following biopsy. In addition to being an invasive procedure, identification of metastatic cells in lymph node biopsies can be challenging and lead to an increase in false negatives [2]. Early detection of metastasis requires tools that can recognize the metastatic potential of cancer cells derived from the primary tumor or liquid biopsies. As primary tumors grow, a small fraction of the cancer cells termed circulating tumor cells (CTC) undergo epithelial to mesenchymal transition (EMT), locally invade the surrounding stroma, intravasate and are shed into the bloodstream leveraging their enhanced motility [3]. A few of these cells survive in the circulation to extravasate, locally invade and form premetastatic niches in secondary organs (e.g. lungs in breast cancer), and colonize through re-acquisition of epithelial characteristics via mesenchymal to epithelial transition (MET) [3]. While our understanding of the processes involved in metastasis has improved substantially in recent years, detecting phenotypic subtypes with metastatic competence has proven to be elusive due in part to the substantial heterogeneity observed in these cell populations [4]. Furthermore, our understanding of what imparts metastatic potential remains rudimentary, and biomarkers that can recognize such competence across different carcinomas are still lacking.

Early genomic analyses of tumors revealed additional organ-specific mutations in metastatic tumors despite sharing common ancestors [5]. Similarly, transcriptional analyses of breast cancer metastasis to various organs including lungs and brain have also identified largely distinct signatures characteristic of organotropism [6, 7]. However, these population-based analyses require elaborate sample preparation and fail to capture the variations in phenotypes at a single-cell level. Recently, we and others have also investigated the physical properties associated with the differences in metastatic phenotypes in specialized microfluidic platforms [8-15]. The pursuit of isolating CTC from blood to determine the course of metastatic disease has resulted in the proliferation of several cell labeling methods that leverage known epithelial markers for identification [16-21]. However, the use of epithelial markers may not be sufficient to detect CTC that undergo EMT to acquire mesenchymal properties, particularly in triple-negative breast cancers [22]. Also, the sensitivity of these methods is challenged by the small number of known markers of metastatic progression that can be targeted simultaneously.

To address these challenges, several techniques based on optical microscopy and imaging have attempted label-free phenotyping of cancer cells [23-27]. For example, phenotypic changes of 4T1 murine breast cancer cells in response to drug treatment were characterized in terms of morphological parameters extracted from fluorescence images in 3D cultures [25]. Rohde and co-workers have developed an automated platform for morphological analysis of cellular phenotypes using transport-based morphometry [24], which we recently used to analyze the quantitative phase images of activated and naïve CD8$^+$ T cells [28]. Similar optical methods have also been leveraged for single-cell analysis of cancer phenotypes [29, 30], which often require large datasets for building robust prediction models.

In this study, we employed 3D optical diffraction tomography (ODT) and label-free Raman spectroscopy to quantitatively investigate both morphological and molecular differences between isogenic breast cancer cells of varying metastatic potential. We used a set of three isogenic cell lines composed of the parental MDA-MB-231 triple-negative breast cancer cell line (P231), circulating tumor cells (CTC), and lung metastatic cells (LM) where the latter two were derived respectively from the circulation and lungs of a mouse bearing parental P231 cells [31, 32]. Compared to the widely used qualitative phase imaging methods, such as Zernike phase contrast and differential interference contrast, quantitative phase imaging methods recover the phase delay caused by the sample, decoupled from absorption information. ODT is a form of quantitative phase imaging that allows morphological analysis of single-cells based on their 3D refractive index (RI) profiles [33-35]. In addition to providing traditional measures of morphology such as area and aspect ratio, the RI information allows a label-free and non-contact route for the determination of cell dry mass and local thickness of specimens with nanometric sensitivity [36]. The additional morphological insights provided by optical diffraction tomography have been increasingly exploited for label-free and stain-free *in vitro* analysis of cells and tissues [33, 34, 37]. Yet, most of the cellular studies have focused on either visualization of morphological dynamics in response to external stimuli such as drug exposure in single-cells or rapid identification of cells such as bacteria and white blood cells using deep learning by leveraging large datasets [38, 39]. Its utility in assessment of phenotypic differences among closely related mammalian cancer cells, particularly in data-limited settings, remains largely unexplored.

Raman spectroscopy, on the other hand, provides a label-free route for assessment of biological specimens with exquisite molecular specificity [40, 41]. This optical technique, based on the inelastic scattering of light, probes vibrational modes of molecules and allows direct profiling of molecular composition of biological specimens including live cells and tissues in their native states [40-42]. The simple integration of Raman spectroscopy with optical microscopy facilitates seamless vibrational spectroscopic imaging at diffraction-limited spatial resolution with subcellular resolution. Several groups including our own laboratory have exploited the high resolution and rich molecular information afforded by Raman spectroscopic imaging to study the molecular progression of cancer [23, 41, 43-45]. Due to the low likelihood of spontaneous Raman scattering, most single-cell imaging studies have focused on employing nanoparticles for plasmonic enhancement of signals and selective tagging of subcellular regions of interest [46, 47]. Therefore, only a few studies have attempted label-free characterization of cells for studying biological processes associated with physiological changes, disease progression, and drug response [48-50]. Our recent label-free Raman investigation in pellets of isogenic breast cancer cell lines that exhibit organotropism to brain, liver, lung, and spine revealed distinct metastatic organ-specific spectral signatures that were confirmed by metabolomics analysis [23]. Due to the long acquisition time, label-free Raman spectroscopic studies have either exploited high-resolution single-cell maps for analysis of limited cells or bulk sampling of cell populations that permits the use of machine learning algorithms for classification problems by generating large datasets at the cost of spatial information. This tradeoff between obtaining higher spatial resolution maps and acquiring sufficiently large datasets amenable for machine learning has largely prevented the use of machine learning techniques to learn and predict cellular phenotypes from Raman images with single-cell analytical resolution.

Therefore, in this study, we sought to test whether morphological attributes encoded by ODT and biomolecular insights obtained using Raman spectroscopy can predict the phenotype of the closely related isogenic cell lines P231, CTC and LM of varying metastatic potentials with statistical confidence. Using the 3D RI profiles obtained from ODT of single-cells, we compared the distributions of morphological parameters such as area, aspect ratio, and dry mass across the three classes, and used their combination to train and test random forest classifiers for automated identification. By leveraging coarse Raman sampling of single cancer cells to reduce the acquisition time and obtain spectral maps from a larger number of cells, we explored the intersection of the abovementioned resolution-sampling tradeoff to find a solution for identifying metastatic phenotype of cancer cells with single-cell analytical resolution. To show that coarse Raman maps capture sufficient information for achieving single-cell phenotyping, we used random forests to iteratively test spectral maps of individual cells against classifiers trained on the data from the remaining cells in the dataset. Furthermore, we used multivariate curve resolution alternating least squares (MCR-ALS) analysis to identify the spectra from subcellular compartments and test the utility of random forest classification in predicting the metastatic phenotype when only spectra from either nucleus or cytoplasm are available. The ability to use specific subcellular regions of the cell for chemical imaging is expected to further reduce the spectral acquisition time and boost the number of cells in the training dataset to capture population heterogeneity. Such label-free identification of cells with high metastatic competence would not only have a profound impact on the prediction of a patient's risk of developing metastasis but also inform the design of optimal, personalized therapeutic treatments.

Materials and methods

Cell culture

An isogenic panel of varying metastatic potential derived from the human breast cancer cell line MDA-MB-231 was used in this study. In addition to the td-Tomato expressing parental MDA-MB-231 cells (P231), the panel consisted of CTC and LM cells previously obtained after orthotopic implantation of the parental cells in the fourth right mammary fat pad of female athymic nu/nu female mouse (NCI) as detailed in our previous publications [31, 32]. The three cell lines were cultured in RPMI-1640 media supplemented with 10% fetal bovine serum (FBS), 100 U/ml penicillin, and 100 μg/ml streptomycin and maintained at 37 ^{0}C and 5% CO_2 in a humidified incubator.

Optical diffraction tomography and data analysis

The three cell lines were seeded in glass coverslip-bottom Petri dishes for tomography. The morphological assessment of the cells was performed on an ODT system (HT-1H, Tomocube Inc., Republic of Korea) comprised of a 60X water-immersion objective (1.2 NA), an off-axis Mach-Zehnder interferometer with a 532 nm laser and a digital micromirror device (DMD) for tomographic scanning of each cell [51]. The 3-D RI distribution of the cells was reconstructed from the interferograms using the Fourier diffraction theorem as described previously [52]. TomoStuido (Tomocube Inc, Republic of Korea) was used to perform reconstruction and visualization of 3D RI maps and their 2D maximum intensity projections (MIP). The 2D MIP images were segmented using CellProfiler^{TM} (v3.1.9) software to isolate single-cells using Otsu two-class thresholding and neglecting the partial cells at the boundaries of raw images [53, 54]. After segmentation, we obtained 57, 35, and 44 cells in P231, CTC, and LM classes respectively. The area of each cell was calculated by counting the number of non-zero pixels in their corresponding segmentation masks generated by the CellProfiler^{TM} software [55]. Similarly, the perimeter and aspect ratio were calculated respectively as the number of non-zero pixels at the edges of the masks and the ratio of major and minor axes lengths [55]. The cell dry mass was calculated from the 3D RI profile [56]. The morphological parameters were used to train a random forest classifier with 100 trees using the MATLAB TreeBagger class and inspect the out-of-bag-error.

Raman spectroscopic imaging

The cells from three different passages (biological repeats) for each class were seeded on quartz slides (1 in x 1 in) coated with poly-lysine and incubated overnight to facilitate cell attachment for Raman imaging. The cells were fixed using 4% paraformaldehyde and washed prior to imaging in phosphate buffered saline (PBS) at room temperature. Five cells from each slide (technical repeats) were randomly selected for Raman mapping. The coarse single-cell Raman imaging experiments were performed on a HORIBA XploRA PLUS confocal Raman microscope. A 532 nm diode laser was used for excitation and delivered to the sample via a 60X water immersion objective (1.2 NA). The backscattered Raman light was dispersed using an 1800 lines/mm diffraction grating and imaged on a thermoelectrically cooled CCD coupled to the microscope. The spectra were acquired from the points on a coarse rectangular grid overlaid on each single cell to obtain spectra from various subcellular regions and capture intracellular spatial heterogeneity. Each spectrum in the fingerprint region (600-1950 cm^{-1}) was acquired by exposing the sample to a laser power *ca.* 1 mW at each point for 2.5s (5 accumulations of 0.5s exposure).

Raman data analysis

All the Raman spectral analysis was performed in MATLAB 2017b (Mathworks) environment. Spectroscopic imaging of each cell provided a hyperspectral dataset, where each pixel on the rectangular mapping grid corresponds to a Raman spectrum. The hyperspectral datasets from all the cells were unfolded (by preserving the spatial information and cell identity) and concatenated to form a combined spectral dataset for further analysis. The spectra in the fingerprint region were subjected to background subtraction using a fifth-order best-fit polynomial-based fluorescence removal method and cosmic ray removal using median filtering on the groups of spectra from each cell. Next, the points on the mapping grid exterior of the imaged cell were identified by Otsu thresholding on the 1452 cm^{-1} peak (CH_2 bending mode of proteins) intensity and labeled separately for further analysis. The spectra were finally vector normalized to remove the variations in laser power across the experiments.

To identify spectra from specific subcellular regions, we performed MCR-ALS analysis for decomposing each spectrum into its constituents by iterative fitting under nonnegativity constraints on the obtained component spectra (loadings) and their contributions (scores) [57]. The components were identified as rich in cytoplasm, nucleus, and background (quartz and water) characteristics and confirmed by re-constructing the score maps for each cell. Each spatial location in the cell is assigned either cytoplasm, nucleus, mixed or background based on the Otsu thresholding of the cytoplasm-like and nucleus-like component scores, which were both negatively correlated [54]. The scores corresponding to cytoplasm-like and nucleus-like constituents were compared across the three cell lines through violin plots with outlier suppression for clarity. The significance of differences between medians was determined using Wilcoxon rank-sum test with the conventional threshold.

Random forest classifiers (bootstrap-aggregated or bagged decision trees) were trained using the TreeBagger class in MATLAB to enable the identification of the metastatic phenotypes. We used a leave-one-cell-out protocol by leaving one cell out each time as a test case and training random forest classifiers on spectra from the remaining cells. Several iterations of training were performed for each test case by selecting randomized subsets of training data to ensure equal membership for all the three classes and to avoid overtraining for the class with high data availability. The spectra of the excluded cell were subjected as a test dataset and the class label for the entire cell was determined according to the following class assignment criterion. Since the test dataset (left-out cell) remained the same for all training iterations, the median of predicted labels for each spectrum was used for decision making at the cell level. For each cell, the majority class was assigned as the predicted class if its membership was at least 30% higher than the random chance prediction and 30% higher than the membership of the second majority class. If these conditions were not met by the majority class, the test cell was labeled unclassified. To verify the ability of cytoplasm and nucleus spectra for the identification of metastatic phenotype, the random forest classifiers were run on the cytoplasm- and nucleus-rich spectra identified by the MCR-ALS analysis, in addition to running them on the entire spectral dataset.

Results and discussion

The availability of isogenic breast cancer cells of varying metastatic potential derived from the same MDA-MB-231 human breast cancer cell line (**Fig. 1A**) enabled us to investigate the utility of label-free optical imaging for the identification of metastatic phenotypes at the single-cell level. We used parental P231 cells along with their circulating (CTC) and lung metastatic (LM) variants to assess the efficacy of ODT (**Fig. 1B**) and Raman spectroscopy (**Fig. 1C**) for capturing the phenotypic differences in terms of their morphological and molecular attributes. Our previous characterization of these cell lines confirmed their distinct metastatic abilities commensurate with the stage and organ from which they were isolated [31]. Our recent investigation of the biophysical properties of these cells revealed that the LM cells are most motile and least stiff, which bestow them with unique invasive capability [8].

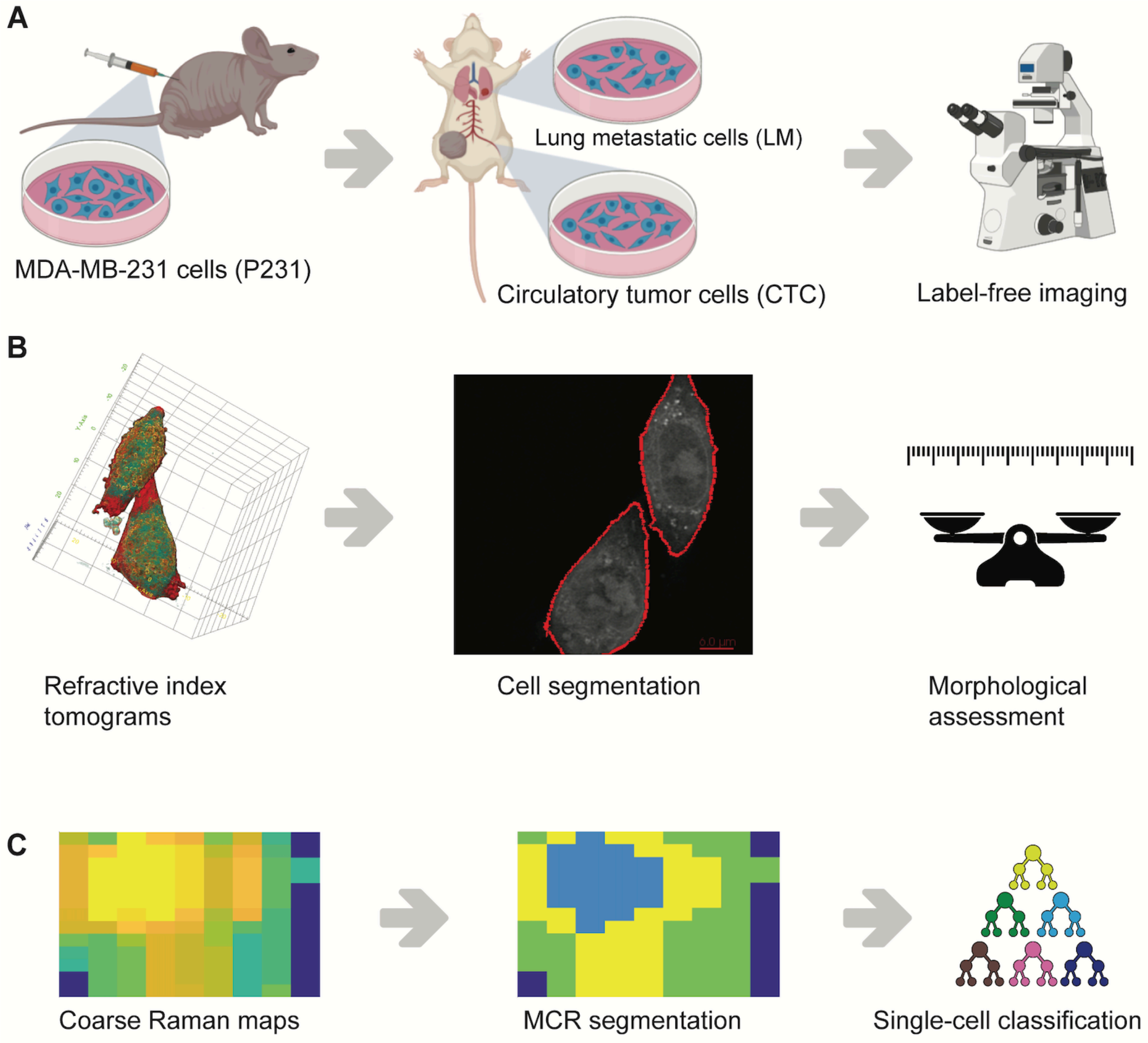

Figure 1. Label-free identification of metastatic phenotypes. (A) Circulating tumor cells (CTC) and lung metastatic cells (LM) used in the study were isolated from the blood and lungs of mice bearing parental MDA-MB-231 (P231) tumor xenografts. **(B)** Refractive index tomograms were segmented to isolate single-cells for morphological assessment. **(C)** Coarse Raman maps of single-cells were subjected to MCR-ALS analysis to identify subcellular regions rich in cytoplasm and nucleus prior to the use of supervised classification using random forests.

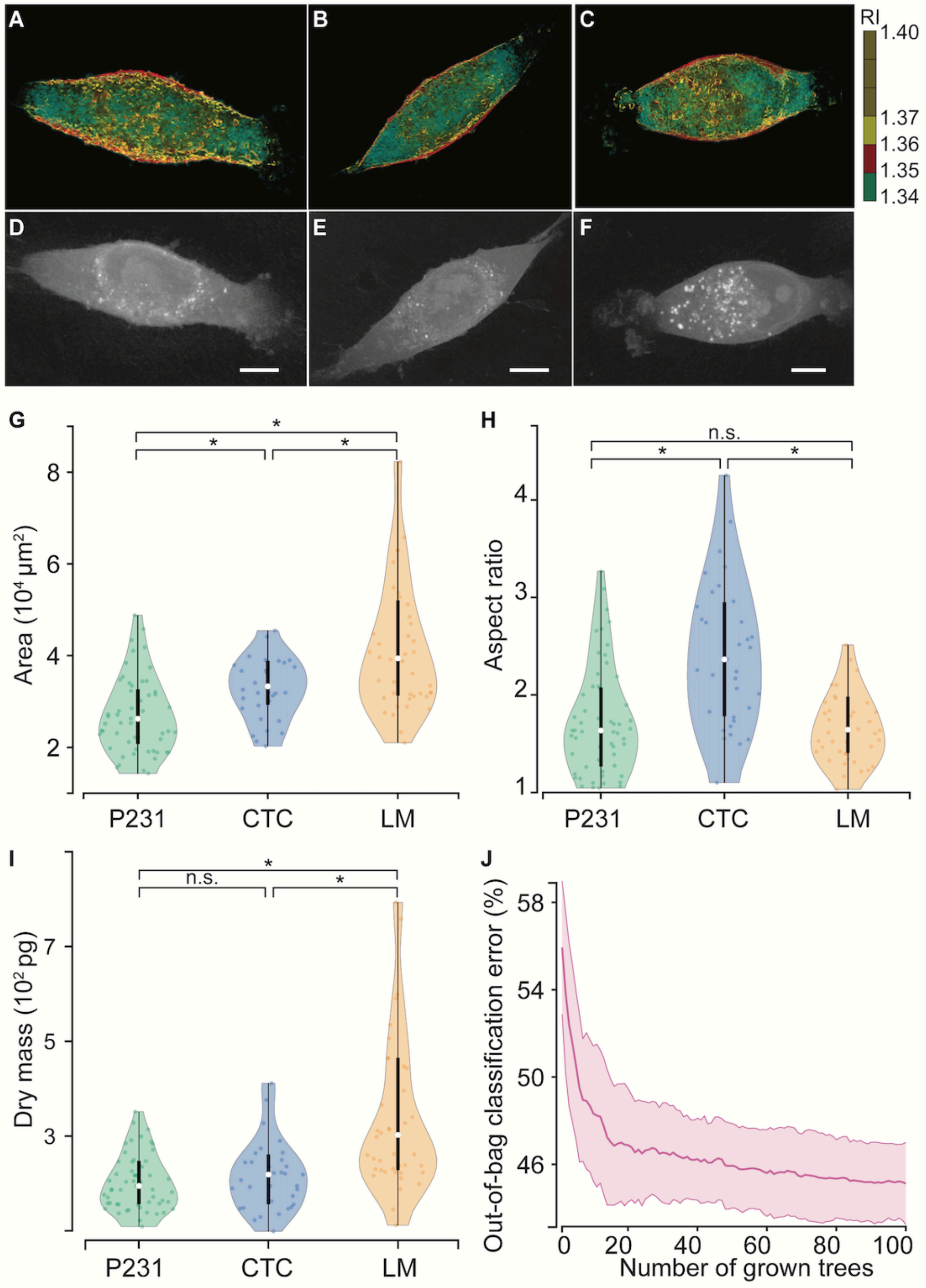

Figure 2. Morphological assessment of metastatic phenotypes. Representative 3D refractive tomograms of **(A)** P231, **(B)** CTC, and **(C)** LM cells show the intracellular variation of the refractive index. The maximum intensity projections of (**D**) P231, (**E**) CTC, and (**F**) LM were used for determining the 2D morphological parameters. The violin plots show the variations in **(G)** area, **(H)** aspect ratio, and **(I)** dry mass across the three classes. **(J)** The out-of-bag classification error plot shows that random forests build on the morphological parameters fail to accurately predict metastatic phenotypes. The scale bars represent 5 µm. * represents statistically significant differences at $p < 0.05$ threshold (Wilcoxon rank-sum test)**.**

To understand the morphological differences among the phenotypically distinct isogenic cell lines, we acquired 3D RI tomograms of single cancer cells belonging to each group. While the RI tomograms of cells from all three cell groups (**Fig. 2A-C**) show expected intracellular heterogeneity arising from RI variations across subcellular compartments, the intercellular differences are not apparent from gross visual inspection. Therefore, the maximum intensity projections of the RI tomograms (**Fig. 2D-F**) were subjected to further assessment using CellProfiler^{TM} software to quantify morphological parameters such as area and aspect ratio. Also, we calculated the cell dry mass directly from the 3D RI tomograms to include an additional dimension in the morphological analysis that cannot be readily measured from brightfield or phase contrast microscopy. As seen in **Fig. 2G**, we observed that the area of the cells increased steadily with the increase in metastatic potential from P231 to LM. However, a significant increase in aspect ratio (**Fig. 2H**) was only observed for the CTC in comparison to the P231 and LM classes, while the differences between the latter were not statistically significant. These observations are consistent with the characteristics of EMT and MET processes in metastasis of P231 cells to lungs that respectively result in the acquisition of a spindle shape by the CTC for enhanced motility to reach the metastatic site and re-acquisition of epithelial shape for promoting the proliferation of LM cells to form metastatic tumors [3]. We observed that the cell dry mass (**Fig. 2I**) increased with the metastatic potential of the isogenic cells, but the difference was statistically significant only for LM cells in comparison to P231 and CTC. The increase in the cell dry mass of the LM cells is consistent with the prior observation of an increased RI and cell dry mass of cancer cells in comparison with normal cells due to the higher accumulation of proteins associated with the higher proliferation of the former group [58]. Our observation expands this idea to the metastatic regime and provides a rationale to explore cell dry mass as a potential biomarker of invasiveness in future studies.

While the phenotypically distinct cell lines showed variable differences in individual morphological parameters, their utility for identifying phenotypes of single cancer cells is dependent on the existence of clear class boundaries between the three classes. The violin and box plots (**Fig. 2G-I**) show the appreciable overlap in the distribution of each morphological parameter across the metastatic potential thus making univariate analyses challenging for class separation. While prior studies have leveraged deep learning methods for cellular classification based on latent morphological features from the complete cell images, they have largely probed simpler systems such as bacteria, blood cells, immune cells, and anthrax spores compared to the current cohort of isogenic breast cancer cells [38, 39]. Since our current study is focused on the detection of the cellular phenotypes within the constraints of small training datasets, we trained random forest classifier to test if supervised models leveraging these three morphological parameters can accurately predict the metastatic phenotype of test samples. The out-of-bag classification error rate (**Fig. 2J**), calculated for each training sample by testing them against the decision trees in the forest that did not use them for training, was found to asymptotically plateau around 46% (compared to 33.3% random chance). These results indicate that while phenotype differences among the isogenic cells show subtle but significant morphological differences, they are not sufficient for robust classification of closely related cells at a single-cell analytical resolution, particularly when the training data is relatively scarce.

Next, we sought to check if molecular information provided by vibrational spectroscopy can enable the identification of metastatic phenotypes at a single-cell analytical resolution. Therefore, to build a dataset large enough for training machine learning models, we performed coarse Raman microscopy of all three isogenic cell lines. The entirety of each cell was mapped coarsely (average of 67 pixels per cell) to capture intracellular heterogeneity. While these coarse Raman images (**Fig. 3A**) do not offer diffraction-limited spatial resolution, they capture enough information for the cell classification task and help significantly reduce the spectral acquisition time for each cell. The mean (+/- 1 s.d.) of the spectra from the three cell lines (**Fig. 3B**) show prominent peaks at 931 cm^{-1}, 1003 cm^{-1}, 1085 cm^{-1}, 1303 cm^{-1}, 1450 cm^{-1}, 1658 cm^{-1} indicative of the common biological constituents of cells and tissues [59]. Since there are no discernible visible differences between the spectra of the three cell lines, we used MCR-ALS analysis to decompose the spectra into component spectra and their scores. MCR-ALS decomposition allows representation of each spectrum in the dataset as a weighted sum of iteratively generated pure component-like basis spectra, without requiring any composition estimates as inputs [57]. In this study, a simple three-component MCR-ALS decomposition provided component loadings harboring features of cytoplasm, nucleus, and quartz background from the slide on which the cells were cultured (**Fig. 3C**). We identified MC1 and MC2 as loadings resembling cytoplasm and nucleus due to the prominence of cytoplasm features at 1003 cm^{-1} (C–C stretching vibration of the aromatic ring in the phenylalanine side chain) and 1639 cm^{-1} (amide I feature in proteins) in the former and nucleic acid features at 788 cm^{-1} (O-P-O stretching in DNA) and 1092 cm^{-1} (symmetric PO_2^- stretching in DNA) in the latter. The assignment is also justified by the strong negative correlation between the MC1 and MC2 scores for each cell, assessed by an average correlation coefficient of -0.98 over all the cells in the study.

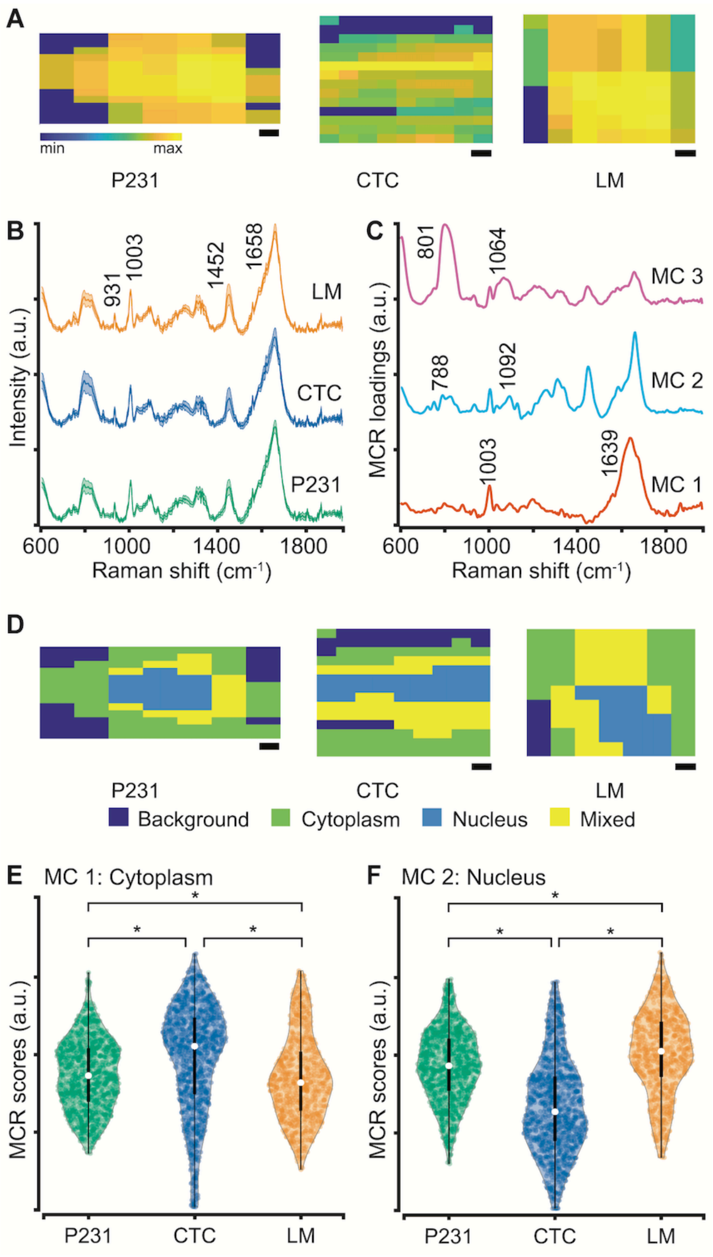

Figure 3. MCR segmentation of single-cell Raman images. (A) Representative coarse Raman maps reconstructed using the 1452 cm^{-1} peak intensity shown for P231, CTC, and LM cells. **(B)** Mean Raman spectra (with the shadow representing 1 s.d. and vertical offset for clarity) are shown and some prominent biological peaks highlighted for the three isogenic cell lines used in the study. **(C)** The three MCR component loadings derived from the combined spectral dataset are shown. MC1, MC2, and MC3 respectively show cytoplasm-like, nucleus-like, and quartz background spectral features. **(D)** The segmentation maps constructed by thresholding on MCR component scores for the cells in panel A are shown. The violin plots with embedded box and whisker plots show the distribution of MCR scores for cytoplasm-like **(E)** and nucleus-like **(F)** loadings. The scale bars represent 2 μm. * represents statistically significant differences at $p < 0.05$ threshold (Wilcoxon rank-sum test)**.**

We further verified the assignment by reconstructing the abundance maps for the scores of MC1 and MC2. We assigned each pixel as cytoplasm, nucleus, or mixed by thresholding on the MC1 and MC2 scores. Using this MCR-ALS decomposition of spectral dataset allows better visualization of the spatial demarcation between cytoplasm and nucleus, which was not apparent in the coarse Raman maps at individual wavenumbers. The identification of pixels as those rich in cytoplasm and nucleus allow us to dissect the heterogeneous single-cell Raman measurements into relatively homogenous subsets for identifying subcellular compartments that capture the information necessary for identification of metastatic phenotypes. The remaining loading MC3, showing features at 801 cm^{-1} and 1064 cm^{-1}, captures the minor contributions of quartz substrate in the cell spectra. We compared the scores of the cytoplasm-like and nucleus-like components to understand the relative abundance of these components in the cells of varying metastatic potential. The violin plots of MC1 and MC2 show that while the median values for the cytoplasm scores are significantly higher for the CTC in comparison to the P231 cells, the median for the LM cells is significantly lower in comparison to both P231 and CTC groups. Since the nucleus scores are negatively correlated with the cytoplasm scores, their medians show an opposite trend. The similarity of P231 and LM scores and their deviation from CTC hint at the ability of Raman spectroscopy to identify the differences associated with the EMT and MET processes that make CTC dissimilar to the P231 and LM cells. These observations are consistent with our prior characterization of these cell lines that showed significant differences in the mRNA expression levels of osteopontin (OPN), CD44, and vimentin (VIM) – genes involved in EMT, cell migration, extracellular matrix organization, and cell adhesion – in CTC and LM cells compared to the P231 cells [31]. Since the MCR scores represent the contribution of the pure component-like spectra to each spectrum irrespective of its location in the cell, the observed differences in the scores across the metastatic cascade provide relatively limited direct biological insights. However, these statistically significant differences in the MC component scores provide a rationale for exploring supervised classification techniques for the determination of metastatic phenotypes in the studied cancer cell lines.

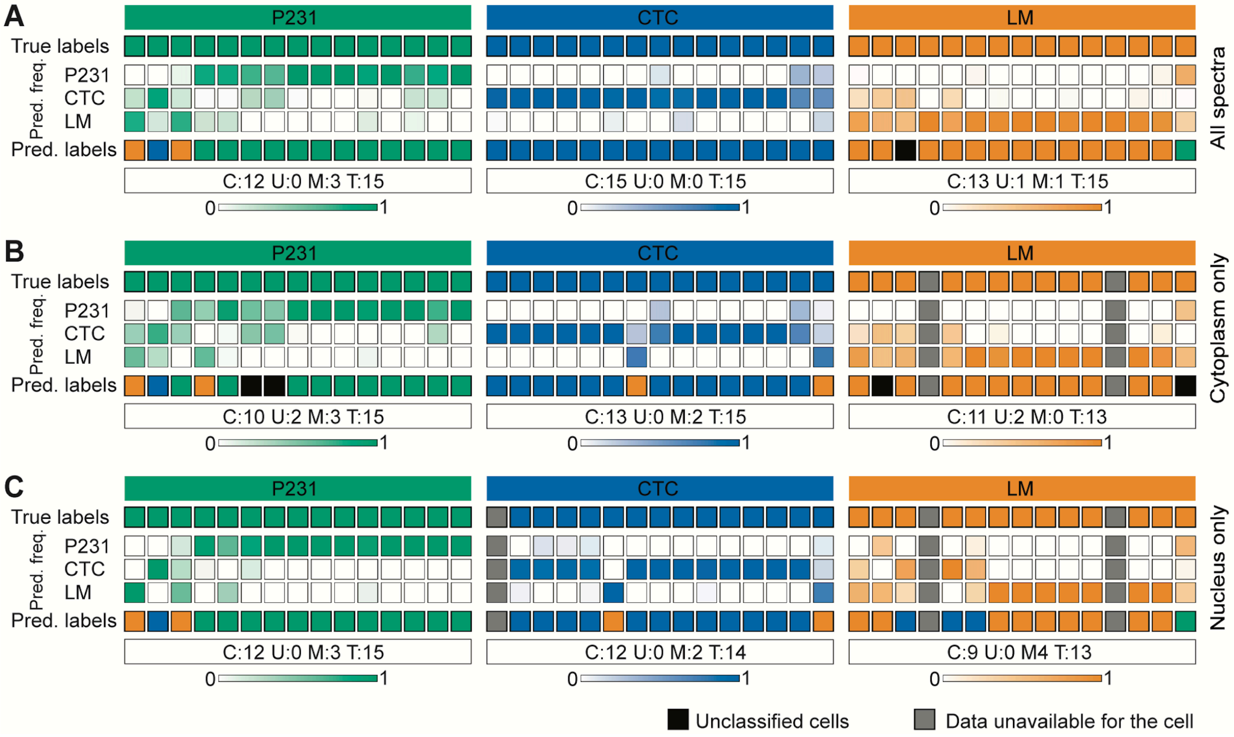

Figure 4. Leave-one-cell-out random forest classification of Raman images at a single-cell level. The leave-one-cell-out random forest predictions are shown for the multiclass classification task by including (**A**) all the spectra in the dataset, (**B**) spectra with high cytoplasm MC scores, and (**C**) spectra with high nucleus MC scores. Each column represents one unique cell, while the top and bottom rows respectively show the true and predicted class labels. The other rows show the normalized prediction frequencies (color bars at the bottom represent color scales) of spectra from each cell into the three classes. The classification results are summarized for each class to include the number of cells correctly classified (C), unclassified (U), and misclassified (M) out of the total (T) cells in the class.

We employed a multiclass random forest classifier to quantify our ability to classify P231, CTC, and LM cells based on the biochemical information encoded in their Raman spectra. Random forests are ensemble classifiers that employ a collection of decision trees constructed by random sampling of instances and variables in each tree to yield fast and generalizable models that are void of dependence on specific features or training instances [60]. Due to these characteristics and the ability to parallelize the tree construction, random forest classifiers are gaining attention in a variety of research areas including image classification and vibrational spectroscopy [61]. First, we subjected the entire spectral data consisting of spectra from all subcellular compartments (cytoplasm, nucleus, and mixed) to a leave-one-cell-out random forest classification task (as described in Methods). Briefly, we trained the classifier iteratively by leaving spectra from one cell at a time from the training dataset and subjecting it as a test dataset to the developed model. We observed satisfactory prediction performance across the three classes with only 4 misclassifications and 1 unclassification among 45 cells (**Fig. 4A**). The majority of misclassifications occurred between P231 and LM classes, while all CTC were classified accurately. This observation is in agreement with the similarity of both cytoplasm and nucleus MCR scores for P231 and LM classes and their deviation from CTC. The misclassifications of P231 cells can also be attributed to the presence of cells in the P231 group that have future propensity to intravasate into circulation and colonize lungs (i.e. future CTC and LM cells). Unlike most of the previous studies where the classification of spectra is done at a bulk level [23, 62, 63], the leave-one-cell-out analysis allowed us to demonstrate not only the ability to identify metastatic potential at a single-cell level but also the robustness of such classification by completely excluding representation of the test data from the training dataset.

While the leave-one-cell-out analysis provided an excellent prediction of phenotype for the cells in all the three classes, the use of the entire dataset comprised of spectra from different subcellular regions introduces substantial intra-class heterogeneity in the training dataset and may make prediction challenging. Such difficulty can be further exacerbated if the target phenotype changes are specifically guided by local molecular variations in particular regions, for example in the nucleus of cells treated with chemotherapeutic drugs. While there is no direct attribution of metastatic phenotypes observed in the isogenic panel employed in this study to specific compartments, we sought to train and test the random forest classifiers using subsets of the spectral dataset from the regions identified as cytoplasm and nucleus using MCR-ALS decomposition. The deviation of the observations from the baseline results obtained by subjecting the entire dataset will provide preliminary insights into specific localization of changes in subcellular regions that render the CTC and LM cells more metastatic in comparison to the parental P231 cells. First, we restricted our analysis to include only spectra that exhibit high scores for cytoplasm-like loading (MC1) in training and test datasets. The leave-one-cell-out analysis of the cytoplasm spectra (**Fig. 4B**) from the three classes yielded similar predictions with a slight improvement in the classification of LM cells, where a previously misclassified cell was now unclassified due to the relative increase of spectral classification into LM group. However, we found new unclassifications and misclassifications, respectively, in the P231 and CTC classes. Next, we performed the leave-one-cell-out analysis on the subset of dataset comprised only of spectra that show high scores of nucleus-like loading (MC2). We observed that the exclusion of cytoplasm spectra (**Fig. 4C**) resulted in the deterioration of performance in CTC and LM classes without affecting the P231 classification. Together, these results show that while the prediction of P231 and LM cells are primarily driven by the spectra acquired from the nucleus and cytoplasm respectively, the classification of CTC is more challenging and requires spectra from both regions. While these observations are preliminary and require further investigation in a larger cohort of cells, the results hint at the sufficiency of spectra from specific subcellular regions to predict subtle phenotypic differences associated with metastatic potential in closely related isogenic cells.

In conclusion, our label-free optical study revealed morphological and molecular differences among isogenic breast cancer cells of progressively increasing metastatic potential. Using 3D RI tomograms, we showed that the parental P231, circulating CTC, and lung metastatic LM cells showed subtle yet significant variations in morphology as assessed by area, aspect ratio, and cell dry mass. The observations were consistent with prior evidence of EMT and MET processes that guide the metastatic progression of these MDA-MB-231 breast cancer cells. To uniquely predict the metastatic potential of these cells with single-cell analytical resolution, we used Raman spectroscopic imaging to capture their biomolecular composition along with the spatial details. The use of MCR-ALS decomposition allowed better visualization and demarcation of the nucleus and cytoplasm despite the low resolution of the coarse Raman images. Finally, our random forest classification models incorporating a leave-one-cell-out strategy provided a route identification of subtle metastatic phenotype of cells at a single-cell level based on the coarse Raman maps. Further classification using the spectra individually from cytoplasm and nucleus regions as identified by MCR-ALS decomposition showed that specific subsets were sufficient for the identification of metastatic phenotypes. Taken together, these studies show that optical imaging and spectroscopy are sensitive to the differences in the cellular states guided by biological processes. We envision that coarse Raman imaging will be leveraged to build large spectral datasets from clinical samples that are amenable to machine learning analysis for determination of biomolecular phenotype/variant at a single-cell resolution to avoid loss of information associated with the population analyses. The imaging protocol and leave-one-cell-out random forest routine can readily be extended to the investigation of a variety of phenomena such as drug response, stem cell differentiation, and immune cell activation.

Acknowledgments

S.K.P. acknowledges the support of the SLAS Graduate Education Fellowship Grant. I.B. acknowledges the support from the National Cancer Institute (R01 CA238025), the National Institute of Biomedical Imaging and Bioengineering (2-P41-EB015871-31) and the National Institute of General Medical Sciences (DP2GM128198). K.G. acknowledges the support from the National Cancer Institute (R01 CA213428, R01 CA213492). The authors thank Tomocube Inc for use of the 3D ODT system. The schematic in Figure 1A was partially created with BioRender.com.

References

[1] C. L. Chaffer and R. A. Weinberg, A perspective on cancer cell metastasis, *Science* **331**(6024), 2011, 1559-1564.

[2] C. Liu, J. Ding, K. Spuhler, Y. Gao, M. Serrano Sosa, M. Moriarty, et al., Preoperative prediction of sentinel lymph node metastasis in breast cancer by radiomic signatures from dynamic contrast-enhanced MRI, *Journal of Magnetic Resonance Imaging* **49**(1), 2019, 131-140.

[3] C. L. Chaffer, B. P. San Juan, E. Lim and R. A. Weinberg, EMT, cell plasticity and metastasis, *Cancer Metastasis Rev* **35**(4), 2016, 645-654.

[4] Y. Hsiao, M. Chou, C. Fowler, J. T. Mason and Y. Man, Breast cancer heterogeneity: Mechanisms, proofs, and implications, *Journal of Cancer* **1**2010, 6.

[5] P. K. Brastianos, S. L. Carter, S. Santagata, D. P. Cahill, A. Taylor-Weiner, R. T. Jones, et al., Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets, *Cancer Discovery* **5**(11), 2015, 1164-1177.

[6] A. J. Minn, G. P. Gupta, P. M. Siegel, P. D. Bos, W. Shu, D. D. Giri, et al., Genes that mediate breast cancer metastasis to lung, *Nature* **436**(7050), 2005, 518-524.

[7] P. D. Bos, X. H. Zhang, C. Nadal, W. Shu, R. R. Gomis, D. X. Nguyen, et al., Genes that mediate breast cancer metastasis to the brain, *Nature* **459**(7249), 2009, 1005-1009.

[8] Z. Liu, S. J. Lee, S. Park, K. Konstantopoulos, K. Glunde, Y. Chen, et al., Cancer cells display increased migration and deformability in pace with metastatic progression, *The FASEB Journal* **34**(7), 2020, 9307-9315.

[9] A. Han, L. Yang and A. B. Frazier, Quantification of the heterogeneity in breast cancer cell lines using whole-cell impedance spectroscopy, *Clin Cancer Res* **13**(1), 2007, 139-143.

[10] C. L. Yankaskas, K. N. Thompson, C. D. Paul, M. I. Vitolo, P. Mistriotis, A. Mahendra, et al., A microfluidic assay for the quantification of the metastatic propensity of breast cancer specimens, *Nature Biomedical Engineering* **3**(6), 2019, 452-465.

[11] J. S. Jeon, S. Bersini, M. Gilardi, G. Dubini, J. L. Charest, M. Moretti, et al., Human 3D vascularized organotypic microfluidic assays to study breast cancer cell extravasation, *Proceedings of the National Academy of Sciences* **112**(1), 2015, 214-219.

[12] A. F. Sarioglu, N. Aceto, N. Kojic, M. C. Donaldson, M. Zeinali, B. Hamza, et al., A microfluidic device for label-free, physical capture of circulating tumor cell clusters, *Nature Methods* **12**(7), 2015, 685-691.

[13] Y. V. Ma, K. Middleton, L. You and Y. Sun, A review of microfluidic approaches for investigating cancer extravasation during metastasis, *Microsystems & Nanoengineering* **4**(1), 2018, 1-13.

[14] A. F. Sarioglu, N. Aceto, N. Kojic, M. C. Donaldson, M. Zeinali, B. Hamza, et al., A microfluidic device for label-free, physical capture of circulating tumor cell clusters, *Nature Methods* **12**(7), 2015, 685-691.

[15] J. Che, V. Yu, E. B. Garon, J. W. Goldman and D. Di Carlo, Biophysical isolation and identification of circulating tumor cells, *Lab on a Chip* **17**(8), 2017, 1452-1461.

[16] B. J. Green, T. Saberi Safaei, A. Mepham, M. Labib, R. M. Mohamadi and S. O. Kelley, Beyond the capture of circulating tumor cells: Next-□generation devices and materials, *Angewandte Chemie International Edition* **55**(4), 2016, 1252-1265.

[17] S. L. Stott, C. Hsu, D. I. Tsukrov, M. Yu, D. T. Miyamoto, B. A. Waltman, et al., Isolation of circulating tumor cells using a microvortex-generating herringbone-chip, *Proceedings of the National Academy of Sciences* **107**(43), 2010, 18392-18397.

[18] E. Ozkumur, A. M. Shah, J. C. Ciciliano, B. L. Emmink, D. T. Miyamoto, E. Brachtel, et al., Inertial focusing for tumor antigen–dependent and–independent sorting of rare circulating tumor cells, *Science Translational Medicine* **5**(179), 2013, 179ra47.

[19] M. Poudineh, P. M. Aldridge, S. Ahmed, B. J. Green, L. Kermanshah, V. Nguyen, et al., Tracking the dynamics of circulating tumour cell phenotypes using nanoparticle-mediated magnetic ranking, *Nature Nanotechnology* **12**(3), 2017, 274-281.

[20] M. Labib, R. M. Mohamadi, M. Poudineh, S. U. Ahmed, I. Ivanov, C. Huang, et al., Single-cell mRNA cytometry via sequence-specific nanoparticle clustering and trapping, *Nature Chemistry* **10**(5), 2018, 489.

[21] X. Hu, P. H. Bessette, J. Qian, C. D. Meinhart, P. S. Daugherty and H. T. Soh, Marker-specific sorting of rare cells using dielectrophoresis, *Proceedings of the National Academy of Sciences* **102**(44), 2005, 15757-15761.

[22] D. Gao, V. Mittal, Y. Ban, A. R. Lourenco, S. Yomtoubian and S. Lee, Metastatic tumor cells–genotypes and phenotypes, *Frontiers in Biology* **13**(4), 2018, 277-286.

[23] P. T. Winnard Jr, C. Zhang, F. Vesuna, J. W. Kang, J. Garry, R. R. Dasari, et al., Organ-specific isogenic metastatic breast cancer cell lines exhibit distinct Raman spectral signatures and metabolomes, *Oncotarget* **8**(12), 2017, 20266.

[24] S. Basu, S. Kolouri and G. K. Rohde, Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry, *Proceedings of the National Academy of Sciences* **111**(9), 2014, 3448-3453.

[25] Z. Di, M. J. Klop, V. Rogkoti, S. E. Le Dévédec, B. van de Water, F. J. Verbeek, et al., Ultra high content image analysis and phenotype profiling of 3D cultured micro-tissues, *PloS One* **9**(10), 2014, e109688.

[26] R. K. Chhetri, Z. F. Phillips, M. A. Troester and A. L. Oldenburg, Longitudinal study of mammary epithelial and fibroblast co-cultures using optical coherence tomography reveals morphological hallmarks of pre-malignancy, *PLoS One* **7**(11), 2012, e49148.

[27] V. K. Lam, T. Nguyen, T. Phan, B. Chung, G. Nehmetallah and C. B. Raub, Machine learning with optical phase signatures for phenotypic profiling of cell lines, *Cytometry* **95**(7), 2019, 757-768.

[28] S. H. Karandikar, C. Zhang, A. Meiyappan, I. Barman, C. Finck, P. K. Srivastava, et al., Reagent-free and rapid assessment of T cell activation state using diffraction phase microscopy and deep learning, *Anal Chem* **91**(5), 2019, 3405-3411.

[29] P. Hai, T. Imai, S. Xu, R. Zhang, R. L. Aft, J. Zou, et al., High-throughput, label-free, single-cell photoacoustic microscopy of intratumoral metabolic heterogeneity, *Nature Biomedical Engineering* **3**(5), 2019, 381-391.

[30] M. G. Kim, J. Park, H. G. Lim, S. Yoon, C. Lee, J. H. Chang, et al., Label-free analysis of the characteristics of a single cell trapped by acoustic tweezers, *Scientific Reports* **7**(1), 2017, 1-9.

[31] A. Rizwan, S. K. Paidi, C. Zheng, M. Cheng, I. Barman and K. Glunde, Mapping the genetic basis of breast microcalcifications and their role in metastasis, *Scientific Reports* **8**2018, 11067.

[32] A. Rizwan, C. Bulte, A. Kalaichelvan, M. Cheng, B. Krishnamachary, Z. M. Bhujwalla, et al., Metastatic breast cancer cells in lymph nodes increase nodal collagen density, *Scientific Reports* **5**2015, 10002.

[33] W. Choi, C. Fang-Yen, K. Badizadegan, S. Oh, N. Lue, R. R. Dasari, et al., Tomographic phase microscopy, *Nature Methods* **4**(9), 2007, 717-719.

[34] K. Kim, K. S. Kim, H. Park, J. C. Ye and Y. Park, Real-time visualization of 3-D dynamic microscopic objects using optical diffraction tomography, *Optics Express* **21**(26), 2013, 32269-32278.

[35] Y. Sung, W. Choi, C. Fang-Yen, K. Badizadegan, R. R. Dasari and M. S. Feld, Optical diffraction tomography for high resolution live cell imaging, *Optics Express* **17**(1), 2009, 266-277.

[36] G. Popescu, Y. Park, N. Lue, C. Best-Popescu, L. Deflores, R. R. Dasari, et al., Optical imaging of cell mass and growth dynamics, *American Journal of Physiology-Cell Physiology* **295**(2), 2008, C538-C544.

[37] S. Lee, H. Park, K. Kim, Y. Sohn, S. Jang and Y. Park, Refractive index tomograms and dynamic membrane fluctuations of red blood cells from patients with diabetes mellitus, *Scientific Reports* **7**(1), 2017, 1-11.

[38] Y. Park, C. Depeursinge and G. Popescu, Quantitative phase imaging in biomedicine, *Nature Photonics* **12**(10), 2018, 578-589.

[39] Y. Jo, H. Cho, S. Y. Lee, G. Choi, G. Kim, H. Min, et al., Quantitative phase imaging and artificial intelligence: A review, *IEEE Journal of Selected Topics in Quantum Electronics* **25**(1), 2018, 1-14.

[40] K. Kong, C. Kendall, N. Stone and I. Notingher, Raman spectroscopy for medical diagnostics — from in-vitro biofluid assays to in-vivo cancer detection, *Adv Drug Deliv Rev* **89**2015, 121-134.

[41] S. K. Paidi, R. Pandey and I. Barman, Medical applications of Raman spectroscopy, *Encyclopedia of Analytical Chemistry* 2020, 1-21.

[42] N. Stone, C. Kendall, N. Shepherd, P. Crow and H. Barr, Near-infrared Raman spectroscopy for the classification of epithelial pre-cancers and cancers, *J Raman Spectrosc* **33**(7), 2002, 564-573.

[43] B. Kann, H. L. Offerhaus, M. Windbergs and C. Otto, Raman microscopy for cellular investigations — from single cell imaging to drug carrier uptake visualization, *Adv Drug Deliv Rev* **89**2015, 71-90.

[44] S. K. Paidi, A. Rizwan, C. Zheng, M. Cheng, K. Glunde and I. Barman, Label-free Raman spectroscopy detects stromal adaptations in premetastatic lungs primed by breast cancer, *Cancer Res* **77**(2), 2017, 247-256.

[45] S. K. Paidi, P. M. Diaz, S. Dadgar, S. V. Jenkins, C. M. Quick, R. J. Griffin, et al., Label-free Raman spectroscopy reveals signatures of radiation resistance in the tumor microenvironment, *Cancer Res* **79**(8), 2019, 2054-2064.

[46] J. W. Kang, P. T. So, R. R. Dasari and D. Lim, High resolution live cell Raman imaging using subcellular organelle-targeting SERS-sensitive gold nanoparticles with highly narrow intra-nanogap, *Nano Letters* **15**(3), 2015, 1766-1772.

[47] W. Xu, S. K. Paidi, Z. Qin, Q. Huang, C. Yu, J. V. Pagaduan, et al., Self-folding hybrid graphene skin for 3D biosensing, *Nano Letters* **19**(3), 2018, 1409-1417.

[48] K. Hamada, K. Fujita, N. I. Smith, M. Kobayashi, Y. Inouye and S. Kawata, Raman microscopy for dynamic molecular imaging of living cells, *J Biomed Opt* **13**(4), 2008, 044027.

[49] M. Okada, N. I. Smith, A. F. Palonpon, H. Endo, S. Kawata, M. Sodeoka, et al., Label-free Raman observation of cytochrome c dynamics during apoptosis, *Proceedings of the National Academy of Sciences* **109**(1), 2012, 28-32.

[50] S. F. El-Mashtoly, H. K. Yosef, D. Petersen, L. Mavarani, A. Maghnouj, S. Hahn, et al., Label-free Raman spectroscopic imaging monitors the integral physiologically relevant drug responses in cancer cells, *Anal Chem* **87**(14), 2015, 7297-7304.

[51] S. Shin, K. Kim, J. Yoon and Y. Park, Active illumination using a digital micromirror device for quantitative phase imaging, *Opt Lett* **40**(22), 2015, 5407-5410.

[52] K. Kim, H. Yoon, M. Diez-Silva, M. Dao, R. R. Dasari and Y. Park, High-resolution three-dimensional imaging of red blood cells parasitized by plasmodium falciparum and in situ hemozoin crystals using optical diffraction tomography, *J Biomed Opt* **19**(1), 2013, 011005.

[53] C. McQuin, A. Goodman, V. Chernyshev, L. Kamentsky, B. A. Cimini, K. W. Karhohs, et al., CellProfiler 3.0: Next-generation image processing for biology, *PLoS Biology* **16**(7), 2018, e2005970.

[54] N. Otsu, A threshold selection method from gray-level histograms, *IEEE Trans Syst Man Cybern* **9**(1), 1979, 62-66.

[55] P. Wu, J. M. Phillip, S. B. Khatau, W. Chen, J. Stirman, S. Rosseel, et al., Evolution of cellular morpho-phenotypes in cancer metastasis, *Scientific Reports* **5**(1), 2015, 1-10.

[56] K. G. Phillips, S. L. Jacques and O. J. T. McCarty, Measurement of single cell refractive index, dry mass, volume, and density using a transillumination microscope, *Phys Rev Lett* **109**(11), 2012, 118105.

[57] J. Felten, H. Hall, J. Jaumot, R. Tauler, A. De Juan and A. Gorzsás, Vibrational spectroscopic image analysis of biological material using multivariate curve resolution–alternating least squares (MCR-ALS), *Nature Protocols* **10**(2), 2015, 217.

[58] W. J. Choi, D. I. Jeon, S. Ahn, J. Yoon, S. Kim and B. H. Lee, Full-field optical coherence microscopy for identifying live cancer cells by quantitative measurement of refractive index distribution, *Optics Express* **18**(22), 2010, 23285-23295.

[59] Z. Movasaghi, S. Rehman and I. U. Rehman, Raman spectroscopy of biological tissues, *Applied Spectroscopy Reviews* **42**(5), 2007, 493-541.

[60] A. Parmar, R. Katariya and V. Patel, A review on random forest: An ensemble classifier, *International Conference on Intelligent Data Communication Technologies and Internet of Things,* Springer, 2018, 758-763.

[61] S. Mittal, K. Yeh, L. S. Leslie, S. Kenkel, A. Kajdacsy-Balla and R. Bhargava, Simultaneous cancer and tumor microenvironment subtyping using confocal infrared microscopy for all-digital molecular histopathology, *Proceedings of the National Academy of Sciences* **115**(25), 2018, E5651-E5660.

[62] S. Qiu, Y. Weng, Y. Li, Y. Chen, Y. Pan, J. Liu, et al., Raman profile alterations of irradiated human nasopharyngeal cancer cells detected with laser tweezer Raman spectroscopy, *RSC Advances* **10**(24), 2020, 14368-14373.

[63] M. Marro, C. Nieva, R. Sanz-Pamplona and A. Sierra, Molecular monitoring of epithelial-to-mesenchymal transition in breast cancer cells by means of Raman spectroscopy, *Biochimica Et Biophysica Acta (BBA)-Molecular Cell Research* **1843**(9), 2014, 1785-1795.