UNIVERSITY OF CALIFORNIA SAN DIEGO

Identifying drivers of phenotype heterogeneity in breast cancer

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Bioengineering

by

Kevin Chen

Committee in charge:

Professor Stephanie I. Fraley, Chair
Professor David Cheresh
Professor Adam J. Engler
Professor Trey Ideker
Professor Christian Metallo

2020

The dissertation of Kevin Chen is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2020

DEDICATION

For my family and my friends, who have supported me through this incredible journey.

EPIGRAPH

Nothing in life is to be feared, it is only to be understood.
Now is the time to understand more, so that we may fear less.

– Marie Curie

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Dr. Stephanie Fraley, for being a continuous source of guidance, encouragement, and support throughout my PhD. I found our conversations and interactions to be immensely valuable, as I have learned much from her unique ways of interpreting, processing, and presenting data.

I would also like to thank my fellow lab members for always being supportive and encouraging. Our brainstorming and discussion sections keep me energized and excited through my scientific pursuits.

The contents of this dissertation contain partial reprints of manuscripts being submitted to journals for publication. The dissertation author was the primary investigator and author of all papers. The following chapters contain material that have been submitted for publication as it may appear in iScience, 2020, Chen, Kevin; Ozturk, Kivilcim; Contreras, Ryne L; Simon, Jessica; McCann, Sean; Chen, Wei Ji; Carter, Hannah; Fraley, Stephanie I, Cell Press, 2020. The following chapters also contain material currently being prepared for submission for publication. Chen, Kevin; Ozturk, Kivilcim; Liefeld, Ted; Reich, Michael; Mesirov, Jill; Carter, Hannah; Fraley, Stephanie I.

VITA

2016	B. S. in Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD
2020	Ph.D. in Bioengineering, University of California San Diego, San Diego, CA

PUBLICATIONS

K. Chen, K. Ozturk, R. L. Contreras, J. Simon, S. McCann, W. J. Chen, H. Carter, S. I. Fraley. Phenotypically supervised single cell sequencing parses within-cell-type-heterogeneity. *iScience*. (2020) (under review).

K. Chen, K. Ozturk, T. Liefeld, M. Reich, J. Mesirov, H. Carter, S. I. Fraley. Parsing within-cell type heterogeneity using phenotypically supervised single cell analysis. Nature Protocols (2020) (in submission)

F. Zheng, B. Tutuncuoglu, K. Ono, D. L. Swaney, M. Kim, E. Silva, S. Liu, J. Park, A. Kratz, M. K. Yu, **K. Chen**, J. Chen, C. Churas, R. Pillich, P. N. Devin, K. Licon, J. F. Kreisberg, D. Pratt, J. S. Gutkind, S. I. Fraley, N. Krogan, T. Ideker. Convergence of cancer mutations on a hierarchy of protein systems. Cell (2020) (under review)

P. Beri, A. Popravko, B. Yeoman, A. Kumar, **K. Chen**, E. Hodzic, A. Chiang, A. Banisadr, J. K. Placone, H. Carter, S. I. Fraley, P. Katira, A. J. Engler, Cell Adhesiveness Serves as a Biophysical Marker for Metastatic Potential. *Cancer Res.* **80**, 901–911 (2020).

W. Zhang, A. Bojorquez-Gomez, D. O. Velez, G. Xu, K. S. Sanchez, J. P. Shen, **K. Chen**, K. Licon, C. Melton, K. M. Olson, M. K. Yu, J. K. Huang, H. Carter, E. K. Farley, M. Snyder, S. I. Fraley, J. F. Kreisberg, T. Ideker, A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet.* **50**, 613–620 (2018).

A. Aralar, Y. Yuan, **K. Chen**, Y. Geng, D. Ortiz Velez, M. Sinha, S. M. Lawrence, S. I. Fraley, Improving quantitative power in digital PCR through digital high-resolution melting. *J. Clin. Microbiol.* **58** (2020), doi:10.1128/JCM.00325-20.

ABSTRACT OF THE DISSERTATION

Identifying drivers of phenotype heterogeneity in breast cancer

by

Kevin Chen

Doctor of Philosophy in Bioengineering

University of California San Diego, 2020

Professor Stephanie I. Fraley, Chair

A major component of cancer's complexity lies in its heterogeneity. Because cancer heterogeneity can manifest across multiple spatiotemporal and biological scales(*1–4*), comprehensive characterization is challenging. Assays that allow controlled experimentation to determine the mechanisms behind cancer heterogeneity remain underdeveloped. Many existing technologies rely on exploitation of predetermined characteristics(*5–7*), which precludes exploration of phenotypes that are ill-defined. Other platforms begin with blind genomics, using post-sequencing experiments for validation(*8, 9*). These approaches inherently require significant differences in the genomics underlying heterogeneity to elucidate potential

mechanisms, since unsupervised clustering relies on deploying mathematics to obtain clear separations(*10–12*). While successful when involving multiple cell types, since their biological profiles lie in separable state spaces(*6*, *8*, *13*, *14*), unsupervised analysis has limitations when evaluating more similar cells, such as within-cell-type heterogeneity, as it becomes difficult to separate profiles that are highly alike. A study on this specialized scope has not yet been done, probably due to the challenges of parsing a more subtly heterogeneous sample.

This dissertation describes the development and application of a platform that enables detailed interrogation of within-cell-type heterogeneous, 3D collective cancer cell migration phenotypes. Collective migration is a process where multiple cells coordinate their movement(*15*, *16*). Recent studies point to the importance of collective cell migration in cancer metastasis(*17–20*). Although various phenotypes have been identified, factors that regulate collective behavior are not fully understood(*15*, *16*, *21*). Using a flexible, microscopy-based platform, I identified and isolated cells that exhibit invasive and non-invasive collective migration. By first defining functional subpopulations, I deployed a supervised approach to determine the mechanisms behind this aspect of heterogeneity. I found that the collectively invasive phenotype is associated with upregulated proliferation, increased stress responses, and the ductal carcinoma subtype, while the collectively non-invasive phenotype was associated with immune-related processes and the luminal carcinoma subtype. Functional perturbation of differentially expressed genes resulted in shifts in migration phenotypes. These results validate the platform I developed for identifying mechanisms of within-cell-type heterogeneity. Furthermore, the results demonstrate a link between migration regulation, stress response, proliferation, and immune response and indicate potential value in exploring how collective invasion may be controlled through these associated modules.

Chapter 1: Development of an automated, high-throughput platform to separate cancer cells based on visual heterogeneity

1.1 Introduction

Tumor heterogeneity is one of the major features of cancer that precludes effective treatment. Since standard treatment plans target a particular receptor or pathway, the existence of subclones presents opportunities for resistance and complicates therapeutic strategy(*22–24*). Experimental approaches for the identification of regulatory factors that modulate tumor heterogeneity can therefore importantly contribute to the design of therapeutic agents for the treatment and/or prevention of cancer.

Advances in sequencing technologies have helped elucidate information on the genetic mechanisms of tumor heterogeneity through evolutionary and cellular differentiation models(*25*). However, functional characterization of tumor heterogeneity remains lacking, and approaches linking omics data to functional outputs remain underdeveloped(*26*).

The elucidation of the exact pathways that result in the manifestation of heterogeneity in cancer has been hampered by the lack of suitable methodology to functionally interrogate the putative roles of protein regulators. In principle, studying the differences between heterogeneous profiles requires isolation of these subpopulations for experimental comparison. This separation is non-trivial, especially for phenotypes that are not well characterized. This problem is further complicated by the rise of 3D culture models, which aim to better recapitulate native physiology. Extraction of cells embedded in 3D culture presents additional challenges. Isolation of cells from the culture platform is not a trivial task, particularly when only specific and possibly rare phenotypes are of interest. Furthermore, these methods need to be designed to minimize

disturbance so that the native biology is preserved after isolation, yet rapid enough to attain the throughput needed for large-scale studies. Additionally, there is a need for this technique to be widely applicable and easy to use.

In recent advancements, transcriptome measurements have been integrated with multiple omics, genotype, cell electrophysiology, lineage tracing, and spatial information(*27–32*). In other protocols, rare cell subpopulations are functionally sorted using physical or image-guided techniques in an attempt to link phenotype to omics data(*5*, *33*). However, these approaches are often limited in their capacity to explore undefined phenotypes that have no known biomarkers. In other cases, the technique can directly tackle particular phenotypes but is highly technical and specialized, resulting in low throughput when isolating subpopulations of interest. This results in a need to amplify the cell population after collection to obtain enough genetic material for downstream experiments, which presents problems when investigating phenomena that may change between cell cycles.

To address the need to develop a quick, flexible, high-throughput platform for interrogating cellular heterogeneity, I have developed a method that enables fluorescent tagging of groups of cells within 3D culture. I can then release the cells from 3D culture by gel digestion and collect the targeted population using fluorescence activated cell sorting (FACS) while preserving their viability. This platform enables to collection of thousands of cells over the course of a few hours and is widely adaptable for use in interrogating any visual manifestation of heterogeneity.

1.2 Results

1.2.1 Dendra2 photoconversion enables selective phototagging of cells in 3D culture

As the basis for our technique, I virally transduced Dendra2, a protein known to exhibit green-to-red photoconversion, into our cells. Post transduction, cells were sorted for expression and maintained a baseline green fluorescence in culture. To obtain cells with the highest Dendra2 signal, we sorted for the 15% of cells with the highest green fluorescence. We outfitted our microscope with a laser box and galvanometer scanner to enable fine spatial control of a 405 nm laser beam. Exposure of cells to 405 nm light at 50% laser power and 30 µs dwell time caused the cells of interest to express red fluorescence, while untargeted cells only expressed green fluorescence.

We then sought proof of concept in 3D culture. Cells were embedded in 3D collagen, and then exposed to 405 nm light. We tried various different laser power strengths and dwell times to optimize the exposure needed for photoconversion while minimizing phototoxicity effects. In the x-y plane, we had no observation of unintended photoconversion. A laser power of 70% and a dwell time of 30 µs was sufficient in producing a distinctive red fluorescence post-exposure (Figure 1.1). In the z-axis, cells directly underneath the target cell would exhibit photoconversion if the distance was less than 200 µm. Thus, to minimize unintended phototagging, we avoided targeting cells that containing overlapping cells in the z-axis less than 300 µm apart.

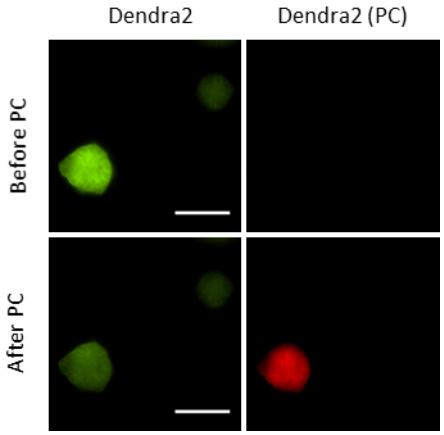


Figure 1.1. Cells transduced with Dendra2 are able to be selectively phototagged. Representative FITC and TRITC images of cells transduced with Dendra2 pre and post stimulation. Only the targeted cells (bottom left in field of view) were phototagged and expressing red fluorescence (photoconverted Dendra2).

1.2.2 Collagenase and trypsin treatments release the cells from the matrix and preserves the photoconverted signature

Post-phototagging, cells then needed to be isolated from 3D culture. Treatment of the gels with 1 mg of collagenase per mg collagen for 15 minutes, aided with mechanical disruption, caused the gel to fully dissolve. The solution was then diluted with cell culture media and centrifuged at 400 xg to pellet the cells. The supernatant was aspirated to remove the collagenase, and the cells were resuspended in trypsin to dissociate them into a single cell suspension for 10 minutes. Post digestion, the cells remain viable and retain red fluorescence imparted by phototagging (Figure 1.2).

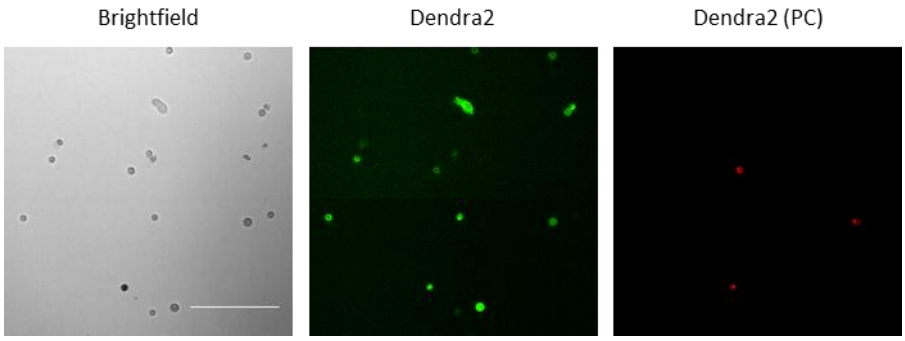


Figure 1.2. BRCA cells remain viable and phototagged post gel digestion. Representative brightfield, FITC, and TRITC image of MDA-MB-231 cells after extraction from a 3D collagen gel. Almost all cells express green fluorescence (Dendra2), while only a select few are expressing red fluorescence from phototagging (Dendra2 (PC)). Scale bar 200 µm.

1.2.3 Fluorescence activated cell sorting (FACS) enables purification of phototagged cells

Post-gel digestion, cells then needed to be purified to isolate the cells that were phototagged. To preserve viability and the transcriptome state, cells were sorted in a chilled environment. Cells were resuspended in a buffer containing BSA and EDTA to minimize aggregation and sorted at a low flow to minimize stress. Sorting the cells based on red fluorescence enables capture of the phototagged cells (Figure 1.3). Post FACS, the cells remain viable and all cells retain red fluorescence imparted by phototagging (Figure 1.4).

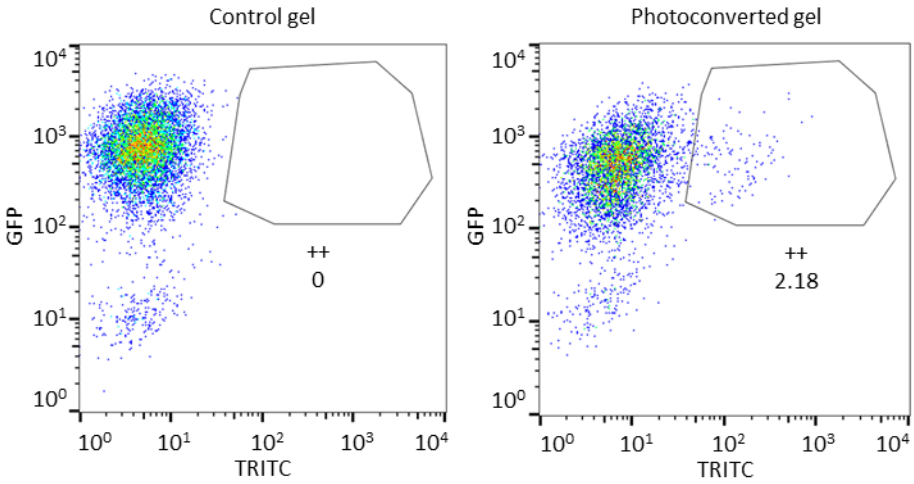


Figure 1.3. FACS enables separation and recovery of phototagged cells. The fluorescent profile of a control gel (not photoconverted) is used to gate the red fluorescence level. Cells in the photoconverted gel are then collected if they fall within the gate. A fraction of cells exhibits greater red fluorescence compared to the control.

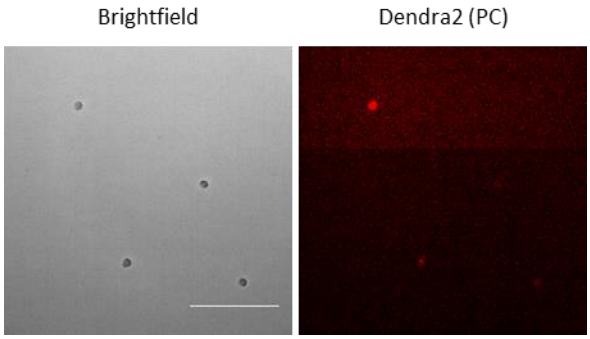


Figure 1.4. BRCA cells remain viable and phototagged post FACS. Representative brightfield and TRITC image of MDA-MB-231 cells after sorting via FACS. All cells express red fluorescence from phototagging (Dendra2 (PC)) and contain a pure population of the phototagged cells. Scale bar 200 µm.

1.3 Discussion

1.3.1 Development of the protocol

We developed a workflow that enables investigation of molecular mechanisms of heterogeneity in cancer through transcriptomic sequencing on functionally sorted populations. While adaptable to many contexts, we chose to focus on isolating collectively invasive cancer cells and collectively non-invasive cancer cells that spontaneously form in our 3D culture model(*21*,

34). This method is based on the fluorescent tagging of the subpopulation of interest through photoconversion. Dendra2 was chosen as the photoconvertible protein, which can be transduced into any cell line, rendering them green-fluorescent. Upon exposure to 405 nm light, cells will start to become red-fluorescent. Dendra2 was originally used to track intracellular protein movement, but we found that it could also be used as a general fluorescent tag to mark cells within 3D culture, enabling its use in any system that is amenable to transduction or transfection.

Combining this fluorescent tagging technique with gel digestion and FACS, we were able to isolate out populations of cells that exhibit invasive and non-invasive modes of collective cancer migration. This approach of using an agnostic fluorescent tag to perform phenogenomic sequencing is a powerful way to link functional output to transcriptomic data without the need for prior knowledge on biomarkers that define heterogeneous profiles.

Accordingly, the protocol presented here can be used for high-throughput separation of phenotypically heterogeneous cells in 3D culture based on any visual indicator. We describe a photoconversion-based platform that enables fluorescent tagging of visual phenotypes based on morphological characteristics. We also attach a streamlined 3D culture method, gel digestion and FACS protocol.

1.3.2 Comparison with other methods

Current methods developed for phenogenomic sequencing often require more stringent and less flexible methods of identifying the phenotype of interest(*28–30*). The application of these techniques are limited to situations where established biomarkers already define the subpopulation to be isolated. Other agnostic labeling methods using confocal microscopy provide more flexibility but are often limited in throughput and scale(*33*). The amount of cells that can be tagged typically range in the tens or hundreds, and the time required to tag the cells at such a fine resolution

precludes high throughput, making these techniques unsuitable for studies above the single cell level. These methods often require growth-based amplification after sorting to obtain enough genetic material for sequencing, which is problematic if cells are plastic and can redefine their transcriptome between sorting and extraction. Furthermore, no published method has yet described a tagging platform based on flexible recognition of heterogeneous visual phenotypes.

1.3.3 Applications of the method

The presented methodology enables researchers to interrogate the molecular basis of biological heterogeneity that can be defined by any visual parameter. Although this technique was used to probe heterogeneity in the collective migration profiles of MDA-MB-231 cells, other cell lines and other aspects of biological heterogeneity can also be investigated, provided the heterogeneity manifests in a visual manner. Because our fluorescent tagging method is based on transduction of Dendra2, any cell amenable to viral transduction can be used with our method. Combined with downstream FACS and transcriptome sequencing, an insight into the mechanisms of the heterogeneous property in question can be answered at the transcriptome level. This can then lead to functional perturbations at the protein level to confirm the inferences made from sequencing outputs. Our platform can also be adapted to probe gene expression perturbation strategies downstream to investigate their impact on the biological heterogeneity in question.

1.3.4 Limitations

To increase throughput, our platform uses a wide-field fluorescent microscope at lower magnification to photoconvert a larger amount of cells with a shorter amount of time compared to confocal microscopy. However, this loss in resolution could potentially lead to unintended photoconversion of nearby cells, particularly above and below the plane where the targeted cells

lie. This limitation can be overcome by avoiding photoconversion of cells that contain nearby neighbors, and to more sparsely seed in 3D culture to minimize overlap of cells in the vertical axis. With our system, we calculated that objects farther than 10 μm in x-y and farther than 200 μm in z do not receive sufficient exposure to 405 nm light to be photoconverted. We also validated this experimentally.

Separation of the photoconverted cell population requires the use of collagenase and trypsin, which may destroy surface markers. If surface proteins need to be preserved for downstream applications, enzymes with less disruptive mechanisms can be used, or more sophisticated techniques, such as laser microdissection, may need to be used.

The protocol as presented requires manual phototagging of phenotypes, hampering throughput. I have already developed preliminary image acquisition and processing algorithms to automate the phototagging process. I can obtain volumetric images and identify spheroidal collective cell structures within the 3D gel and draw a proper ROI for stimulation. Future work will focus on integrating the communication with the laser unit to allow for stimulation, as well as developing additional image processing algorithms to expand on phenotypes available for targeting.

1.3.5 Experimental Design

The experimental workflow of the procedures is described in Figure 1.5, which depicts the different stages of the Procedure. In broad terms, the Procedure consists of four main sections, which are detailed below.

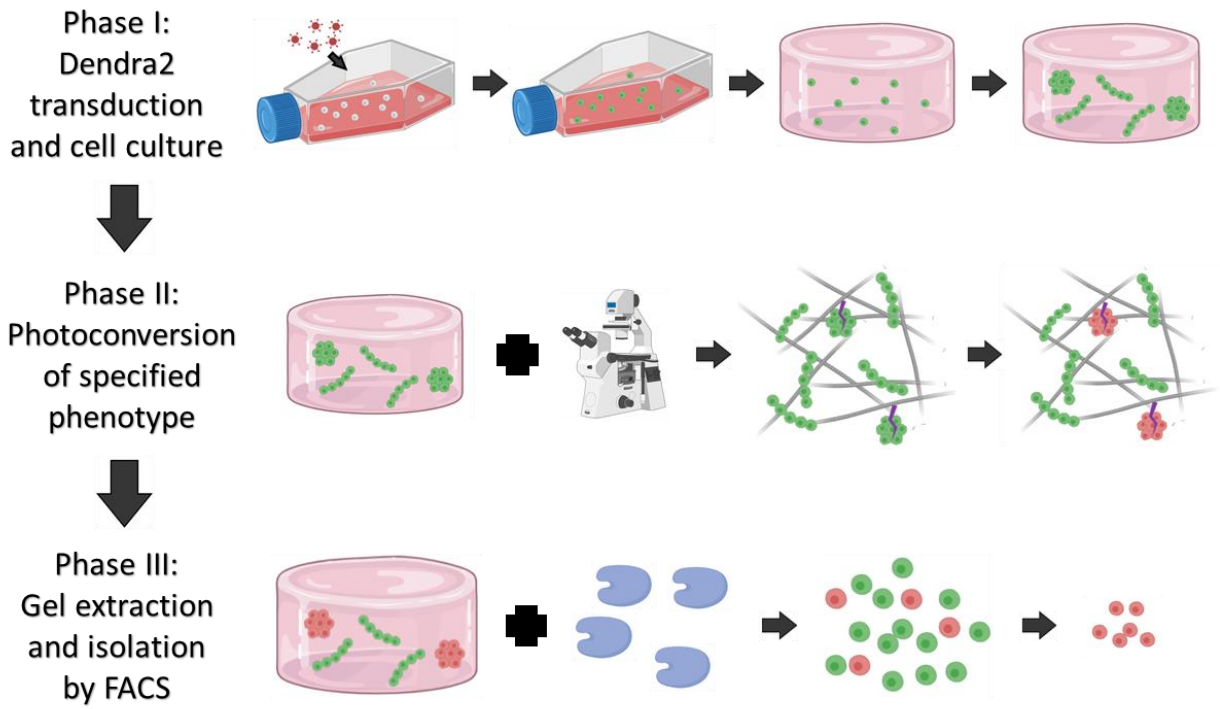


Figure 1.5. Experimental workflow of the phenotypic sorting platform. Cells are first transduced with Dendra2 and cultured in a 3D collagen matrix, where they are grown for 7 days to enable the development of collective migration phenotypes. The gel is then transferred to a microscope, where select phenotypes are identified in 3D space and photoconverted. The gel is then digested using collagenase and cells are recovered using FACS.

Choice of cell line

Successful transduction of a photoconvertible protein into the experimental cell line is the foundation of this technique. Thus, selection of a cell line amenable to gene editing is critical. While we present a method of lentiviral transduction to induce expression of Dendra2 in our protocol, other methods of induced gene expression and other photoconvertible proteins may be used as well.

Culturing platform

The presented protocol separates cells expressing heterogeneous collective migration profiles within 3D culture, specifically in 3D COLI. While heterogeneity can be investigated in a variety of contexts, there are specific considerations to keep in mind for 3D cultures to maintain an appropriate context. The biophysical properties of the environment, including choice of

material, should be adapted according to the type of heterogeneity being studied. Tissue specific studies, for instance, may need to be matched with particular extracellular matrix proteins at specific densities and stiffnesses to ensure biological relevance.

Fluorescent tagging by photoconversion

The choice of imaging strategy is not limited to the specific microscopes and settings presented in this protocol. However, for optimal configuration, a widefield microscope should be equipped with at least a laser line at wavelengths suitable for photoconversion, a galvanometric scanner, and climate control (temperature, humidity, and CO2). The code and macros we provide are specifically optimized for image sequences obtained using our microscope setup and software. However, this can be easily adapted to other platforms. The choice of lenses should be optimized to focus the laser light as much as possible to minimize off-target photoconversion, while also taking in consideration the physical size of the attributes used to determine the modes of heterogeneity being studied and maximizing throughput (subpopulation photoconversion).

Preparation for FACS

Proper digestion and preparation of cells into a single cell suspension is crucial for the efficient collection of the fluorescently tagged population. We have minimized the processing time to preserve the transcriptional signature as much as possible. In our case, since we culture our cells in a Col I hydrogel, we chose to use collagenase as our gel digestion enzyme. 3D cultures using other materials should use their respective appropriate enzyme(s) and be optimized for a short processing time while minimizing adverse effects on the cells. Since we study collective phenotypes, we also had an incubation phase with trypsin followed by straining to further dissociate the cells. This may not be necessary in other cases where cells are less adherent to each other after gel digestion. In addition, if preservation of cell surface markers are important for the

study, the choice of digestive enzymes to be used will need to be carefully selected to prevent loss during processing. We describe a standard buffer for FACS that was amenable to the survival of our cells. Cells that cannot survive in this basic flow sorting buffer may require other supplements or growth media while sorting.

The flow cytometer to be used must have the proper lasers and filters to detect the emission spectrum of the photoconvertible protein, in its native and photoconverted state. The equipment we use allows for stringent gating to ensure the collection of a phenotypically pure population. Users can adjust the gate depending on the stringency of their experiment. The flow sorter we use comes equipped with liquid chilling to help preserve our sample, although this may not be necessary in all cases.

1.3.6 Controls

It is important to verify that the biological heterogeneity being studied is not altered through expression of the photoconvertible protein. Cells expressing the photoconvertible protein should be compared to wild-type cells to ensure the same modes of heterogeneity exist and at similar frequencies or modes.

Proper controls must also be used to ensure collection of purified photoconverted populations. A non-photoconverted sample exhibiting biological heterogeneity should be used to gate the baseline fluorescence of the photoconverted channel. To ensure purity of the sample, a strict gate should be applied where all cells collected express higher fluorescence in the photoconverted channel compared to the negative control. Users can adjust the strictness of this gate to their application on the demands for the purity of their enriched population.

1.4 Materials and Methods

1.4.1 Biological materials

- MDA-MB-231 (ATCC, cat. no. HTB-26)
- Lenti-X 293T (Takara Bio, cat. no. 632180)

1.4.2 Reagents

Cell culture

- Dulbecco's Modified Eagle Medium (Life Technologies, cat. no. 11995073)
- Fetal bovine serum (Fisher Scientific, cat. no. MT35010CV)
- Phosphate-Buffered Saline (Life Technologies, cat. no. 10010-031)
- Gentamicin (Life Technologies, cat. no. 15750060)
- Trypsin (Life Technologies, cat. no. 25200056)

3D Culture

- Rat tail collagen I (Fisher Scientific, cat. no. CB354249)
- Sodium Hydroxide (Fisher Scientific, cat. no. S318-500)
- HEPES, Free Acid (Millipore Sigma, cat. no. 5310-OP)
- Sodium Bicarbonate (MP Biomedicals, cat. no. 02119484783)
- Polyethylene Glycol (Sigma-Aldrich, cat. no. P5413-500G)
- Wet Ice

Viral Particles Generation and Viral Transduction

- Dendra-2-Lifeact-7 plasmid (Addgene, cat. no. 54694)
- pSin-EF2-Nanog-Pur plasmid (Addgene, cat. no. 16578)
- ECORI (NEB, cat. no. R0101S)

- SPEI (NEB, cat. no. R0133S)
- Forward Primer (5’ TAAGCAACTAGTGGTTTAGTGAACCGTCAGA 3’, IDT)
- Reverse Primer (5’ GGTGCTTAGAATTCGTAAAACCTCTACAAATGTGG 3’, IDT)
- Kanamycin Sulfate (Fisher Scientific, cat. no. 11-845-024)
- Tryptone (Neogen, cat. no. NCM0211A)
- Nutrient Agar (Neogen, cat. no. NCM0269)
- Ampicillin (Fisher Scientific, cat. no. BP1760-5)
- Bacto™ Yeast Extract (ThermoFisher, cat. no. 288620)
- Sodium Chloride (Promega, cat. no. H5271)
- Viral packaging/envelope plasmids (Addgene, cat. no. 12260, 12259)
- Chemically competent DH5a
- Lipofectamine 3000 kit (ThermoFisher, cat. no. L3000008)
- 0.45 µm sterile filter (VWR, cat. no. 28137-938)

Flow Cytometry

- Collagenase (Sigma-Aldrich, cat. no. C0130)
- Nanopure water
- Bovine Serum Albumin (Fisher Scientific, cat. no. BP671-10)
- EDTA (BioPioneer, cat. no. MB1010)

Single cell RNAseq

- Chromium Chip B Single Cell Kit (10x Genomics, cat. no. 1000154)
- Chromium i7 Multiplex Kit (10x Genomics, cat. no. 120262)
- Chromium Single Cell 3’ Library & Gel Bead Kit v3 (10x Genomics, cat. no. 1000092)
- Chromium Single Cell 3’ Library Construction Kit v3 (10x Genomics, cat. no. 1000078)

1.4.3 Equipment

Culture

- Inverted contrast microscope for routine examination of cell cultures (Nikon DIAPHOT)
- Water bath at 37°C (Fisher Scientific, cat. no. 15-462-20Q)
- Incubator at 37°C with 5% CO2
- Centrifuge for 15-ml, 50-mL and Eppendorf tubes (ThermoFisher, cat. no. 750072003)
- Centrifuge tubes (50 mL, sterile, DNAse/RNase free, Corning, cat. no. 430829)
- Centrifuge tubes (15 mL, sterile, DNAse/RNase free, Corning, cat. no. 352095)
- Filtered sterile pipette tips (Neptune Scientific, cat. no. BT10E, BT20, BT200, BT1250)
- Micropipettes (Fisher Scientific, cat. no. 13-675-48, 13-675-49, 13-675-51, 13-675-52)
- Multiple-well plates (6 wells; Corning, cat. no. 353224)
- Multiple-well plates (48 wells; VWR, cat. no. 734-2326)
- Tissue Culture flasks (T75; VWR, cat. no. 734-2313)

Photoconversion

- 35 mm glass cover dish (WPI, cat. no. FD35-100)
- Inverted widefield fluorescence microscope (Nikon, model no. Eclipse Ti-E (MEA53100)) equipped with
- 20x Objective (Nikon, MRD00205)
- High Speed Galvo Miniscanner (Nikon, 99316)
- Nikon LUNA (Nikon, MHF45000)
- Detectors: Hammamatsu Flash 4.0 sCmos camera (Hammamatsu, cat. no. C11440-22CU)
- Light engine for epifluorescence (Nikon, MEE54100)
- Large incubator with heating and CO2 control (Nikon, 77025108)

Flow cytometry

- BD Influx sorter (BD, cat. no. 646500) equipped with
- Lasers (488, 563 nm)

1.4.4 Reagent setup

Complete medium I

For cancer cell expansion, prepare complete medium I by supplementing DMEM (450 mL) with FBS (50 mL), and Gentamicin (500 µL). Sterile filter and store at 4C until needed. Before use, warm up in a water bath (37°C). Complete medium I can be stored at 4C for up to 6 months.

FACS buffer

Mix 0.2 g BSA, 20 µL of EDTA, and 50 mL of nanopure water. Sterile filter and store at 4C until needed. FACS buffer can be stored at 4C for 4 months.

Reconstitution buffer (RB)

Mix 110 mg NaHCO3, 240 mg HEPES free acid, and 5 mL nanopure water to make a stock solution. Sterile filter, aliquot, and store at -20C until needed. RB can be stored at -20C for 1 week.

PEG

Mix PEG with PBS to make a 100 mg/mL stock solution. Sterile filter, aliquot, and store at -20C until needed. PEG can be stored at -20C for 3 weeks.

NaOH

Mix NaOH with nanopure water to make a 1N stock solution. Sterile filter and store at -20C until needed. NaOH can be stored at -20C for 6 months.

Gel digestion buffer

Mix 10 mg of collagenase with 1 mL of PBS. Sterile filter and store at -20C until needed. Gel digestion buffer can be stored at 4C for 6 months

1.4.5 Equipment setup

Photoconversion

Before you begin, make sure to calibrate the laser with the galvo scanner. You will need to determine the appropriate settings for some key simulation parameters, such as dwell time and laser power. Optimize these parameters to obtain high post-photoconversion fluorescence but not expose your cells to phototoxicity.

FACS

Ensure that the 405 nm laser line on the machine can be turned off, as this can result in photoconversion during flow sorting. Ensure that the equipment contains the appropriate lasers and filters to capture both the native and photoconverted fluorescent states to ensure enrichment of the desired population. A machine that can support chilled sorting is preferred to maintain viability of the cells.

1.4.6 Procedure

Cloning of Dendra2 into lentiviral vector

1. Amplify the Dendra2-Lifeact vector using PCR with the custom primers. Thermocycling conditions are listed in Box Y.
2. Digest the pSin plasmid using EcoRI and SpeI.
3. Run the digested plasmid, along with an undigested control, on a 1.5 wt% agarose gel
4. Check for successful digestion, and gel purify the digested backbone using QIAQuick gel extraction kit.
5. Clone the Dendra2-Lifeact fragment into the pSin backbone using T4 Ligase.
6. Transform the plasmid into competent DH5a for amplification.
7. Harvest the lenti-Dendra2 vector using the Promega miniprep kit.
8. Verify the sequence through sequencing with the custom primers.

Box Y

Step	Temp	Time
Initial Denaturation	98°C	30s
Cycling (30x)	98°C 68°C 72°C	10s 20s 20s
Final Extension	72°C	5 min
Hold	4C	

Production of Lentiviruses

9. Culture HEK293T cells in a 6-well plate until 70-80% confluency.
10. Transfect the cells 16–24 h after plating with the lenti-Dendra2 plasmid along with lentiviral packaging and envelope vectors using the Lipofectamine 3000 kit. For the DNA component, add equimolar amounts of each plasmid. Add the transfection mix drop-wise to the cells and gently swirl the plate to mix.

11. Twenty-four hours after transfection, replace the medium with fresh medium. Check the cells, using a fluorescence microscope to determine transfection efficiency.
12. Harvest the virus-containing medium on the 3rd day after transfection. Check the cells again, using a fluorescence microscope to determine virus production efficiency. Collect the medium and filter using the 0.45 μm filter to remove cell debris. Collect the virus in 1.5 mL eppendorf tubes and store at -80C for long term storage, or use immediately for transduction.

Viral transduction of Dendra2 into MDA-MB-231

13. Culture MDA-MB-231 cells in a 6-well plate until 70-80% confluency.
14. Aspirate the media, wash the cells once with PBS, and replace with new growth media.
15. Add 75 μL of collected lenti-Dendra2 dropwise to the well. Gently swirl to mix.
16. Monitor transduction efficiency through fluorescent microscopy. It may take 2-3 days before cells start to fluoresce.
17. Passage cells into larger flasks to prepare for purification through FACS. Using a wild-type control, gate for the cells expressing above background levels of green fluorescence. You may choose to collect only the cells that have the highest fluorescence for ease of identification for downstream experiments.

3D culture of MDA-MB-231 Dendra2 (MDA-Dendra)

18. Thaw out RB, NaOH on ice.
19. Place a 48-well plate in the incubator to preheat to 37°C.
20. Passage MDA-Dendra and count the cells using a hemocytometer. Keep cells on ice.

21. Calculate the amount of reagents required to make the 3D collagen hydrogel. Box Z displays sample calculations for making a 2.5 mg/mL collagen + 10 mg/mL PEG hydrogel, with 50,000 cells/mL embedded. Adjust calculations as necessary.

Box Z

Total desired volume of gel: 200 µL. Make 250 µL to account for reagent loss during pipetting.

Stocks	Desired Concentration or Volume	Final Volume
Collagen: 9.00 mg/mL	2.5 mg/mL	2.5*0.25/9.00 = 69.4 µL
PEG: 100 mg/mL	10 mg/mL	10*0.25/100 = 25 µL
NaOH	6.25% * volume Collagen	0.0625*69.4 = 4.34 µL
RB	Remaining Volume / 2	(250-69.4-25-4.34)/2 = 75.6 µL
Cells (200 Kcells/mL)	50 Kcells/mL	50*0.25/150 = 62.5 µL
Media	Top off until gel volume is reached	250-69.4-25-4.34-75.6-62.5 = 13.2 µL

Note: The following steps must be performed quickly and carefully. Carefully mix the solutions at every step and do not introduce bubbles into the solution. If bubbles form, start over as the architecture of the hydrogel will not be homogenous. Steps must be performed quickly and reagents kept as cold as possible to prevent polymerization before the gel solution is incubated at 37°C.

22. Place the collagen, media, and an empty 1.5 mL Eppendorf tube on ice.
23. Mix the reagents in the empty Eppendorf tube in the following order: cells, media, PEG, RB, Collagen, NaOH.
24. Immediately pipet the gel solution into the preheated well plate and incubate at 37°C for 30 minutes.

25. After 30 minutes, perform 3 washes of PBS for 5 minutes each. The amount of PBS to pipette on top of the gel is the same as the gel volume.
26. After the last wash, aspirate the PBS and add growth media on top of the gel. The amount of growth media to pipette on top of the gel is the same as the gel volume.
27. Leave the 3D culture for a week, changing media every 2-3 days.

Tagging of collective cell phenotypes by photoconversion

28. Transfer the collagen gel to a glass-bottom dish. Add enough media to keep the gel hydrated, but not so much that the gel will float in solution or move around in the dish.
29. Transfer the glass-bottom dish to a fluorescent microscope stage.
30. In the Bruker Miniscanner panel, set the dwell time to 75 us and the 405 nm laser power to 30%.
31. Calibrate the galvanometric scanner.
32. Using a 20x lens, identify the cells you would like to photoconvert.
33. Verify that other cells are not within 10 μm in x-y and not within 200 μm in z
34. Draw an ROI around the cells you would like to photoconvert. Right-click and select to use ROI as a Stimulation ROI.
35. In the Bruker Miniscanner panel, click Stimulate.
36. Repeat steps 24-27 until all cells of interest within the hydrogel have been photoconverted.

Gel extraction and FACS sorting of photoconverted cells

37. Transfer the collagen gel to a 1.5 mL Eppendorf tube.
38. Add 50 μL of 10 mg/mL collagenase to the gel.
39. Use a P1000 pipette tip to gently mash and mechanically disrupt the gel

40. Incubate the gel in the water bath at 37°C for 5 minutes
41. Mix the solution further with the P1000 pipet tip. The entire solution should be pipetted up and down within the pipette tip. Minimize the introduction of bubbles as much as possible.
42. Incubate the gel in the water bath at 37°C for 5 minutes
43. Mix the solution with a P200 pipette tip. The entire solution should be pipetted up and down within the pipette tip. Minimize the introduction of bubbles as much as possible.
44. Incubate the gel in the water bath at 37°C for 5 minutes
45. Centrifuge the solution at 400 xg for 4 minutes
46. Discard the supernatant, add 50 µL of 0.25% trypsin, and resuspend the pellet.
47. Incubate in the water bath at 37°C for 5 minutes.
48. Mix the solution with a P200 pipette tip.
49. Incubate in the water bath at 37°C for 5 minutes.
50. Centrifuge the solution at 400 xg for 4 minutes.
51. Remove the supernatant and resuspend the pellet with an ice-cold FACS buffer.
52. Strain the cells prior to FACS.
53. Use forward scatter and side scatter to exclude debris.
54. Use un-photoconverted controls to set gates for red-fluorescence
55. Collect cells expressing red-fluorescence higher than gate

1.4.7 Troubleshooting

Problems

- low transduction efficiency
    - increase viral load, or post sort cells for expression by flow cytometry

- cell subpopulations are too closely clustered together
    - seed cells sparser in culture platform
- low fluorescence of photoconverted cells
    - flow sort for the highest Dendra2 expressing cells, or increase the dwell time of the laser at each point
- gel digestion is incomplete
    - increase collagenase concentration, increase mechanical disruption intervals, or increase incubation time with collagenase
- low cell viability post sorting
    - ensure cells are placed on ice at all times and that the flow sorting speed is lower

1.4.8 Timing

Steps 1-8, cloning of Dendra2: 1 week for the provided protocol

Steps 9-12, lentiviral production: 3 - 4 days

Steps 13-17, lentiviral transduction: 4-5 days

Steps 18-27, 3D Culture: 1 week

Steps 28-36, photoconversion: 3 hours

Steps 37 - 55, gel extraction and FACS sorting: 1-2 hours

1.5 Acknowledgments

General: I would like to acknowledge Daniel Ortiz for his assistance in cloning the Dendra2 construct into the lentiviral vector and the Human Embryonic Stem Cell Core for assistance with flow cytometry. I would also like to acknowledge Khoi Le for his assistance in the optimization of this protocol during its developmental stages.

Funding: This work was supported by a Burroughs Wellcome Fund Career Award at the Scientific Interface to S.I.F. (1012027), NSF CAREER Award to S.I.F. (1651855), ACS Institutional Research Grant (15-172-45-IRG) provided through the Moores Cancer Center.

Chapter 2: Stress response, proliferation, and immunologic cellular processes are coordinated with invasive and non-invasive phenotypes

2.1 Introduction

Cellular heterogeneity is indispensable for population-level survival strategies of multicellular organisms, such as bet-hedging in order to achieve a better chance of survival when faced with new stresses(*36*, *37*). In cancer, the genetic, epigenetic, transcriptional, and proteomic differences among tumor cells can give rise to diverse phenotypes, some of which can persist in dysregulated environmental conditions, survive therapeutic attempts, and migrate away from the primary tumor to form metastases(*20*). Linking the heterogeneity observed in cellular genomics with phenotypic heterogeneity has significant potential to inform successful population-level treatments. However, the challenges associated with measuring phenotypic heterogeneity and isolating particular cells within the complex *in vivo* environment make it difficult to identify the most basic transcriptional modules regulating individual cell behaviors.

Fortunately, significant evidence suggests that physiologically relevant phenotypes of BRCA cells can be studied in less complex in vitro systems by embedding the cells in 3D Type I Collagen (COL1) hydrogels. BRCA cell lines, organoids from mouse tumors, and organoids from human tumors embedded in this model system harbor the same pattern of differentiation markers as are observed in studies of mouse mammary tumor histology and human BRCA histology(*33*, *38*). Studies have also shown that BRCA cells cultured in this model system upregulate a conserved transcriptional program of 70 genes that is predictive of poor prognosis in human BRCA and eight additional cancer types, with the highest predictive value in triple negative breast cancer (TNBC) (Hazard Ratio = 3.85, Cox p value = 0.007)(*21*, *38*). Thus, a growing body of evidence suggests

that 3D culture of BRCA cells in COL1 is a relevant model system for studying physiologically relevant cancer phenotypes.

Importantly, BRCA cells embedded in a 3D COL1 matrix maintain heterogeneity. In particular, they can take on a range of migration phenotypes, from non-invasive to single-cell mesenchymal style migration to collective invasion(*21*), with the collective invasion phenotype being linked to the metastatic phenotype in vivo(*39*). To begin to define the molecular programs underlying BRCA cell migration heterogeneity, we sought a method capable of linking cell phenotype to gene expression programs. While advances in single cell omics technologies have significantly improved our ability to characterize cell heterogeneity, these methods involve the sequencing of individual cells from a bulk sample and determining cell clusters solely based on differences in the molecular signature. However, the biological interpretation of these complex data is only at an early stage. Inferring cell state, function, and response to treatment from such data remains highly subjective and dependent on *a priori* knowledge(*25*). Cell subpopulations identified from analyzing sequencing data can only be validated with experiments after clusters have been defined, and this relies heavily on the assumption that transcriptomic data maps well to functional profiles. Partitions made from unsupervised clustering methods could potentially divide the sample into groups that may have no functional biological meaning, particularly for samples that are more similar as a whole, like cells of the same type. While standards and strategies are constantly evolving, there remains a lack of consensus on how to define cell types and subtypes based on sequencing data(*26*). The field of single cell analysis is rapidly moving towards integrative, multi-scale measurements to improve the functional interpretability of single cell data. Thus far, transcriptome measurements have been integrated with multiple omics, genotype, cell electrophysiology, lineage tracing, and spatial information(*27–32*). To more concretely link

phenotype to omics data, rare cell subpopulations may also be functionally sorted using innovative physical or image-guided techniques(*5*, *33*).

Here, we explored whether traditional single cell RNA sequencing (scRNAseq) followed by unsupervised clustering analysis would be capable of correctly inferring migration phenotype. This would inherently require that phenotypic regulators dominate the transcriptome of the cells to enable similarity-based clustering. However, we posited that other processes might dominate single cell transcriptomes such that phenotypic regulators could represent a much smaller signal in the data. So, we used an image-guided phenotypic sorting technique to ask whether phenotypically supervised scRNAseq (hereafter referred to as pheno-scRNAseq for ease of reference) can provide more insight into the heterogeneous cell migration behaviors of MDA-MB-231 BRCA cells (MDAs) than unsupervised scRNAseq. Photoconversion-based cell labeling followed by rapid dissociation into a single cell suspension, fluorescence activated cell sorting (FACS), and scRNAseq enabled direct comparison of phenotype labels with unsupervised transcriptional clustering. Unsupervised clustering was not able to correctly infer migration phenotype. Accordingly, pheno-scRNAseq revealed unique molecular programs associated with the migration state. Functional experiments targeting several identified genes validated that they play an active role in regulating migration behaviors. Specifically, perturbing *HSP90AB1*, *DEK*, and *F3* regulated the collective invasion phenotype. Pheno-scRNAseq further revealed that collectively invasive cells exist in a "go and grow" state, where biosynthetic processes, proliferation, oxidative stress responses, and ER stress responses are upregulated. However, non-invasive cells limit proliferation and biosynthesis and are dominated by redox homeostasis and immunomodulatory gene expression programs. These relationships were recapitulated in mouse 4T1 BRCA cells to confirm that the results were not cell line specific. Our phenotypic sorting

approach also enabled reseeding experiments that probed the stability of these cellular states, revealing that invasive cells remain in a stable migration state while non-invasive cells are plastic and capable of repopulating both phenotypes. Studies of other phenotypes and cell types of interest may benefit from the unique information provided by the pheno-scRNAseq approach.

2.2 Results

2.2.1 BRCA cells exhibit heterogeneous migration phenotypes

MDA-MB-231 (MDA) cells embedded in 3D COL1 matrix take on at least two distinct collective phenotypes, which develop from single cells over the course of 7 days (Fig. 2.1A). The majority of cells, approximately 81%, formed collectively invasive cell structures (Fig. 2.1B), a morphology characterized by a low circularity index (Fig. 2.1D). A smaller subset of tumor cells, approximately 19%, did not invade (Fig. 2.1C), characterized by a high circularity index (Fig. 2.1D). Based on their morphology, we termed the invasive structures as "networks" and the non-invasive structures as "spheroids." Confocal microscopy revealed that collectively invading networks were tightly packed with cells (Fig. 2.1B) while spherical structures were capable of forming hollow lumens reminiscent of normal breast epithelial acini (Fig. 2.1C). Similar phenotypes were observed for mouse 4T1 mammary carcinoma cells embedded in 3D COL1 matrix (Fig. 2.1E).

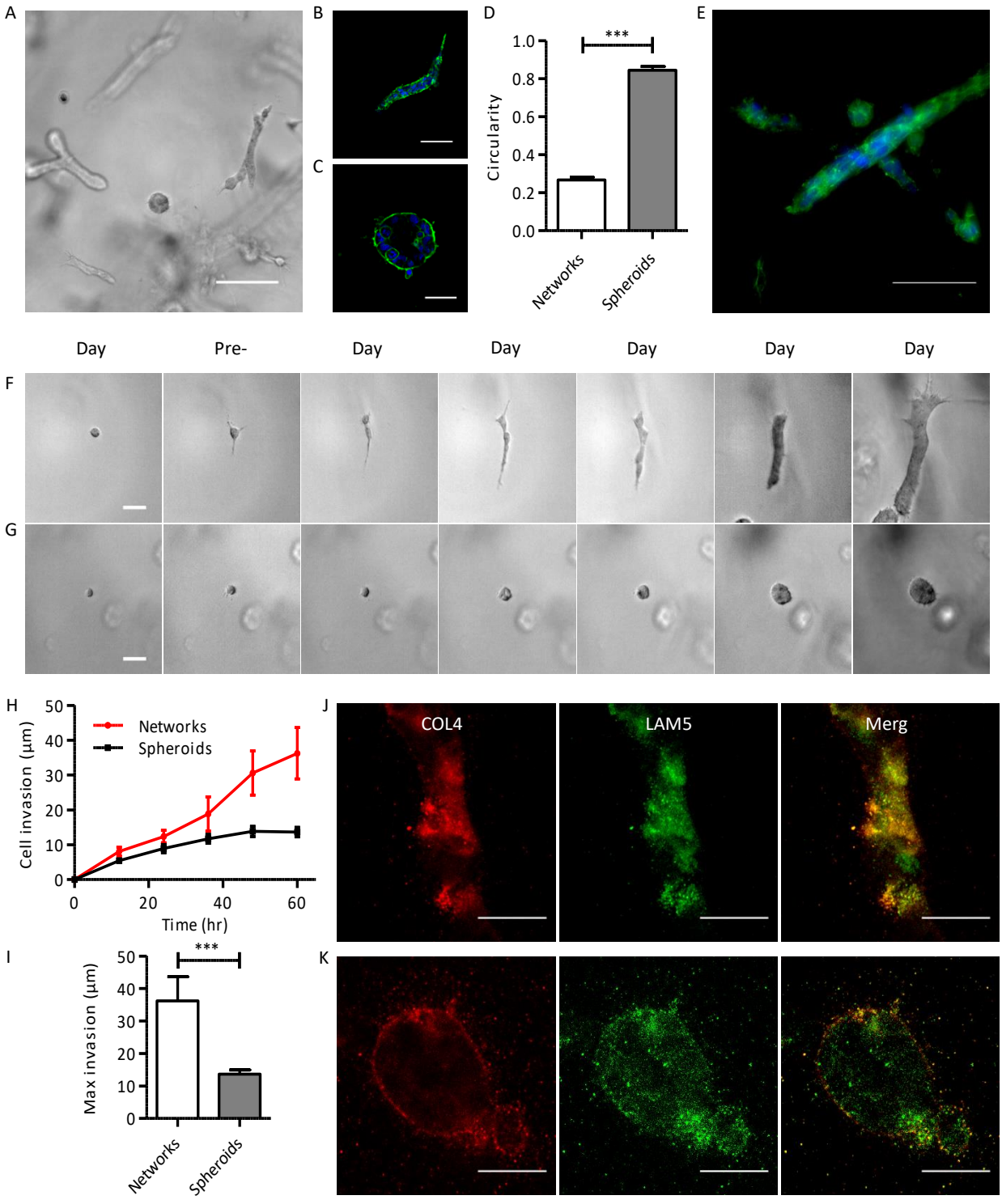
Timelapse microscopy revealed that cells that developed into network structures (Movie S1) began as single cells undergoing uniaxial elongation at early time points (Fig. 2.1F). Proliferation appeared to support the eventual creation of a smooth, continuous, collectively invasive clonal network (Fig. 2.1F). Cells that developed into spherical structures did not elongate

at early time points (Fig. 2.1G). These cells maintained a predominantly rounded morphology. In this case, proliferation enabled the eventual creation of a smooth, rounded spheroid (Fig. 2.1G).

The collective phenotypes also displayed differential aptitudes for invading into the local matrix. The network structures were significantly more invasive than the spheroids, displaying increased spreading away from the initial seeding point over the course of several days (Fig. 2.1H). By 2.5 days, the cells that composed network structures invaded roughly three times farther into their surroundings compared to cells in spheroid structures (Fig. 2.1I).

Since the formation of hollow acini by normal breast epithelial cells has previously been linked to basement membrane deposition(*32*), we assessed whether MDA cancer cells deposited matrix proteins into the COL1 microenvironment. Immunofluorescent staining revealed that both invasive and non-invasive subpopulations deposited their own cell-derived matrix consisting of basement membrane proteins Laminin-5 (LAM5) and type IV collagen (COL4A1) (Fig. 2.1J, K). COL4A1 appeared to be localized more towards the cell-extracellular matrix interface in spheroids compared to invasive networks, where localization was more heterogeneous and intracellular. For LAM5, expression in spheroids was low, while invasive networks displayed heterogeneous, intracellular expression.

Figure 2.1. BRCA cells exhibit heterogeneous migration phenotypes. (**A**) Representative brightfield image of MDA-MB-231 cells cultured in a 3D COL1 matrix after 7 days of culture. Scale bar 200 μm. (**B**) Confocal z-slice of the network and **(C)** spheroid phenotypes. Scale bar 100 μm. (**D**) Quantification of the circularity of heterogeneous collective phenotypes. (**E**) Similar phenotypes are observed in 4T1 cells cultured in 3D type I collagen. (**F** and **G**) Timelapse microscopy depicting the different patterns of growth and morphogenesis of two structurally distinct multicellular phenotypes. (**F**) Single cells that eventually develop into networks display growth and migration that lead to eventual fusion into a multicellular network. (**G**) Single cells that eventually develop into spherical structures display localized growth and development with continual maintenance of the spherical shape. Scale bar 50μm. (**H**) Quantification of the invasion of cells into the local ECM depending on their collective phenotype. (**I**) Maximum invasion of each phenotype from the initial seeding point after 60 hours of culture. (**J** and **K**) Representative immunofluorescence z-slice images of networks (**J**) and spheroids (**K**) stained for COL4A1 and LAM5.


A
B
C
D
E
1.0
0.8
0.6
0.4
0.2
0.0
Circularity
***
Networks
Spheroids
Day
Pre-
Day
Day
Day
Day
Day
F
G
H
50
40
30
20
10
0
Cell invasion (µm)
Networks
Spheroids
0
20
40
60
Time (hr)
J
COL4
LAM5
Merg
I
***
Max invasion (µm)
Networks
Spheroids
K

2.2.2 Phenotypic cell sorting improves transcriptome-phenotype coupling

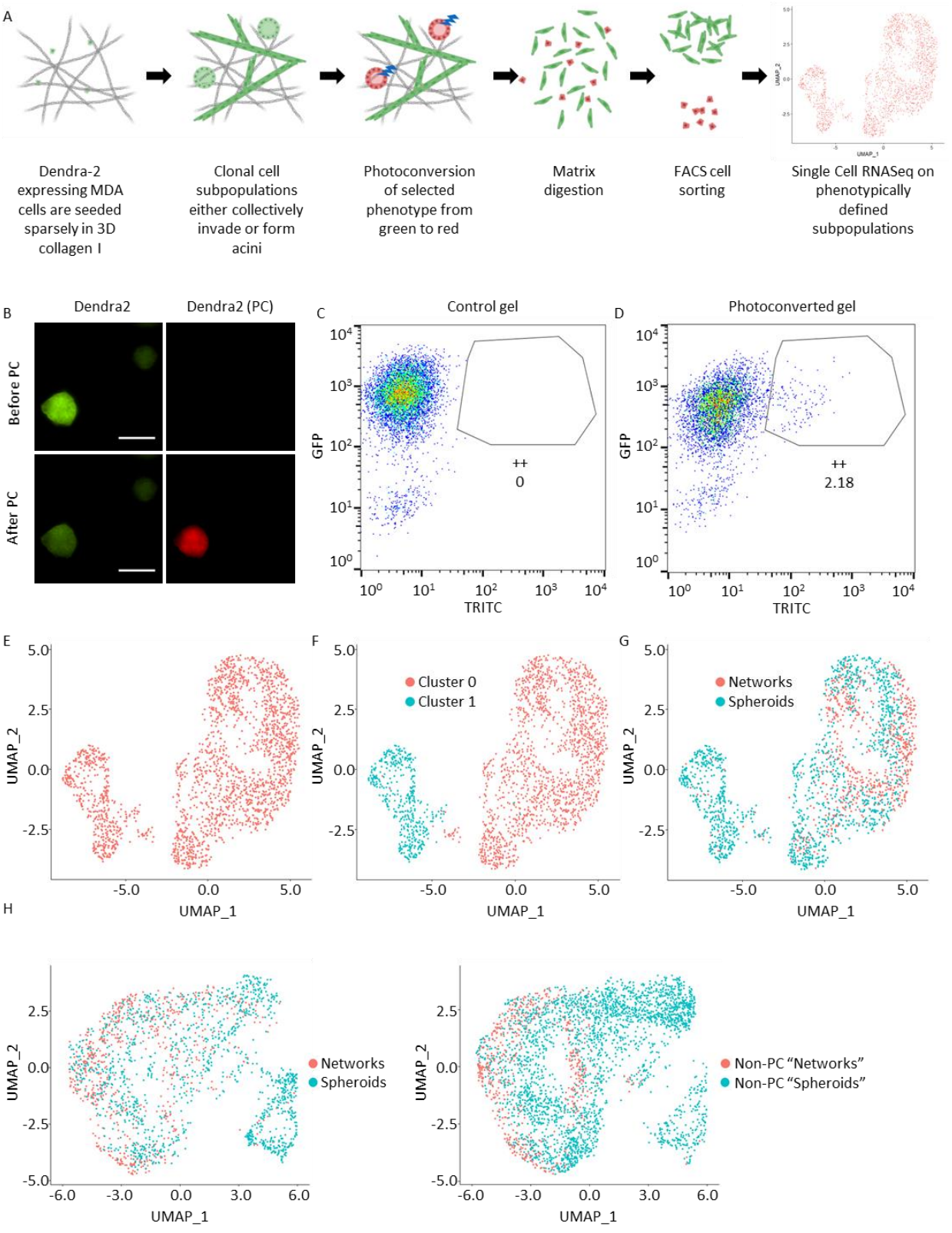
To begin to identify the mechanisms underlying collective migration heterogeneity in MDA cells, we next sought to separate the network cells from the spheroid cells and analyze their molecular differences. Morphologically, these invasive and non-invasive tumor cell subpopulations could be clearly differentiated based on their circularity (Fig. 2.1D), so we subsequently used circularity as a metric by which we distinguished the two phenotypes. To sort the cells based on their phenotype for direct scRNAseq analysis, we devised a technique that enables fluorescent tagging of cells of interest (Fig. 2.2A). Building on a method recently described by Konen et al.(*33*), MDAs were first transduced with Dendra2-Lifeact (MDA-Dendra). Dendra is a photoconvertible protein that changes from green to red fluorescence upon stimulation with a 405 nm laser, facilitating targeted red-fluorescent tagging of specific cells. MDA-Dendra cells were embedded sparsely and cultured in COL1 matrices for one week, allowing the development of clonal collective structures. Transduction with Dendra2-Lifeact did not significantly impact collective cell migration capabilities compared to wild-type MDA cells(*21*). To enable higher throughput photoconversion of each cell phenotype of interest while maintaining fine spatial resolution to target individual cell structures (see Methods for details), we constructed a custom widefield microscope with a galvanometer scanner and laser power source. In one experiment, spheroid structures were selectively stimulated with the 405 nm laser, inducing red fluorescence, while unexposed cells continued to fluoresce green (Fig. 2.2B). In a separate experiment and sample, network cell structures were similarly selectively photoconverted. Subsequent rapid digestion of the matrices and sorting of the cells by FACS enabled recovery of phenotypically pure populations for direct molecular analysis (Fig. 2.2C, D).

Network and spheroid cells isolated by photoconversion were directly subjected to scRNAseq. A UMAP plot of the scRNAseq data from both cell phenotypes aggregated together was constructed (Fig. 2.2E). Conventionally, unsupervised clustering would be used to identify two transcriptionally distinct subpopulations of cells (Fig. 2.2F). However, with our phenotypic labels applied to the data (Fig. 2.2G), it was evident that agnostic transcriptional clusters were not predictive of the functional cell migration phenotypes (compare Fig. 2.2F to 2.2G). Even though there was a significant enrichment of network cells in the larger cluster (cluster 0) and of spheroid cells in the smaller cluster (cluster 1) (Fisher's exact test: odds ratio = 12.51, p-value = 5.37e-62), there was still a high number of spheroid cells that clustered together with network cells. Analysis on the basis of our phenotypic labels revealed a set of 178 genes that were differentially expressed (DE) between network and spheroid cells (Table S2.1), whereas analysis on unsupervised clusters highlighted a set of 528 DEGs. Importantly, 70 of the genes identified using phenotypic labeling were not identified using unsupervised clustering (Table S2.1, highlighted genes).

To control for the effects of our labeling and sorting process, we also sequenced MDA-Dendra cells extracted from COL1 matrices that were not stimulated or sorted. Integrating the sorted and non-sorted cell datasets by normalizing for sequencing depth and correcting for batch effects demonstrated that both have a similar data structure (Figure 2.2H). We then constructed a metagene from the 178 DEGs between the network and spheroid cells and applied this metagene to score and label the non-sorted cells. The 178 gene signature placed phenotypic labels on the non-sorted cells in a similar state space compared to their sorted counterparts (Figure 2.2H), adding further evidence that stimulation and sorting did not significantly change the transcriptional profiles of the cells. We also investigated the DEGs between the sorted and non-sorted cells, which represent some combination of sequencing batch effects and photoconversion/sorting effects. This

revealed 1127 DEGs, of which only 8 were associated with a response to UV. Further analysis showed that only 2 of these 8 UV response associated genes were contained in our list of 178 DEGs that differentiated the network and spheroid cells. This suggested that the effects of the photoconversion and sorting process are negligible for our analysis.

Figure 2.2. Phenotypic cell sorting improves transcriptome-phenotype coupling. (**A**) Schematic overview of our workflow for phenotypic cell sorting. MDA-Dendra cells are cultured in type I collagen, photoconverted, released from the matrix, and sorted based on red fluorescence for immediate scRNAseq or other downstream experiments. (**B**) Images of multicellular MDA-Dendra structures before photoconversion (left) and after photoconversion (right). Scale bar 100 μm. (**C**) Fluorescent profile of a control gel, where no cells were photoconverted. (**D**) Fluorescent profile of cells released from a gel after photoconversion. A fraction of cells exhibits greater red fluorescence compared to the control. (**E**) A UMAP plot generated from the pooled transcriptomic signatures of the cells isolated by phenotypic cells sorting. (**F**) Clusters identified based on unsupervised clustering methods. (**G**) Clusters labeled by phenotype. (**H**) Comparison of cells labeled by photoconversion with non-photoconverted cells that were scored by the metagene derived from the differentially expressed genes of the labeled cells. Data was corrected for batch effects and sequencing depth prior to UMAP projection.


A
Dendra-2 expressing MDA cells are seeded sparsely in 3D collagen I
Clonal cell subpopulations either collectively invade or form acini
Photoconversion of selected phenotype from green to red
Matrix digestion
FACS cell sorting
Single Cell RNASeq on phenotypically defined subpopulations
B
Dendra2
Dendra2 (PC)
Before PC
After PC
C
Control gel
D
Photoconverted gel
GFP
TRITC
++
0
2.18
E
F
G
UMAP_1
UMAP_2
Cluster 0
Cluster 1
Networks
Spheroids
H
Non-PC "Networks"
Non-PC "Spheroids"

2.2.3 Biological processes that differentiate collective cell phenotypes are conserved

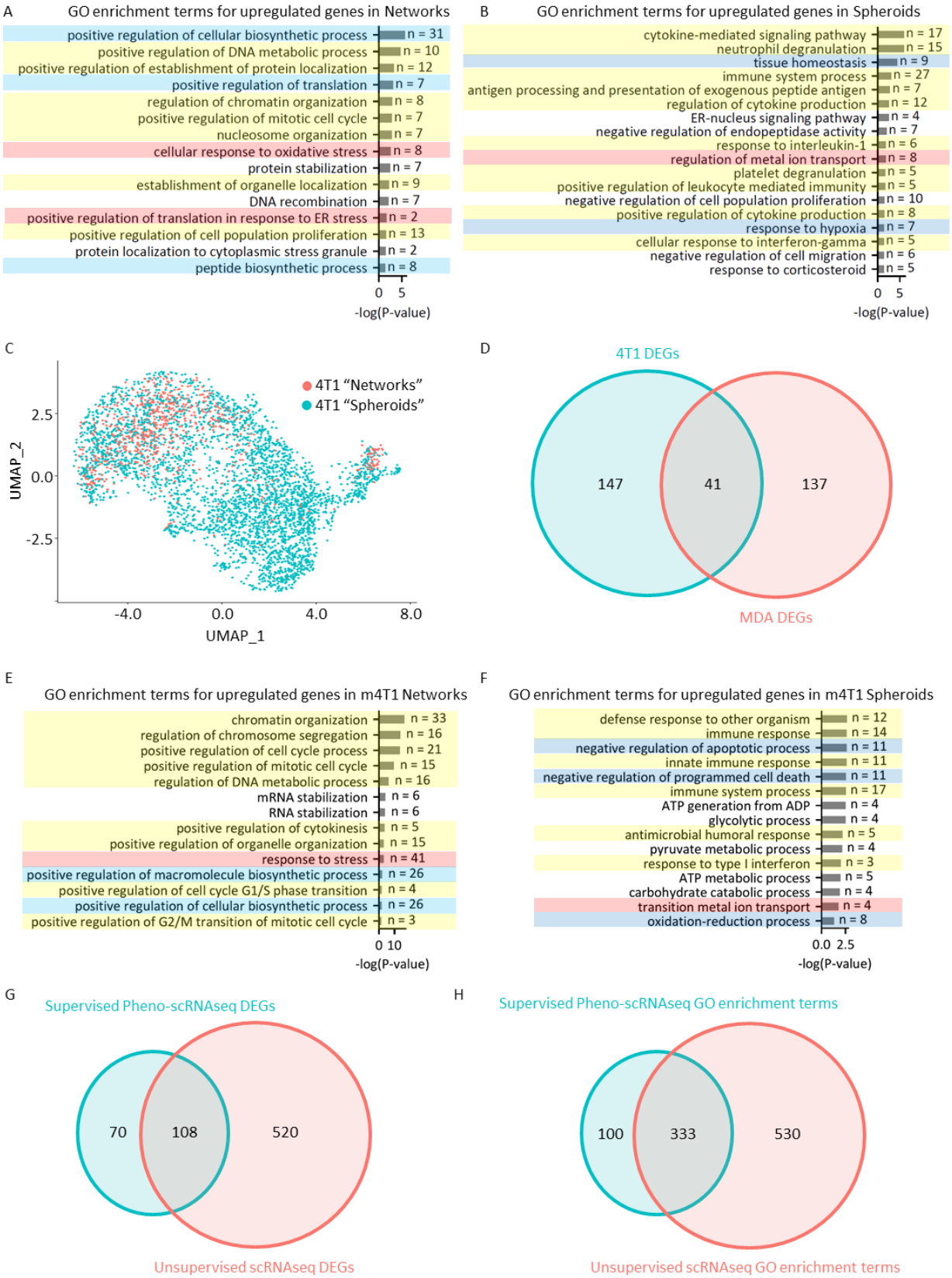
Gene expression analysis based on phenotypic labels revealed that 101 of the 178 DE genes were significantly upregulated in invasive network cells compared to non-invasive spheroid cells. Gene ontology (GO) analysis of these genes identified several enriched processes, including positive regulation of biosynthesis, translation in response to endoplasmic reticulum (ER) stress, response to oxidative stress, and positive regulation of proliferation (Fig. 2.3A). Interestingly, the remaining 77 genes upregulated in the spheroid cells were enriched for several immunomodulatory processes, negative regulation of proliferation and migration, transcriptional regulation through ER-nucleus signaling pathways, and processes promoting homeostasis, including redox equilibrium (Fig. 2.3B).

To determine whether our findings extended to other BRCA cells, we performed scRNAseq on 4T1 murine mammary carcinoma cells. As shown in Figure 2.1E, 4T1s take on very similar migration phenotypes as MDAs, forming network and spheroid structures over seven days. UMAP projection and labeling by metagene scoring reveal a similar pattern to the MDA cells, where the spheroid phenotype is the predominant phenotype in the smaller cluster, while the network phenotype occupies the state space farthest from the lone cluster (Fig 2.3C). Differential expression analysis revealed 166 DEGs that describe the differences between scored 4T1 networks and spheroids with 1:1 human orthologs. When compared to the 149 MDA DEGs that had 1:1 mouse orthologs, 40 genes are shared between the mouse and human datasets (Fig. 2.3D). This overlap was statistically significant (Fisher's exact test: odds ratio = 33.92, p-value = 6.50e-41). GO enrichment analyses of the murine cells also reveal similarly themed biological processes being enriched compared to their human counterparts, as shown in Figure 2.3E and F.

2.2.4 Pheno-scRNAseq provides unique and selective information

Since pheno-scRNAseq identified a smaller set of DEGs (178) than unsupervised scRNAseq (528), we hypothesized that phenotypic labeling may provide a more selective and more functionally relevant gene set than analyses on populations identified from unsupervised clustering. To further explore this idea, we asked whether certain biological processes were uniquely enriched by phenotype labeling or consistently enriched between supervised and unsupervised analyses. Such differences could provide important guidance for prioritizing gene modules to target in functional studies aimed at identifying meaningful associations with migration behaviors. Figure 2.3G shows a Venn Diagram comparing the detected DEGs using supervised pheno-scRNAseq analysis to those found using unsupervised scRNAseq analysis. Figure 2.3H displays a Venn Diagram comparing the significant GO enrichment terms detected using DEGs from supervised pheno-scRNAseq analysis to those found using DEGs from unsupervised scRNAseq analysis. A detailed list of the GO enrichment analysis is provided in Table S2.2. These plots reveal two distinct features of supervised analysis compared to unsupervised analysis. First, for our dataset, supervised analysis narrows down the number of DEGs and GO enrichment terms that are found to be statistically significant (adjusted p-value < 0.05, FDR $< 5\%$). The second feature can be seen visually in the left most partition in each Venn Diagram. These are DEGs and GO terms that were uniquely detected using supervised analysis and not in unsupervised analysis. Together, these results suggest that supervised analysis is indeed more selective and provides unique information.

Figure 2.3. Biological processes that differentiate collective cell phenotypes are conserved. (A) Highlighted significant GO enrichment terms based on the list of upregulated genes in the network cell population. **(B)** Highlighted significant GO enrichment terms based on the list of upregulated genes in the spheroid cell population. **(C)** UMAP of 4T1 mouse cells, labeled after scoring with the metagene derived from the DEGs between the MDA network and spheroid cells. **(D)** Overlap of the orthologs of the 4T1 DEGs between labeled networks and spheroids and the 178 DEGs found between the MDA networks and spheroids. **(E)** Highlighted significant GO enrichment terms based on the list of upregulated genes in the labeled 4T1 network cell population. Similarly colored highlights between **(A)** and **(E)** denote similarly themed processes that were enriched. **(F)** Highlighted significant GO enrichment terms based on the list of upregulated genes in the labeled 4T1 spheroid cell population. Similarly colored highlights between **(B)** and **(F)** denote similarly themed processes that were enriched. **(G)** Venn Diagram displaying the unique and overlapping DEGs when comparing supervised to unsupervised analysis. Unsupervised analysis identified a larger number of significant DEGs. However, many of these DEGs were not significant when compared to supervised analysis. **(H)** Venn Diagram displaying the unique and overlapping GO enrichment terms when comparing supervised to unsupervised analysis. Unsupervised analysis identified a larger number of significant GO enrichment terms. However, many of these GO enrichment terms were not significant when compared to supervised analysis.


A
GO enrichment terms for upregulated genes in Networks
positive regulation of cellular biosynthetic process n = 31
positive regulation of DNA metabolic process n = 10
positive regulation of establishment of protein localization n = 12
positive regulation of translation n = 7
regulation of chromatin organization n = 8
positive regulation of mitotic cell cycle n = 7
nucleosome organization n = 7
cellular response to oxidative stress n = 8
protein stabilization n = 7
establishment of organelle localization n = 9
DNA recombination n = 7
positive regulation of translation in response to ER stress n = 2
positive regulation of cell population proliferation n = 13
protein localization to cytoplasmic stress granule n = 2
peptide biosynthetic process n = 8
0 5
-log(P-value)
B
GO enrichment terms for upregulated genes in Spheroids
cytokine-mediated signaling pathway n = 17
neutrophil degranulation n = 15
tissue homeostasis n = 9
immune system process n = 27
antigen processing and presentation of exogenous peptide antigen n = 7
regulation of cytokine production n = 12
ER-nucleus signaling pathway n = 4
negative regulation of endopeptidase activity n = 7
response to interleukin-1 n = 6
regulation of metal ion transport n = 8
platelet degranulation n = 5
positive regulation of leukocyte mediated immunity n = 5
negative regulation of cell population proliferation n = 10
positive regulation of cytokine production n = 8
response to hypoxia n = 7
cellular response to interferon-gamma n = 5
negative regulation of cell migration n = 6
response to corticosteroid n = 5
0 5
-log(P-value)
C
4T1 "Networks"
4T1 "Spheroids"
2.5
0.0
-2.5
UMAP_2
-4.0
0.0
4.0
8.0
UMAP_1
D
4T1 DEGs
147
41
137
MDA DEGs
E
GO enrichment terms for upregulated genes in m4T1 Networks
chromatin organization n = 33
regulation of chromosome segregation n = 16
positive regulation of cell cycle process n = 21
positive regulation of mitotic cell cycle n = 15
regulation of DNA metabolic process n = 16
mRNA stabilization n = 6
RNA stabilization n = 6
positive regulation of cytokinesis n = 5
positive regulation of organelle organization n = 15
response to stress n = 41
positive regulation of macromolecule biosynthetic process n = 26
positive regulation of cell cycle G1/S phase transition n = 4
positive regulation of cellular biosynthetic process n = 26
positive regulation of G2/M transition of mitotic cell cycle n = 3
0 10
-log(P-value)
F
GO enrichment terms for upregulated genes in m4T1 Spheroids
defense response to other organism n = 12
immune response n = 14
negative regulation of apoptotic process n = 11
innate immune response n = 11
negative regulation of programmed cell death n = 11
immune system process n = 17
ATP generation from ADP n = 4
glycolytic process n = 4
antimicrobial humoral response n = 5
pyruvate metabolic process n = 4
response to type I interferon n = 3
ATP metabolic process n = 5
carbohydrate catabolic process n = 4
transition metal ion transport n = 4
oxidation-reduction process n = 8
0.0 2.5
-log(P-value)
G
Supervised Pheno-scRNAseq DEGs
70
108
520
Unsupervised scRNAseq DEGs
H
Supervised Pheno-scRNAseq GO enrichment terms
100
333
530
Unsupervised scRNAseq GO enrichment terms

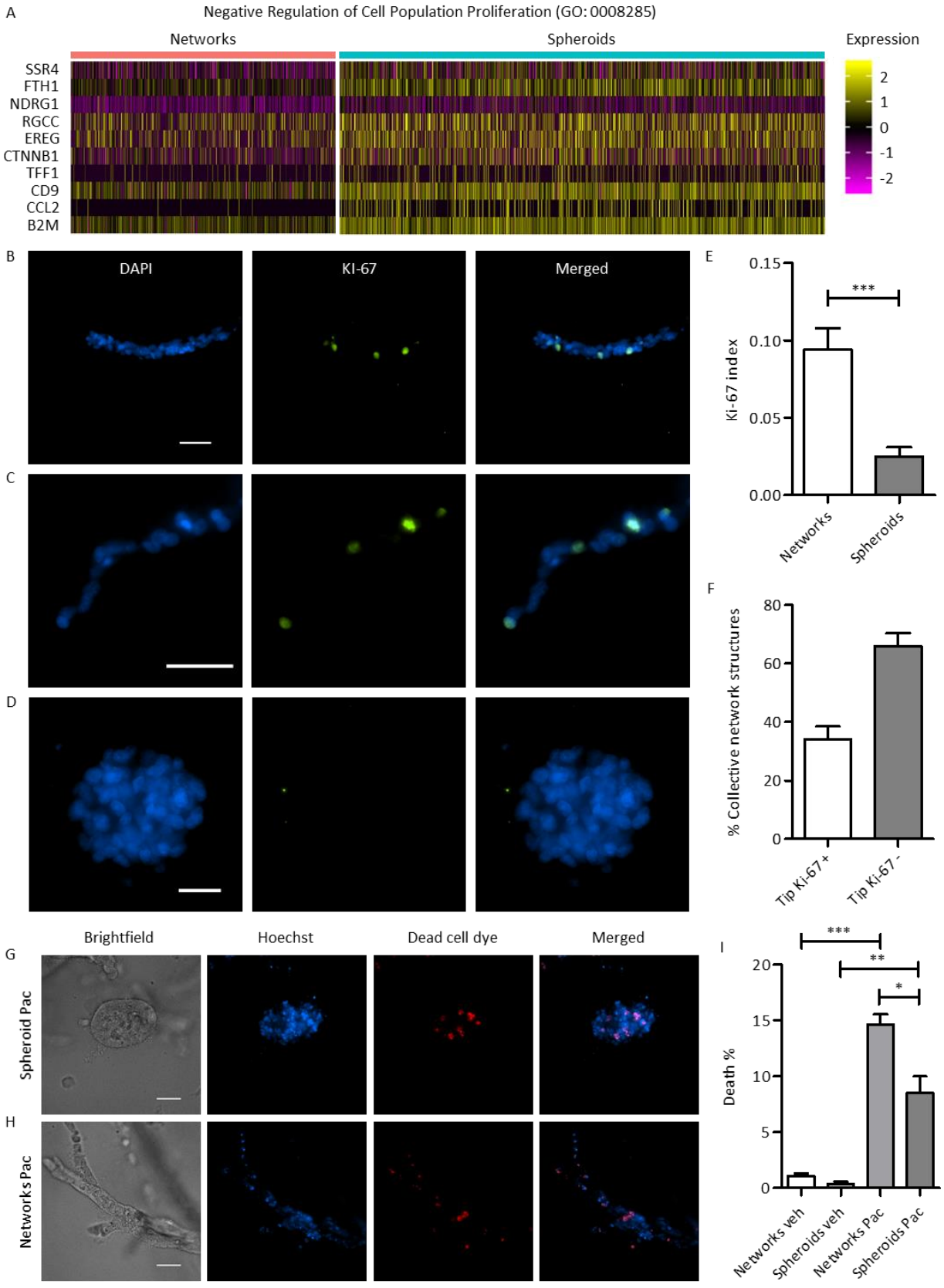
2.2.5 Invasive network cells are more proliferative and more sensitive to chemotherapy treatment

We set out to validate that gene modules identified by supervised analysis were functionally relevant. Of the GO terms that were uniquely identified by supervised analysis, we focused on "negative regulation of cell population proliferation" (GO:0008285, Figure 2.4A). Expression of this group of genes were higher in the spheroid population compared to the network population of cell. Furthermore, *MKI67* (Ki-67), a known marker of proliferation, was uniquely identified as differentially regulated by supervised analysis (Table S2.1) and higher expressed in the network subpopulation. Immunostaining and quantification of Ki-67 showed that the percentage of actively proliferating cells was significantly higher in invasive networks (Fig. 2.4B-E). Cell cycle scoring on the transcriptional markers reveal that the spheroid subpopulation also occupies multiple cell cycle states and has not simply exited the cell cycle (Figure S2.1). Intriguingly, 34% of invasive networks contained leader cells that were Ki-67 positive (Fig. 2.4C, F), suggesting that tip cells can be both invasive and proliferative (Fig. 2.4C). The remaining 66% of networks had Ki-67 positive cells located randomly throughout the network (Fig. 2.4F). Spherical structures were less proliferative, and many exhibited no staining for Ki-67 (Fig. 2.4D, E). These results support our suggestion that supervised analysis provides distinct and important information that can effectively guide follow-up experiments compared to unsupervised analysis.

To further validate the functional relevance of this difference in proliferative state, we performed a cytotoxicity assay with a widely used chemotherapy drug that acts on the cell division process: Paclitaxel. Paclitaxel stabilizes tubulin polymerization, which results in cell cycle arrest(*43*). We hypothesized that the spheroid cells, which express relatively lower Ki-67 and exhibit less Ki-67 staining, should be less sensitive to this drug. We cultured MDAs in 3D COL1

for one week, followed by three days of drug treatment with either 1 μM Paclitaxel or vehicle control. Figure 2.4G and H, shows representative images of dead cell staining in treated networks versus spheroids. Representative images of vehicle control cells are shown in Supplementary Figure S2.2. Cell death was significantly greater in drug treated conditions compared to vehicle, and network cells died more than the spheroid cells (Fig. 2.4I; Fig. S2.2). These data further confirm that the spheroid subpopulation is less proliferative and more resistant to chemotherapy (Fig. 2.4). Taken together, these experiments serve as validation of the functional relevance of a biological process that was uniquely identified by supervised analysis, not by unsupervised analysis.

Figure 2.4. Invasive network cells are more proliferative. (**A**) A heatmap of the list of genes detected by phenotypically guided DE analysis that are in the GO:0008285 term. Spheroids display upregulation of genes associated with "Negative Regulation of Cell Population Proliferation". (**B** and **C**) Immunofluorescent staining of Ki-67 in the network cell population. Scale bar 50 μm. (**C**) Some network structures display Ki-67 staining at the tips of the structures. (**D**) Immunofluorescent staining of Ki-67 in the spheroid cell population. Many spheroids displayed no staining. Scale bar 50 μm. (**E**) Quantification of the percentage of cells in each collective phenotype that stained positively for Ki-67. (**F**) Quantification of the percent of networks that had a tip cell which stained positively for Ki-67. **(G and H)** Brightfield and fluorescence images after treatment with Paclitaxel of spheroids (**G**) and networks (**H**). (**I**) Quantification of cell death after treatment with Paclitaxel. Spheroids show a statistically significant decrease in sensitivity compared to networks.


A
Negative Regulation of Cell Population Proliferation (GO: 0008285)
Networks
Spheroids
Expression
SSR4
FTH1
NDRG1
RGCC
EREG
CTNNB1
TFF1
CD9
CCL2
B2M
2
1
0
-1
-2
B
DAPI
KI-67
Merged
C
D
E
Ki-67 index
0.15
0.10
0.05
0.00
***
Networks
Spheroids
F
% Collective network structures
80
60
40
20
0
Tip Ki-67+
Tip Ki-67 -
Brightfield
Hoechst
Dead cell dye
Merged
G
Spheroid Pac
H
Networks Pac
I
Death %
20
15
10
5
0
***
**
*
Networks veh
Spheroids veh
Networks Pac
Spheroids Pac

2.2.6 Functional validation of additional processes identified by pheno-scRNAseq

Of the 77 genes upregulated in the spheroid cells compared to network cells, GO analysis indicated significant enrichment of several immunomodulatory cellular processes (Fig. 2.3B), including components of the innate, adaptive, and cytokine immune signaling machinery. A heatmap of the differentially expressed genes corresponding to the GO term “immune system process” (GO: 0002376) is shown Figure 2.5A. In particular, Human Leukocyte Antigen Class I (HLA-I) gene expression was significantly upregulated in spheroid cells and downregulated in network cells (Table S2.1). Unexpectedly, immunofluorescence staining of HLA-A revealed distinct patterns of localization in each cell phenotype (Fig. 2.5B, C). Spheroid cells appeared to properly localize HLA-A to their plasma membrane (Fig. 2.5B), whereas invasive cells seemed to localize HLA-A near the nucleus (Fig. 2.5C). Quantification of this staining pattern demonstrated that spheroid cells had approximately three-fold less perinuclear co-localization of HLA-A compared to network cells (Fig. 2.5D).

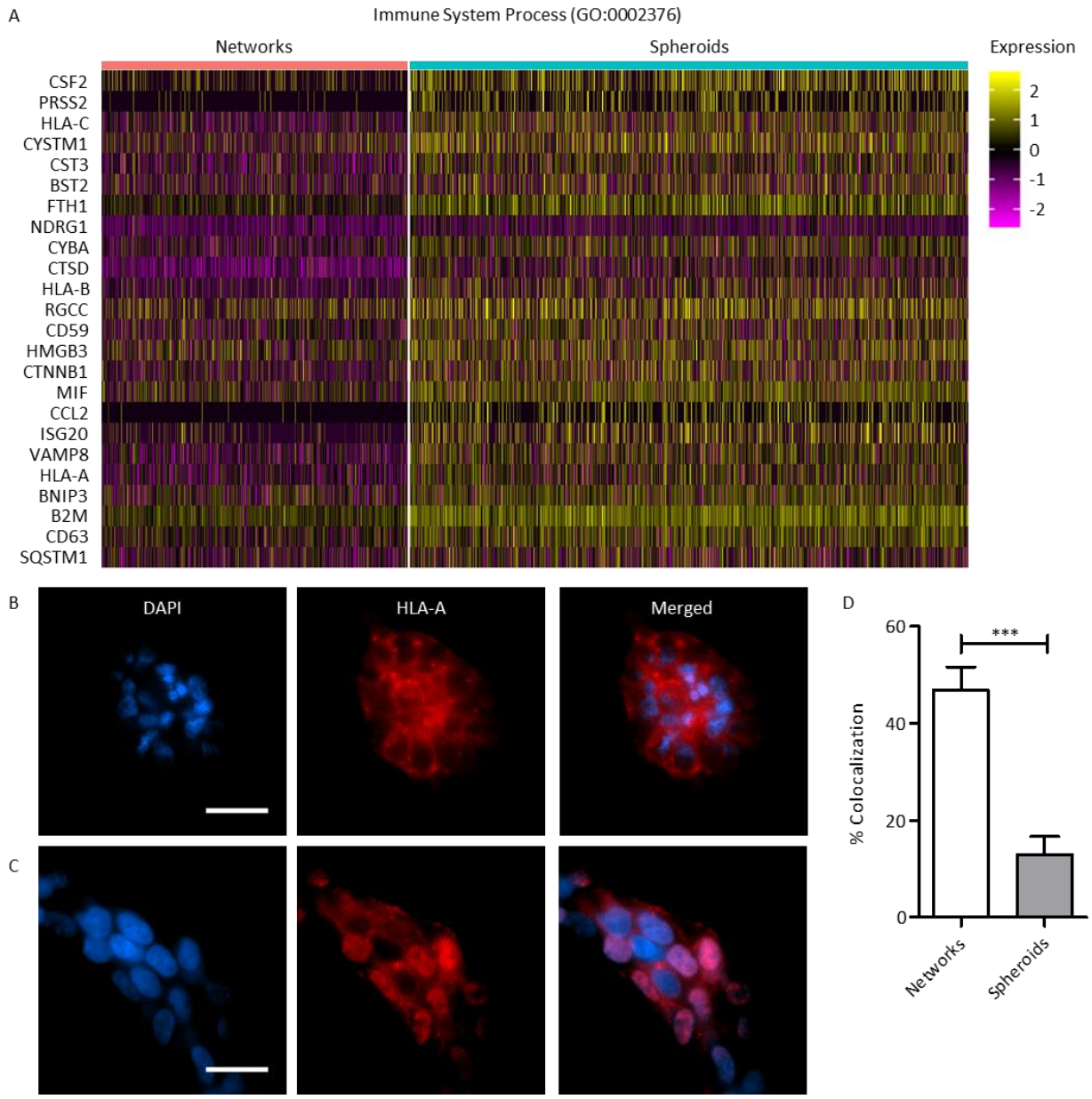


Figure 2.5. Spheroid cells display proper antigen localization and are more plastic. (**A**) A heatmap of the list of genes detected by phenotypically guided DE analysis that are in the GO:0002376 term. Spheroids display upregulation of genes associated with "Immune System Process". (**B**) Immunofluorescent staining of HLA-A in the spheroid cell population. Spheroid cells display membrane localization of HLA-A. Scale bar 50 µm. (**C**) Immunofluorescent staining of HLA-A in the network cell population. Many network cells display perinuclear staining of HLA-A. Scale bar 50 µm. (**D**) Quantification of the perinuclear staining of HLA-A within each collective phenotype.

Of the 101 genes upregulated in the network cells compared to the spheroid cells, we selected three targets for functional studies: *HSP90AB1*, *DEK*, and *F3*. This was motivated by the fact that GO enrichment analysis based on supervised DEGs identified many stress response terms,

many of which contained *HSP90AB1* or *F3*. (Table S2.1) In addition, several GO enrichment terms identified were associated with the regulation of gene expression, of which *DEK* is functionally associated. Based on their upregulation in networks, we hypothesized that their inhibition would negatively impact cell invasion and/or network formation, while their activation would positively promote cell invasion or network formation. Indeed, this is what we observed. Inhibition of *HSP90AB1* by Radicicol and *DEK* by Cordycepin significantly reduced the invasive potential of MDA cells and also reduced the rate of formation of collectively invasive networks (Figure 2.6). Conversely, supplementing the culture media with recombinant *F3*, which was upregulated in the network cells, increased the invasiveness of MDA cells (Figure 2.6A, B). However, adding recombinant *F3* did not increase the rate of network formation (Figure 2.6C, D). Thus, pheno-scRNAseq supervised gene expression analysis successfully predicted the roles that several DEGs played in regulating cell migration behavior.

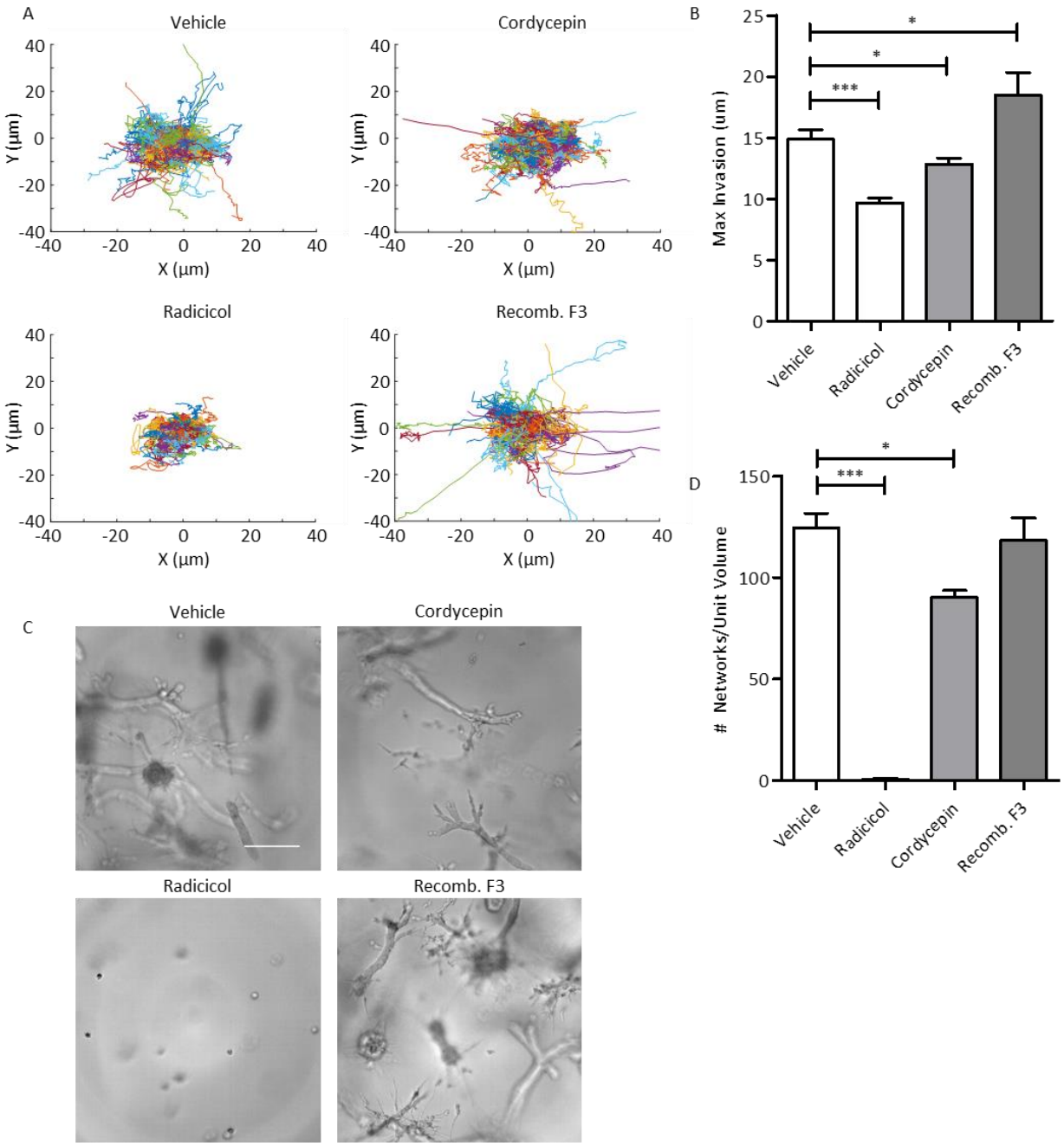


Figure 2.6. Inhibition of upregulated genes in the network phenotype reduces invasion. (A) Spider plots of cell trajectories during drug treatment. **(B)** Quantification of the maximum invasion of cells within each drug condition. Inhibition of the upregulated genes in the network phenotype reduced invasion. Adding recombinant *F3*, an upregulated network gene, increased invasion. **(C)** Representative brightfield images of the cells after 7 days of drug treatment. **(D)** Quantification of the number of network structures after 7 days of drug treatment.

2.3 Discussion

Our study demonstrates that unsupervised transcriptional clustering analysis does not necessarily separate cells according to specific functions, and thus may represent lost opportunities to decipher functional intra-tumour heterogeneity. Precision sorting prior to scRNA-seq, as implemented here via Dendra2-based phenotype marking, can help overcome limitations of unsupervised clustering. Analysis based on unsupervised clustering resulted in the identification of 528 DEGs, whereas analysis based on phenotypic labeling of each cell subpopulation resulted in the identification of 178 DEGs, and only 108 genes overlapped. Thus, roughly 80% of the DEGs identified by unsupervised clustering were not specific to the observed cell phenotypes. Importantly, phenotypic labeling allowed for the identification of 70 unique DEGs that were not detected by unsupervised clustering. We validated the functional relevance of one of these genes, Ki-67. This suggests that phenotypic labeling allows for a more direct and relevant approach to investigating the molecular regulators of functional heterogeneity within a given cell type. However, the results of our sorting and reseeding experiments could support the view that traditional unsupervised clustering analysis is more likely to give information about how cells could behave, rather than how they are currently behaving(*32*).

Phenotypically supervised scRNAseq based on the invasion phenotype of BRCA cells enabled us to gain a systems level view of distinct cancer cell states. Compared to confocal microscopy-based approaches, our setup significantly increased the number of cells that can be labelled and sorted within the limited timeframe before photoconverted protein turnover. A previous study was only able to photoconvert and sort tens of cells(*33*), demanding growth-based amplification of cells prior to omics analyses. Given the plasticity and rapid adaptation of cancer cells in response to changes in their environment, which occurs on short timescales that cannot be

explained by genetic evolution or clonal selection, the requirement for cell amplification could skew omics data and complicate our ability to directly link omics data to phenotypes(*46*). Our custom setup enabled us to photoconvert thousands of cells, resulting in a ten-fold increase in throughput while maintaining high spatial precision through the optimization of the photoconversion parameters. This higher throughput approach powered our statistical analysis, enabling us to directly compare the differences between unsupervised and phenotypically supervised scRNAseq, and led us to discover distinct biological processes that are linked to specific migration states. Both our transcriptional and functional evidence suggests that collective invasion is associated with anabolic metabolism, proliferation, redox stress, and ER stress. Conversely, collective acini formation is associated with lower levels of proliferation, quality control mechanisms, homeostasis, and immunomodulatory functions. Such knowledge may provide opportunities to design multiplexed therapeutic cocktails that take advantage of the susceptibilities of each complex cancer phenotype.

In conclusion, the field of single cell analysis is rapidly moving towards integrative, multi-scale measurements to improve the interpretability and actionable value of single-cell data. Our approach to enabling phenotypically supervised scRNAseq of BRCA cells reveals specific stress response and immunologic cellular processes that are coordinated with invasive and non-invasive phenotypes. Further, our 3D culture system and phenotypic cell sorting approach enable *in vitro* modelling of key aspects of phenotypic plasticity, which may provide a useful platform for the mechanistic dissection of these processes and identification of strategies that could effectively treat heterogeneous tumors. Understanding this plasticity may reveal ways to selectively target seemingly non-proliferative, non-invasive tumor cells while leaving quiescent noncancerous cells (such as normal stem cells) unharmed. For example, given the plasticity of the non-proliferative

spheroid cells, it may be possible to identify mechanisms capable of sensitizing them to chemotherapies or preventing a phenotypic switch to a more invasive, more proliferative state. Our system also offers the opportunity to dissect the mechanisms underlying tumor cell plasticity in the expression and localization of MHC Class I molecules, which could inform therapeutic strategies complementary to immunotherapies.

2.4 Methods

2.4.1 Cell Culture

MDA-MB-231 cells were a gift from Adam Engler (UCSD Bioengineering) and 4T1 cells were obtained from ATCC (Manassas, VA). All cells were cultured in high glucose Dulbecco's modified Eagle's medium supplemented with 10% (v/v) fetal bovine serum (FBS, Corning, Corning, NY) and 0.1% gentamicin (Gibco Thermofisher, Waltham, MA) and maintained at 37^{o}C and 5% CO_2 in a humidified environment during culture and imaging. The cells were passaged every 2-3 days. Cells were tested for mycoplasma contamination using the Mycoalert kit (Lonza, Basel, Switzerland).

To generate MDA-MB-231 cells that express Dendra2, we generated viral particles by cloning a Dendra2-Lifeact-7 plasmid (Addgene #54694, Watertown, MA) into a lentiviral vector. We then transfected the plasmid into lentiX293 T cells (Clonetech, Mountain View, CA. Cat #632180) along with packaging expressing plasmid (psPAX2, Addgene #12260) and envelope expressing plasmid (pMD2.G, Addgene #12259). Viral particles were collected at 48 h after transfection and they were purified by filtering through a 0.45 μm filter. MDA-MB-231 cells were then transduced with the viral particles in the presence of polybrene (Allele Biotechnology, San Diego, CA).

2.4.2 3D culture in type I collagen hydrogels

Cells embedded in 3D collagen matrices were prepared by mixing cells suspended in culture medium and 10× reconstitution buffer, 1:1 (v/v). Polyethylene glycol (PEG, Sigma-Aldrich, St. Louis, MO) was diluted in phosphate buffered saline (PBS, Gibco Thermofisher, Waltham, MA) and added to achieve a final concentration of 10 mg/mL. Soluble rat tail type I collagen in acetic acid (Corning, Corning, NY) was added to achieve a final concentration of 2.5 mg/mL. 6.25% of 1 M NaOH (volume of NaOH / volume of type I collagen) was used to normalize pH and the mixture was polymerized at 37 °C.

2.4.3 Phenotypic cell sorting

Collagen gels containing MDA Dendra cells were transferred to a microscope stage top incubator. Collective cell structures were identified using a Nikon TiE fluorescent microscope (Nikon Instruments Inc., Melville, NY). Regions of interest were outlined using NIS-Elements software, and a Galvo Miniscanner (Nikon Instruments Inc.) was used to control the exposure of 405 nm laser from a Nikon LUnA power source (Nikon Instruments Inc.) to the outlined region to photoconvert the selected cells. 25% laser power with a 300 us dwell time were used to photoconvert the cells. This results in less than 1mJ of energy delivered to each multicellular structure. To ensure fidelity of converting only the desired phenotype, we did not photoconvert overlapping cell structures. The collagen gel was then digested using collagenase for 15 minutes at 37^{o}C (Sigma-Aldrich) and the cells were resuspended in FACS buffer (1% BSA, 0.5 mM EDTA in PBS). A gel with cells that were not photoconverted was used as a sorting control. Cells were sorted at the stem cell core of Sanford Consortium of Regenerative Medicine (La Jolla, CA) using a BD Influx cell sorter (BD, Franklin lakes, NJ). The cells from the control gel were used

to establish a negative gate, and cells expressing red fluorescence above that gate were collected for re-culture or sequencing.

2.4.4 Single cell sequencing and analysis

RNA extraction and library construction were performed using the Chromium Single Cell 3' v3 kit (10x genomics, Pleasanton, CA). At least 700 cells were extracted for the network and spheroid cells, with one independent experiment each. These two experiments were pooled together for sequencing. The non-sorted MDA cells and the 4T1s were also extracted from one independent experiment each, and these were pooled together for sequencing. RNA integrity was verified using RNA Analysis ScreenTape (Agilent Technologies, La Jolla, CA) before sequencing. The RNA was sequenced on the Illumina HiSeq 4000 at a depth of > 20,000 reads per cell. The cellRanger analysis pipeline was used to construct the human reference genome GRCh38 and align reads. Differential gene expression analysis was performed using Seurat(*47*). Cell expression data was filtered, log-normalized, and scaled prior to differential expression analysis using the non-parametric Wilcoxon rank sum test. Differentially expressed genes were filtered for those that had an absolute log-fold change of > 0.25 and expressed in at least 10% of either subpopulation with an adjusted $P < 0.05$ (Bonferroni).

2.4.5 Gene ontology term overrepresentation analysis

Gene ontology enrichment analysis was performed using Panther with the GO biological process complete annotation set(*15*). The Fisher's Exact test was used and the significance level was set at 0.05. The false discovery rate (FDR) was calculated to correct for multiple testing. The fold enrichment is the observed number of genes in our dataset associated with the term divided by the expected number of genes associated with the term.

2.4.6 Cell tracking and phenotypic analysis

Collagen gels containing cells were imaged every 10 min for 60 hours. Coordinates of the cell location at each time frame were determined by tracking single cells using image recognition software (Metamorph/Metavue, Molecular Devices, Sunnyvale, CA). All cell tracking data comes from three independent experiments performed on different days and with different cell passages. Morphologic measurements and phenotype quantification were performed in NIS-Elements using at least three independent experiments performed on different days with different cell passages.

2.4.7 Immunofluorescence and cell imaging

For cell imaging after 7 days of culture to visualize collective phenotypes, collagen gels were fixed with 4% PFA for 30 min at room temperature. F-actin was stained using AlexaFluor® 488 Phalloidin (Cell Signaling Technology, Danvers, MA) and the nuclei were counterstained with DAPI. For immunofluorescence staining the gels were incubated with the primary antibody for 24 hrs at 4^{o}C. The antibodies used were anti-COL4A1 (1:200 dilution, NB120-6586, Novus Biologicals, Littleton, CO) , anti-LAMC2 (1:200 dilution, MAB19562, Millipore Sigma), anti-Ki-67 (1:400 dilution, 8D5, Cell Signaling Technology), anti-HLA-A (1:100 dilution, ab52922, Abcam, Cambridge, UK).

2.4.8 Paclitaxel treatment and susceptibility analysis

After 7 days of culture in collagen gels, MDAs were treated with 1 μM Paclitaxel (Cell Signaling Technolgy) or DMSO vehicle control. Cell death was assessed 3 days later using a Live/Dead Cell Assay (Abcam, Cambridge, UK) with Hoescht counter-stain. Z-stack imaging was conducted and the total number of nuclei forming either spheroids or networks were

quantified as well as the number of nuclei with red co-localization (dead cell dye). % Death is calculated as the number of nuclei with co-localized red fluorescence over the total number of nuclei. Measurements were repeated across 3 biological replicates per condition equating to ~300 cells analyzed per condition.

2.4.9 Radicicol, Cordycepin, and recombinant F3 treatment

MDAs were treated with 10 μM Radicicol (Cayman Chemical, Ann Arbor, MI), 50 μM Cordycepin (MedChemExpress, Monmouth Junction, NJ), 1 μg/mL F3 (BioLegend, San Diego, CA), or DMSO vehicle control (Sigma-Aldrich) after embedment into 3D COL1 matrices. Cell invasion and the formation of the network phenotype was quantified as mentioned above. Measurements were repeated across 3 biological replicates per condition equating to at least 100 cells analyzed per condition.

2.4.10 Statistical Analysis

All measurements were analyzed using Prism (Graphpad, San Diego, CA). Significance (P) was indicated within the figures using the following scale: *P<0.05, **P<0.01, ***P<0.001. Unless otherwise noted, bar graphs show mean and s.e.m. of quantified variables. Statistical analyses between two groups were performed with Student's unpaired t tests and between three or more groups with one-way analysis of variance (ANOVA), followed by Tukey's multiple comparison post hoc test to determine significance. Additional relevant information is detailed in the figure captions.

2.5 Supplementary material

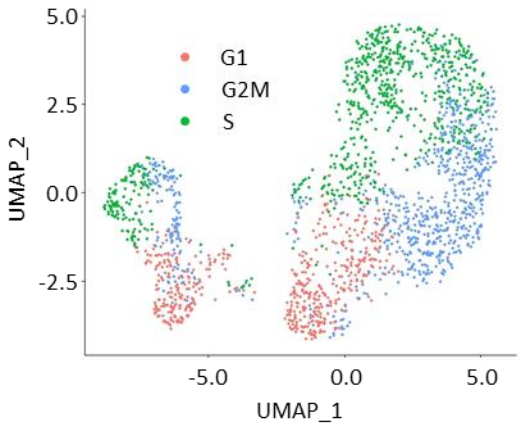


Figure S2.1. Cell cycle state analysis of MDA cells. Spheroid cells lie in multiple cell cycle states and are not simply a subpopulation that have withdrawn from the cell cycle.

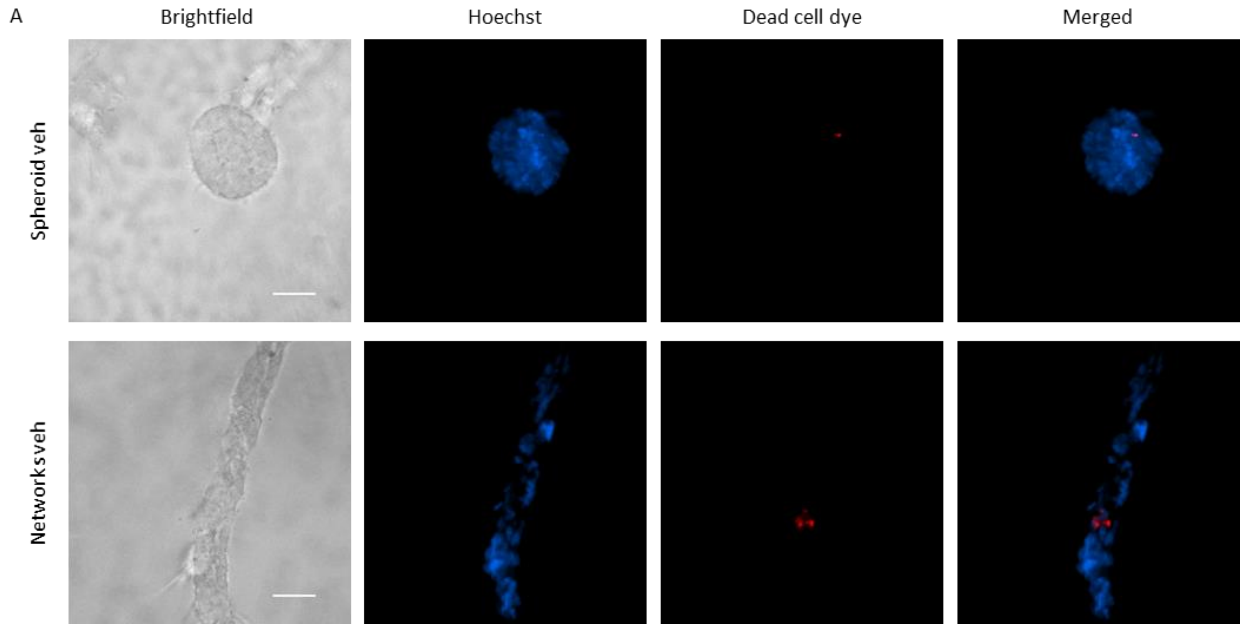


Figure S2.2. Brightfield and fluorescence images after treatment with vehicle control of spheroids and networks corresponding to the Paclitaxel experiments. Scale bar 50 µm.

Table S2.1. Differential Gene Expression Based on Phenotypic Labels. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels.

Gene	Average log (Fold Change)	Fraction of Network Cells in which Gene was Detected	Fraction of Spheroid Cells in which Gene was Detected	Adjusted P Value
ABCE1	0.347206818	0.914	0.755	4.87E-39
ACO1	0.26708818	0.711	0.456	1.04E-28

Table S2.1. Differential Gene Expression Based on Phenotypic Labels. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

Gene	Average log (Fold Change)	Fraction of Network Cells in which Gene was Detected	Fraction of Spheroid Cells in which Gene was Detected	Adjusted P Value
AGR2	-0.474297729	0.094	0.321	1.99E-28
AHCTF1	0.328163761	0.717	0.443	1.97E-38
ALCAM	0.292156653	0.9	0.721	1.33E-31
ANLN	0.339629831	0.781	0.567	1.95E-21
ANP32B	0.25347614	0.984	0.947	2.50E-33
AREG	-0.413443519	0.804	0.848	8.28E-07
ASPM	0.349340874	0.717	0.535	1.58E-14
ATAD2	0.404334004	0.785	0.541	9.09E-31
B2M	-0.34555291	0.996	0.999	5.07E-56
BNIP3	-0.269784306	0.984	0.985	7.87E-20
BRCA2	0.259307656	0.633	0.391	2.07E-24
BST2	-0.2684989	0.543	0.626	7.68E-08
CACYBP	0.259391683	0.978	0.919	7.76E-30
CAVIN3	-0.339054269	0.938	0.945	7.19E-23
CCL2	-0.749475941	0.01	0.19	8.22E-28
CCNA2	0.290266153	0.636	0.468	4.83E-14
CCT3	0.268504626	0.982	0.94	1.38E-33
CCT5	0.274809261	0.992	0.969	8.88E-39
CCT6A	0.30469092	0.99	0.971	1.27E-46
CD46	0.281689121	0.892	0.722	1.11E-29
CD59	-0.261255954	0.891	0.889	1.25E-12
CD63	-0.273323731	0.989	0.988	9.40E-32
CD82	-0.393028534	0.267	0.479	1.94E-23
CD9	-0.360062881	0.926	0.949	1.14E-34
CDC37	0.288304529	0.947	0.824	2.16E-31
CENPF	0.351900186	0.873	0.742	6.02E-16
CENPU	0.277812138	0.785	0.591	4.87E-21
CEP55	0.268813628	0.739	0.557	2.42E-14
CHML	0.311510528	0.695	0.432	2.40E-33
CLTC	0.298014182	0.926	0.787	5.48E-35
CNN3	0.318594305	0.982	0.927	2.92E-41
COL8A1	0.297120783	0.674	0.468	6.69E-20
CSF2	-0.500910691	0.19	0.38	1.17E-18
CST1	-0.446218659	0.304	0.447	3.18E-08
CST3	-0.403390583	0.982	0.993	3.95E-28

Table S2.1. Differential Gene Expression Based on Phenotypic Labels. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

Gene	Average log (Fold Change)	Fraction of Network Cells in which Gene was Detected	Fraction of Spheroid Cells in which Gene was Detected	Adjusted P Value
CST4	-0.491148213	0.272	0.452	4.08E-15
CTGF	-0.410392483	0.828	0.839	1.42E-08
CTNNB1	-0.280733264	0.573	0.646	1.68E-11
CTSD	-0.61140053	0.856	0.942	2.75E-56
CYBA	-0.356448119	0.923	0.946	6.94E-35
CYSTM1	-0.291234014	0.529	0.659	2.10E-17
DDX3X	0.391475003	0.926	0.793	3.99E-53
DEK	0.260632314	0.993	0.982	8.46E-35
DHX9	0.350386545	0.829	0.583	3.30E-42
DLGAP5	0.294108732	0.752	0.614	7.51E-08
DNAJC9	0.252727121	0.866	0.694	3.25E-20
DNMT1	0.334549273	0.951	0.831	2.38E-40
DSP	0.33297117	0.77	0.512	2.25E-35
DST	0.364716501	0.984	0.919	1.65E-52
DYNC1H1	0.287017494	0.977	0.919	5.94E-32
ECH1	-0.26986806	0.679	0.735	1.04E-12
ECT2	0.28831074	0.839	0.694	1.16E-17
EIF4G1	0.250803504	0.922	0.796	1.22E-24
EIF4G2	0.282150388	0.986	0.946	1.32E-36
EREG	-0.346952505	0.46	0.614	2.96E-14
ETFB	-0.341502218	0.743	0.829	2.56E-24
EXOC5	0.274229527	0.744	0.487	6.63E-31
F3	0.485697637	0.847	0.61	1.74E-37
FAT1	0.417340258	0.832	0.586	1.24E-51
FOS	0.276366938	0.863	0.817	1.69E-06
FTH1	-0.376801826	1	0.999	1.11E-48
G0S2	0.512380975	0.959	0.855	1.05E-36
G3BP2	0.269202064	0.926	0.767	2.94E-27
GLRX	-0.259483568	0.781	0.825	1.47E-11
HIF1A	0.289935623	0.874	0.701	7.32E-28
HIST1H1C	0.354032061	0.888	0.797	1.06E-20
HIST1H1E	1.098513765	0.885	0.556	6.18E-89
HIST1H2AC	0.310698099	0.544	0.419	0.000233978
HLA-A	-0.358114507	0.962	0.968	1.08E-35
HLA-B	-0.395354336	0.487	0.641	5.48E-22

Table S2.1. Differential Gene Expression Based on Phenotypic Labels. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

Gene	Average log (Fold Change)	Fraction of Network Cells in which Gene was Detected	Fraction of Spheroid Cells in which Gene was Detected	Adjusted P Value
HLA-C	-0.429845022	0.644	0.783	3.50E-35
HMGB3	-0.256795934	0.787	0.808	8.34E-06
HNRNPAB	0.253644515	0.941	0.864	1.73E-21
HNRNPD	0.331844329	0.988	0.931	3.42E-56
HNRNPU	0.52257376	0.997	0.939	1.95E-114
HPCAL1	-0.329427981	0.906	0.932	2.81E-19
HSP90AA1	0.25010694	0.999	1	4.23E-39
HSP90AB1	0.273648191	1	0.994	8.37E-49
HSPA4	0.387806247	0.952	0.831	4.82E-51
HSPD1	0.301665804	0.988	0.94	2.53E-37
IARS	0.263170065	0.798	0.601	1.35E-23
IGFBP1	-0.545270718	0.814	0.858	1.73E-19
IGFBP4	-0.520106708	0.97	0.976	5.15E-46
ILF3	0.331310214	0.943	0.819	4.88E-43
IPO5	0.27434702	0.804	0.588	2.54E-30
ISG20	-0.38074709	0.124	0.359	4.58E-29
KCTD20	0.252870202	0.804	0.645	1.13E-20
KDM1A	0.311170882	0.773	0.51	1.99E-36
KIF11	0.286489068	0.636	0.438	3.63E-17
KRT19	-0.290049977	0.978	0.978	1.73E-10
KRT81	-0.458348079	0.148	0.321	6.97E-16
LBR	0.418727156	0.821	0.555	3.86E-48
LMNB1	0.268602966	0.694	0.475	5.78E-21
LRP10	-0.252457563	0.687	0.732	2.39E-08
LRPPRC	0.26829432	0.705	0.461	5.53E-26
MAGED2	-0.330929038	0.464	0.609	2.72E-17
MALAT1	-0.401932229	0.996	0.996	2.28E-34
MAP4	0.258236162	0.962	0.861	7.94E-26
MARCKSL1	-0.365377223	0.917	0.921	5.14E-10
MCM4	0.266754101	0.632	0.372	1.57E-26
MIF	-0.291815881	0.876	0.917	6.44E-32
MKI67	0.30526856	0.822	0.667	2.06E-14
MSH6	0.317505341	0.673	0.481	1.06E-18
MT-ND6	0.251673496	0.886	0.756	2.98E-18
MYADM	-0.348677494	0.813	0.872	3.23E-26

Table S2.1. Differential Gene Expression Based on Phenotypic Labels. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

Gene	Average log (Fold Change)	Fraction of Network Cells in which Gene was Detected	Fraction of Spheroid Cells in which Gene was Detected	Adjusted P Value
MYOF	0.258004758	0.9	0.787	3.67E-21
NCL	0.404556063	0.996	0.979	1.98E-71
NDRG1	-0.318450145	0.49	0.617	1.43E-14
NEAT1	-0.337521898	0.944	0.932	1.63E-05
NFKBIA	-0.304129907	0.711	0.746	0.00029272
NPM1	0.405487433	0.995	0.997	1.21E-90
NUDCD1	0.312156487	0.821	0.588	1.59E-30
OGFRL1	0.252895192	0.829	0.679	7.45E-18
PABPC1	0.26610874	0.999	0.974	1.39E-39
PABPC4	0.267822305	0.845	0.656	1.88E-27
PAICS	0.306392867	0.94	0.801	1.10E-34
PARP1	0.312274018	0.891	0.729	5.50E-35
PAWR	0.293207369	0.865	0.728	3.08E-23
PGK1	-0.286153853	0.944	0.96	3.86E-27
PHLDA2	-0.354885605	0.993	0.996	6.69E-39
PLAT	-0.253475682	0.622	0.688	0.00014821
PLS3	0.377871026	0.921	0.722	4.23E-45
PNN	0.272055701	0.963	0.899	1.75E-23
PNPT1	0.263850665	0.663	0.461	7.74E-20
PRDX5	-0.336608283	0.955	0.965	1.78E-41
PRKDC	0.353014368	0.948	0.838	3.35E-46
PRPF4B	0.265643014	0.921	0.777	3.34E-23
PRSS2	-0.501911066	0.026	0.204	2.20E-24
PRSS23	-0.397352707	0.814	0.892	1.13E-26
PTPN11	0.310242687	0.914	0.767	2.30E-35
RABAC1	-0.2863141	0.848	0.892	2.65E-18
RAD21	0.256105486	0.926	0.819	8.29E-18
RGCC	-0.275200396	0.409	0.535	5.53E-08
RPN2	-0.251292729	0.892	0.896	2.22E-16
RRAD	-0.408790811	0.399	0.569	3.05E-14
S100A10	-0.250126001	0.996	0.999	3.65E-22
S100A4	-0.403308415	0.956	0.919	1.30E-05
SACS	0.283234153	0.65	0.374	3.13E-34
SAT1	-0.36329149	0.707	0.784	1.52E-09
SEC61G	-0.288944238	0.966	0.97	5.75E-33

Table S2.1. Differential Gene Expression Based on Phenotypic Labels. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

Gene	Average log (Fold Change)	Fraction of Network Cells in which Gene was Detected	Fraction of Spheroid Cells in which Gene was Detected	Adjusted P Value
SELENOS	-0.282263325	0.698	0.783	7.02E-15
SLC25A24	0.259890193	0.84	0.648	1.49E-22
SLCO4A1	-0.583997806	0.706	0.826	2.26E-35
SLCO4A1-AS1	-0.517171935	0.352	0.568	3.32E-27
SMARCA5	0.345306426	0.889	0.726	4.74E-42
SNRNP200	0.260360973	0.769	0.558	4.99E-25
SPANXB1	-1.20541394	0.094	0.414	4.24E-52
SPTAN1	0.302157264	0.795	0.563	2.00E-29
SQSTM1	-0.281830289	0.843	0.852	1.92E-05
SRPK1	0.331147302	0.977	0.904	6.46E-45
SSR4	-0.27144288	0.944	0.952	6.05E-24
STC1	-0.372804746	0.332	0.456	5.46E-09
SUPT16H	0.270895059	0.933	0.803	4.86E-28
SUZ12	0.258708104	0.866	0.684	2.82E-25
TARS	0.260755944	0.943	0.839	1.56E-26
TFF1	-0.304905865	0.066	0.183	1.20E-09
TFF2	-0.447642564	0.137	0.328	6.35E-18
TFF3	-0.398193906	0.105	0.289	2.55E-18
TFPI	-0.261055313	0.68	0.731	3.98E-09
TFRC	0.306500229	0.685	0.474	6.83E-26
TIMP1	-0.515981814	0.911	0.945	1.70E-45
TLN1	0.291201795	0.96	0.852	3.11E-36
TM4SF1	-0.380554119	0.951	0.974	7.77E-35
TMEM158	-0.459986631	0.845	0.852	7.61E-09
TMEM30A	0.253429377	0.737	0.503	1.00E-23
TMEM59	-0.29794336	0.863	0.905	5.05E-25
TMPO	0.414854545	0.952	0.819	7.91E-44
TMSB10	-0.336427758	0.999	1	3.68E-97
TMSB4X	-0.32898135	0.999	0.999	3.47E-65
TPI1	-0.280628212	0.982	0.991	1.01E-41
TPM4	0.255413058	1	0.993	3.98E-43
TPR	0.30721659	0.978	0.912	5.85E-40
TPX2	0.374838719	0.832	0.679	2.64E-17
TSC22D1	-0.262388537	0.949	0.94	4.98E-11
VAMP8	-0.406914311	0.938	0.96	6.67E-50
XPO1	0.372002023	0.866	0.67	4.28E-44

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
activation of protein kinase activity (GO:0032147)	100	2.84
aggrephagy (GO:0035973)	35.9	2.67
aging (GO:0007568)	0.0942	3.74E-03
ameboidal-type cell migration (GO:0001667)	100	2.29
amide biosynthetic process (GO:0043604)	95.5	1.06E-08
amide transport (GO:0042886)	27.1	3.43E-03
anaphase (GO:0051322)	8.25	2.42
anaphase-promoting complex-dependent catabolic process (GO:0031145)	100	4.42
anatomical structure development (GO:0048856)	0.0113	0.108
anatomical structure formation involved in morphogenesis (GO:0048646)	27.4	0.0601
anatomical structure homeostasis (GO:0060249)	0.0234	29.3
angiogenesis (GO:0001525)	17.4	4.63E-04
animal organ development (GO:0048513)	0.832	0.0274
animal organ morphogenesis (GO:0009887)	96	4.79
antigen processing and presentation (GO:0019882)	0.296	0.0156
antigen processing and presentation of endogenous antigen (GO:0019883)	0.721	1.91
antigen processing and presentation of endogenous peptide antigen (GO:0002483)	0.283	3.78
antigen processing and presentation of endogenous peptide antigen via MHC class I (GO:0019885)	0.282	3.77
antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway (GO:0002484)	0.679	4.16
antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway, TAP-independent (GO:0002486)	0.682	4.16
antigen processing and presentation of exogenous antigen (GO:0019884)	0.087	5.21E-03
antigen processing and presentation of exogenous peptide antigen (GO:0002478)	0.0627	2.99E-03
antigen processing and presentation of exogenous peptide antigen via MHC class I (GO:0042590)	0.721	0.0162
antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent (GO:0002479)	0.56	9.44E-03
antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent (GO:0002480)	0.0411	0.661
antigen processing and presentation of peptide antigen (GO:0048002)	0.115	8.96E-03
antigen processing and presentation of peptide antigen via MHC class I (GO:0002474)	1.71	0.025
apoptotic process (GO:0006915)	5.3	0.0101
aromatic compound biosynthetic process (GO:0019438)	70.6	0.0174

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
aromatic compound catabolic process (GO:0019439)	100	2.08E-08
ATP-dependent chromatin remodeling (GO:0043044)	40.1	0.931
biological adhesion (GO:0022610)	77.4	3.21
biological phase (GO:0044848)	20.6	0.0596
biological regulation (GO:0065007)	3.36E-05	3.26E-13
biological_process (GO:0008150)	7.18E-03	1.32E-12
biosynthetic process (GO:0009058)	59	6.89E-07
blood coagulation (GO:0007596)	5.2	2.78
blood coagulation, extrinsic pathway (GO:0007598)	4.14	12.1
blood vessel development (GO:0001568)	70.6	0.0156
blood vessel morphogenesis (GO:0048514)	40	6.66E-04
bone development (GO:0060348)	100	1.41
bone morphogenesis (GO:0060349)	100	0.926
branch elongation of an epithelium (GO:0060602)	91.9	3.27
branching morphogenesis of an epithelial tube (GO:0048754)	100	2.09
bronchus development (GO:0060433)	71.6	0.871
catabolic process (GO:0009056)	100	1.57E-05
cell activation (GO:0001775)	3.11E-05	7.08E-11
cell activation involved in immune response (GO:0002263)	3.22E-04	8.79E-08
cell adhesion (GO:0007155)	100	4.25
cell communication (GO:0007154)	3.15	0.117
cell cycle (GO:0007049)	7.33E-03	0.239
cell cycle phase (GO:0022403)	20.6	0.0592
cell cycle process (GO:0022402)	4.25E-03	0.307
cell death (GO:0008219)	0.867	7.70E-04
cell development (GO:0048468)	3.32	0.411
cell differentiation (GO:0030154)	1.36E-03	2.96
cell division (GO:0051301)	0.327	0.274
cell migration (GO:0016477)	80.5	0.111
cell motility (GO:0048870)	100	1.12
cell population proliferation (GO:0008283)	24	0.273
cell redox homeostasis (GO:0045454)	7.21	2.27E-04
cell surface receptor signaling pathway (GO:0007166)	1.23	1.27E-03
cellular amide metabolic process (GO:0043603)	100	3.85E-08
cellular aromatic compound metabolic process (GO:0006725)	7.96	7.26E-10
cellular biosynthetic process (GO:0044249)	62.7	1.23E-06

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
cellular catabolic process (GO:0044248)	100	9.81E-06
cellular component assembly (GO:0022607)	4.82E-03	5.10E-05
cellular component biogenesis (GO:0044085)	5.21E-03	1.78E-07
cellular component disassembly (GO:0022411)	35.6	0.313
cellular component organization (GO:0016043)	5.42E-07	2.71E-09
cellular component organization or biogenesis (GO:0071840)	1.30E-06	1.40E-11
cellular detoxification (GO:1990748)	100	4.16
cellular developmental process (GO:0048869)	1.00E-04	0.979
cellular localization (GO:0051641)	3.02E-05	2.89E-16
cellular macromolecule biosynthetic process (GO:0034645)	17	2.54E-06
cellular macromolecule catabolic process (GO:0044265)	100	9.95E-06
cellular macromolecule localization (GO:0070727)	0.956	2.93E-06
cellular macromolecule metabolic process (GO:0044260)	1.48	9.93E-10
cellular metabolic process (GO:0044237)	7.57	3.39E-12
cellular nitrogen compound biosynthetic process (GO:0044271)	15.8	2.01E-12
cellular nitrogen compound catabolic process (GO:0044270)	100	7.63E-09
cellular nitrogen compound metabolic process (GO:0034641)	14.2	1.99E-12
cellular process (GO:0009987)	1.10E-04	1.27E-24
cellular protein localization (GO:0034613)	0.91	2.56E-06
cellular protein metabolic process (GO:0044267)	42.2	9.77E-05
cellular protein-containing complex assembly (GO:0034622)	6.01E-03	4.93E-07
cellular response to biotic stimulus (GO:0071216)	9.69	0.322
cellular response to chemical stimulus (GO:0070887)	2.63E-05	4.73E-15
cellular response to chemical stress (GO:0062197)	6.86E-03	0.782
cellular response to cytokine stimulus (GO:0071345)	1.74E-05	3.22E-11
cellular response to decreased oxygen levels (GO:0036294)	7.13	0.14
cellular response to endogenous stimulus (GO:0071495)	5.68	0.066
cellular response to growth factor stimulus (GO:0071363)	47.9	9.24E-03
cellular response to heat (GO:0034605)	1.23	8.13
cellular response to hormone stimulus (GO:0032870)	8.07	4.15
cellular response to hypoxia (GO:0071456)	6.02	0.0842
cellular response to inorganic substance (GO:0071241)	1.19	12.5
cellular response to interferon-gamma (GO:0071346)	27.1	3.16
cellular response to interleukin-1 (GO:0071347)	30.9	0.0327
cellular response to interleukin-7 (GO:0098761)	100	0.788
cellular response to lipid (GO:0071396)	5.72	0.105

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
cellular response to lipopolysaccharide (GO:0071222)	5.41	0.501
cellular response to molecule of bacterial origin (GO:0071219)	6.45	0.292
cellular response to nitrogen compound (GO:1901699)	6.98	3.37
cellular response to organic cyclic compound (GO:0071407)	2.87	1.52
cellular response to organic substance (GO:0071310)	1.07E-04	1.91E-14
cellular response to oxidative stress (GO:0034599)	6.39E-03	0.319
cellular response to oxygen levels (GO:0071453)	9.52	0.309
cellular response to oxygen-containing compound (GO:1901701)	0.569	9.42E-03
cellular response to peptide (GO:1901653)	65.2	1.91
cellular response to reactive oxygen species (GO:0034614)	1.23	5.45
cellular response to stimulus (GO:0051716)	0.078	2.26E-08
cellular response to stress (GO:0033554)	1.36E-03	4.97E-06
cellular response to topologically incorrect protein (GO:0035967)	20.8	4.52E-03
cellular response to transforming growth factor beta stimulus (GO:0071560)	100	0.0731
cellular response to tumor necrosis factor (GO:0071356)	59	0.92
cellular response to type I interferon (GO:0071357)	2.21	1.72
cellular response to unfolded protein (GO:0034620)	13.8	3.53E-03
CENP-A containing chromatin organization (GO:0061641)	8.04	3.28
CENP-A containing nucleosome assembly (GO:0034080)	8.08	3.29
central nervous system development (GO:0007417)	12.1	3.35
centromere complex assembly (GO:0034508)	2.04	1.62
chaperone-mediated protein complex assembly (GO:0051131)	0.444	5.55
chaperone-mediated protein folding (GO:0061077)	100	0.718
chromatin assembly (GO:0031497)	6.64	0.171
chromatin assembly or disassembly (GO:0006333)	3.28	0.174
chromatin organization (GO:0006325)	0.156	4.34
chromatin remodeling (GO:0006338)	63.8	0.376
chromatin remodeling at centromere (GO:0031055)	9.52	4.43
chromosome organization (GO:0051276)	1.41E-03	0.101
circulatory system development (GO:0072359)	25.5	0.0752
coagulation (GO:0050817)	5.41	2.85
common-partner SMAD protein phosphorylation (GO:0007182)	100	2.53
cotranslational protein targeting to membrane (GO:0006613)	100	9.47E-10
craniofacial suture morphogenesis (GO:0097094)	100	3.76
cytokine-mediated signaling pathway (GO:0019221)	0.0169	1.14E-07

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
cytoplasmic sequestering of NF-kappaB (GO:0007253)	0.676	4.15
cytoplasmic sequestering of transcription factor (GO:0042994)	2.84	15.7
cytoplasmic translation (GO:0002181)	100	1.60E-03
cytoskeleton organization (GO:0007010)	0.0351	6.22
defense response (GO:0006952)	1.06	0.225
detection of chemical stimulus (GO:0009593)	100	0.165
detection of chemical stimulus involved in sensory perception (GO:0050907)	100	0.0426
detection of chemical stimulus involved in sensory perception of smell (GO:0050911)	41.3	0.116
detection of stimulus involved in sensory perception (GO:0050906)	100	0.0854
developmental growth (GO:0048589)	0.338	0.394
developmental process (GO:0032502)	7.78E-05	7.65E-03
DNA biosynthetic process (GO:0071897)	31.2	2.73
DNA conformation change (GO:0071103)	0.181	1.54E-03
DNA damage response, detection of DNA damage (GO:0042769)	100	1.77
DNA metabolic process (GO:0006259)	2.06	0.0327
DNA packaging (GO:0006323)	5.65	0.0233
DNA recombination (GO:0006310)	3.73	12.3
DNA replication-independent nucleosome assembly (GO:0006336)	13.1	1.95
DNA replication-independent nucleosome organization (GO:0034724)	2.53	2.13
embryo implantation (GO:0007566)	53.6	2.52
endochondral bone morphogenesis (GO:0060350)	100	2
endoderm development (GO:0007492)	94.4	0.804
endoderm formation (GO:0001706)	63.8	1.45
endoplasmic reticulum unfolded protein response (GO:0030968)	25.5	1.63E-03
endothelial cell differentiation (GO:0045446)	42	3.34
endothelium development (GO:0003158)	51.9	0.809
entry into host (GO:0044409)	27.1	1.69
enzyme linked receptor protein signaling pathway (GO:0007167)	45.4	0.253
epithelial cell differentiation (GO:0030855)	12.3	2.96
epithelial cell migration (GO:0010631)	100	1.61
epithelial cell proliferation (GO:0050673)	18.6	0.566
epithelium development (GO:0060429)	23.9	1.71
epithelium migration (GO:0090132)	100	1.92
ERBB signaling pathway (GO:0038127)	2.34	0.493
ERBB2 signaling pathway (GO:0038128)	8.97	0.897

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
ER-nucleus signaling pathway (GO:0006984)	12.4	1.77
establishment of Golgi localization (GO:0051683)	100	4.15
establishment of localization (GO:0051234)	5.23E-04	3.64E-12
establishment of localization in cell (GO:0051649)	1.28E-04	4.96E-16
establishment of organelle localization (GO:0051656)	1.19	31.2
establishment of protein localization (GO:0045184)	11.2	4.67E-04
establishment of protein localization to endoplasmic reticulum (GO:0072599)	100	2.13E-10
establishment of protein localization to membrane (GO:0090150)	100	2.35E-05
establishment of protein localization to organelle (GO:0072594)	28.8	1.69E-06
establishment of RNA localization (GO:0051236)	0.155	23.8
exocrine system development (GO:0035272)	100	3.49
exocytosis (GO:0006887)	1.10E-04	1.01E-10
export from cell (GO:0140352)	2.72E-03	1.20E-09
extracellular matrix organization (GO:0030198)	100	1.7
extracellular structure organization (GO:0043062)	100	1.72
gene expression (GO:0010467)	1.51	9.79E-09
generation of precursor metabolites and energy (GO:0006091)	73.9	1.91
gland development (GO:0048732)	75.4	0.242
gland morphogenesis (GO:0022612)	100	1.35
glial cell development (GO:0021782)	100	2.5
gliogenesis (GO:0042063)	4.63	0.567
global genome nucleotide-excision repair (GO:0070911)	100	2.14
granulocyte activation (GO:0036230)	1.84E-04	2.62E-08
granulocyte chemotaxis (GO:0071621)	100	3.97
growth (GO:0040007)	0.115	0.446
hair cycle (GO:0042633)	100	2.03
hair cycle process (GO:0022405)	94.2	2.66
hair follicle development (GO:0001942)	90.6	2.24
hematopoietic or lymphoid organ development (GO:0048534)	11	0.0856
hemopoiesis (GO:0030097)	7.11	0.21
hemostasis (GO:0007599)	5.56	3.02
heterocycle biosynthetic process (GO:0018130)	71.8	0.0156
heterocycle catabolic process (GO:0046700)	100	7.27E-09
heterocycle metabolic process (GO:0046483)	7.5	2.30E-10
histone exchange (GO:0043486)	19.6	0.231

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
histone H3-K27 methylation (GO:0070734)	1.23	91.5
homeostatic process (GO:0042592)	7.04	0.142
immune effector process (GO:0002252)	8.84E-04	5.24E-07
immune response (GO:0006955)	0.385	1.05E-03
immune system development (GO:0002520)	3.97	0.0301
immune system process (GO:0002376)	1.35E-03	9.46E-09
immunoglobulin production involved in immunoglobulin mediated immune response (GO:0002381)	34.5	2.41
inflammatory response (GO:0006954)	34.4	1.62
integrated stress response signaling (GO:0140467)	14.3	1.41
integrin-mediated signaling pathway (GO:0007229)	24	3.38
interaction with host (GO:0051701)	10.7	0.215
interferon-gamma-mediated signaling pathway (GO:0060333)	10.6	1.97
interleukin-1-mediated signaling pathway (GO:0070498)	100	0.868
interspecies interaction between organisms (GO:0044419)	1.61E-05	1.79E-13
intracellular protein transport (GO:0006886)	31	3.97E-05
intracellular transport (GO:0046907)	7.27	3.96E-05
IRE1-mediated unfolded protein response (GO:0036498)	100	0.105
iron ion homeostasis (GO:0055072)	4.21	96.3
kinetochore organization (GO:0051383)	95.2	3.77
leukocyte activation (GO:0045321)	4.47E-04	8.52E-10
leukocyte activation involved in immune response (GO:0002366)	3.15E-04	7.26E-08
leukocyte degranulation (GO:0043299)	2.06E-04	1.07E-07
leukocyte differentiation (GO:0002521)	24.5	1.41
leukocyte mediated immunity (GO:0002443)	5.31E-04	2.39E-07
localization (GO:0051179)	2.43E-03	5.26E-13
localization of cell (GO:0051674)	100	1.12
lung development (GO:0030324)	100	1.97
lymphocyte activation (GO:0046649)	71.2	2.45
M phase (GO:0000279)	11	0.226
macromolecule biosynthetic process (GO:0009059)	19.3	1.00E-06
macromolecule catabolic process (GO:0009057)	100	1.13E-07
macromolecule localization (GO:0033036)	0.181	1.82E-05
macromolecule metabolic process (GO:0043170)	2.19	3.37E-08
macrophage differentiation (GO:0030225)	34.6	2.14
maintenance of gastrointestinal epithelium (GO:0030277)	3.53	100

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
maintenance of location (GO:0051235)	3.45	0.194
maintenance of location in cell (GO:0051651)	6.13	0.869
maintenance of protein location (GO:0045185)	6.76	1.11
maintenance of protein location in cell (GO:0032507)	33.1	4.47
maturation of LSU-rRNA (GO:0000470)	100	0.591
maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) (GO:0000463)	100	2.41
membrane organization (GO:0061024)	0.626	2.37
metabolic process (GO:0008152)	2.81	7.07E-12
microtubule cytoskeleton organization involved in mitosis (GO:1902850)	1.51	29.5
microtubule-based process (GO:0007017)	3.13	64.8
mitochondrial transport (GO:0006839)	26.5	4.94
mitotic anaphase (GO:0000090)	8.24	2.42
mitotic cell cycle (GO:0000278)	0.0168	2.39
mitotic cell cycle phase (GO:0098763)	20.6	0.0594
mitotic cell cycle process (GO:1903047)	0.0386	1.97
mitotic M phase (GO:0000087)	10.9	0.225
mitotic prometaphase (GO:0000236)	3.63	1.7
mitotic spindle organization (GO:0007052)	0.81	35.6
modulation by symbiont of entry into host (GO:0052372)	100	1.1
modulation by symbiont of host defense response (GO:0052031)	63.6	0.485
molting cycle (GO:0042303)	100	2.02
molting cycle process (GO:0022404)	94.2	2.66
morphogenesis of a branching epithelium (GO:0061138)	100	2.7
morphogenesis of a branching structure (GO:0001763)	100	1.62
movement in host environment (GO:0052126)	28.3	0.725
mRNA catabolic process (GO:0006402)	100	3.27E-09
mRNA metabolic process (GO:0016071)	33.2	2.15E-06
mRNA splicing, via spliceosome (GO:0000398)	5.58	3.07
mRNA stabilization (GO:0048255)	4.29	36.4
mucus secretion (GO:0070254)	2.07	46.4
multicellular organism development (GO:0007275)	0.169	0.193
multicellular organismal process (GO:0032501)	9.58E-04	1.56
MyD88-independent toll-like receptor signaling pathway (GO:0002756)	100	4.42
myeloid cell activation involved in immune response (GO:0002275)	3.08E-04	2.57E-07
myeloid cell differentiation (GO:0030099)	4.53	0.209

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
myeloid leukocyte activation (GO:0002274)	1.21E-04	3.03E-09
myeloid leukocyte differentiation (GO:0002573)	3.78	1.44
myeloid leukocyte mediated immunity (GO:0002444)	2.75E-04	1.85E-07
ncRNA metabolic process (GO:0034660)	100	0.246
ncRNA processing (GO:0034470)	100	0.918
negative regulation of amyloid fibril formation (GO:1905907)	100	2.52
negative regulation of apoptotic process (GO:0043066)	3.03	8.61E-07
negative regulation of apoptotic signaling pathway (GO:2001234)	11.6	0.0739
negative regulation of biological process (GO:0048519)	1.54E-07	7.39E-21
negative regulation of biosynthetic process (GO:0009890)	4.08	7.62E-05
negative regulation of blood coagulation, extrinsic pathway (GO:2000264)	4.18	12.3
negative regulation of catabolic process (GO:0009895)	2.75	1.4
negative regulation of catalytic activity (GO:0043086)	40.5	0.0411
negative regulation of cell activation (GO:0050866)	100	0.386
negative regulation of cell adhesion (GO:0007162)	75.5	0.235
negative regulation of cell communication (GO:0010648)	28.5	1.81E-05
negative regulation of cell cycle (GO:0045786)	1.13	3.72
negative regulation of cell cycle phase transition (GO:1901988)	28.6	3.15
negative regulation of cell cycle process (GO:0010948)	3.11	6.96
negative regulation of cell death (GO:0060548)	0.114	7.60E-08
negative regulation of cell development (GO:0010721)	100	1.83
negative regulation of cell differentiation (GO:0045596)	24.5	0.0213
negative regulation of cell population proliferation (GO:0008285)	12.9	0.187
negative regulation of cell-cell adhesion (GO:0022408)	66.2	0.0527
negative regulation of cellular amide metabolic process (GO:0034249)	24.1	2.33
negative regulation of cellular biosynthetic process (GO:0031327)	5.54	8.52E-05
negative regulation of cellular catabolic process (GO:0031330)	0.836	4.69
negative regulation of cellular component organization (GO:0051129)	2.54E-04	1.78E-05
negative regulation of cellular macromolecule biosynthetic process (GO:2000113)	4.04	6.44E-05
negative regulation of cellular metabolic process (GO:0031324)	1.54E-03	6.55E-10
negative regulation of cellular process (GO:0048523)	8.50E-07	6.86E-14
negative regulation of cellular protein metabolic process (GO:0032269)	0.0796	1.34E-07
negative regulation of cellular response to transforming growth factor beta stimulus (GO:1903845)	100	3.75
negative regulation of chromatin organization (GO:1905268)	1.75	11.9

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
negative regulation of chromatin silencing (GO:0031936)	1.92	11.5
negative regulation of chromosome organization (GO:2001251)	9.92E-03	0.331
negative regulation of cytokine production (GO:0001818)	73.8	1.13
negative regulation of cytoskeleton organization (GO:0051494)	8.1	2.33
negative regulation of developmental process (GO:0051093)	0.0705	8.46E-03
negative regulation of endopeptidase activity (GO:0010951)	6.64	2.75
negative regulation of gene expression (GO:0010629)	0.108	2.52E-13
negative regulation of glycoprotein biosynthetic process (GO:0010561)	82.4	2.05
negative regulation of glycoprotein metabolic process (GO:1903019)	21.9	0.518
negative regulation of hemopoiesis (GO:1903707)	88.5	3.44
negative regulation of immune effector process (GO:0002698)	12.4	3.74
negative regulation of immune system process (GO:0002683)	24	1.82
negative regulation of intracellular signal transduction (GO:1902532)	50.8	0.975
negative regulation of kinase activity (GO:0033673)	100	1.28
negative regulation of leukocyte activation (GO:0002695)	100	0.394
negative regulation of leukocyte cell-cell adhesion (GO:1903038)	100	1.75
negative regulation of lymphocyte activation (GO:0051250)	100	0.766
negative regulation of lymphocyte differentiation (GO:0045620)	100	1.86
negative regulation of macromolecule biosynthetic process (GO:0010558)	3.31	3.15E-04
negative regulation of macromolecule metabolic process (GO:0010605)	4.03E-03	7.14E-18
negative regulation of metabolic process (GO:0009892)	1.97E-03	1.25E-17
negative regulation of mitotic cell cycle (GO:0045930)	14.7	2.03
negative regulation of mitotic cell cycle phase transition (GO:1901991)	24.8	1.91
negative regulation of mitotic nuclear division (GO:0045839)	2.7	8.73
negative regulation of molecular function (GO:0044092)	19.3	5.38E-04
negative regulation of multicellular organismal process (GO:0051241)	0.115	4.24E-04
negative regulation of nervous system development (GO:0051961)	100	2.27
negative regulation of neurogenesis (GO:0050768)	100	0.898
negative regulation of neuron apoptotic process (GO:0043524)	48.3	4.56
negative regulation of nitrogen compound metabolic process (GO:0051172)	0.0295	4.28E-10
negative regulation of nuclear division (GO:0051784)	4.55	15
negative regulation of nucleic acid-templated transcription (GO:1903507)	36.3	0.0699
negative regulation of nucleobase-containing compound metabolic process (GO:0045934)	6.68	0.0592
negative regulation of organelle organization (GO:0010639)	4.40E-05	9.68E-04

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
negative regulation of ossification (GO:0030279)	40.6	0.0509
negative regulation of peptidase activity (GO:0010466)	7.54	4.88
negative regulation of phosphate metabolic process (GO:0045936)	100	0.099
negative regulation of phosphorus metabolic process (GO:0010563)	100	0.101
negative regulation of phosphorylation (GO:0042326)	100	0.069
negative regulation of programmed cell death (GO:0043069)	3.24	6.31E-07
negative regulation of protein catabolic process (GO:0042177)	80.6	0.786
negative regulation of protein kinase activity (GO:0006469)	100	1.46
negative regulation of protein metabolic process (GO:0051248)	0.0781	4.62E-10
negative regulation of protein modification by small protein conjugation or removal (GO:1903321)	100	0.527
negative regulation of protein modification process (GO:0031400)	62	5.77E-05
negative regulation of protein phosphorylation (GO:0001933)	100	0.0285
negative regulation of protein ubiquitination (GO:0031397)	100	0.803
negative regulation of protein-containing complex assembly (GO:0031333)	17.4	0.353
negative regulation of proteolysis (GO:0045861)	2.03	0.493
negative regulation of response to stimulus (GO:0048585)	6.92	2.33E-07
negative regulation of RNA biosynthetic process (GO:1902679)	36.3	0.0713
negative regulation of RNA catabolic process (GO:1902369)	2.03	54.9
negative regulation of RNA metabolic process (GO:0051253)	5.16	0.0302
negative regulation of signal transduction (GO:0009968)	37.5	6.77E-06
negative regulation of signaling (GO:0023057)	27.2	1.87E-05
negative regulation of supramolecular fiber organization (GO:1902904)	20.8	0.193
negative regulation of T cell activation (GO:0050868)	100	0.976
negative regulation of T cell differentiation (GO:0045581)	100	3.25
negative regulation of transcription by RNA polymerase II (GO:0000122)	88.3	1.71
negative regulation of transcription, DNA-templated (GO:0045892)	28.5	0.0345
negative regulation of transferase activity (GO:0051348)	100	0.927
negative regulation of transforming growth factor beta receptor signaling pathway (GO:0030512)	100	3.34
negative regulation of ubiquitin-dependent protein catabolic process (GO:2000059)	100	4.78
nervous system process (GO:0050877)	62.5	2.98
neuropilin signaling pathway (GO:0038189)	100	1.3
neutrophil activation (GO:0042119)	1.59E-04	1.98E-08
neutrophil activation involved in immune response (GO:0002283)	1.28E-04	3.76E-08

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
neutrophil degranulation (GO:0043312)	1.20E-04	2.96E-08
neutrophil mediated immunity (GO:0002446)	1.52E-04	5.65E-08
nitrogen compound metabolic process (GO:0006807)	19	1.12E-09
nitrogen compound transport (GO:0071705)	1.24	8.93E-03
nuclear-transcribed mRNA catabolic process (GO:0000956)	100	4.58E-10
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay (GO:0000184)	100	8.84E-11
nucleic acid metabolic process (GO:0090304)	2.01	5.17E-09
nucleic acid transport (GO:0050657)	0.145	23.1
nucleobase-containing compound biosynthetic process (GO:0034654)	52.5	0.0327
nucleobase-containing compound catabolic process (GO:0034655)	95.8	5.65E-10
nucleobase-containing compound metabolic process (GO:0006139)	2.54	1.50E-09
nucleobase-containing compound transport (GO:0015931)	0.225	37.4
nucleosome assembly (GO:0006334)	2.92	0.0219
nucleosome organization (GO:0034728)	2.51	2.18E-03
nucleosome positioning (GO:0016584)	2.18	12.8
odontogenesis of dentin-containing tooth (GO:0042475)	41.9	3.34
oogenesis (GO:0048477)	4.13	24.8
organelle localization (GO:0051640)	1.71	9.31
organelle organization (GO:0006996)	4.35E-06	1.54E-05
organic cyclic compound biosynthetic process (GO:1901362)	80.1	0.104
organic cyclic compound catabolic process (GO:1901361)	100	1.45E-07
organic cyclic compound metabolic process (GO:1901360)	15.5	2.36E-09
organic substance biosynthetic process (GO:1901576)	65.8	4.72E-07
organic substance catabolic process (GO:1901575)	100	1.55E-04
organic substance metabolic process (GO:0071704)	5.61	4.99E-10
organic substance transport (GO:0071702)	1.4	0.0591
organonitrogen compound biosynthetic process (GO:1901566)	100	1.12E-06
organonitrogen compound metabolic process (GO:1901564)	84.9	8.47E-06
ossification (GO:0001503)	2.91	0.127
osteoblast differentiation (GO:0001649)	4.14	4.27
oxidation-reduction process (GO:0055114)	100	2.58
peptide biosynthetic process (GO:0043043)	40.7	1.61E-10
peptide metabolic process (GO:0006518)	94.8	7.38E-11
peptide transport (GO:0015833)	25.6	1.69E-03
PERK-mediated unfolded protein response (GO:0036499)	14.2	1.4

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
platelet degranulation (GO:0002576)	4.32	2.93E-07
positive regulation by symbiont of entry into host (GO:0075294)	100	1.11
positive regulation of actin filament bundle assembly (GO:0032233)	30.4	3.71
positive regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains (GO:0002824)	7.71	4.15
positive regulation of angiogenesis (GO:0045766)	100	0.0753
positive regulation of animal organ morphogenesis (GO:0110110)	100	0.567
positive regulation of apoptotic process (GO:0043065)	1.37	6.10E-11
positive regulation of apoptotic signaling pathway (GO:2001235)	100	0.997
positive regulation of autophagy of mitochondrion (GO:1903599)	3.92	54.4
positive regulation of binding (GO:0051099)	1.24	4.40E-05
positive regulation of biological process (GO:0048518)	1.43E-05	3.26E-13
positive regulation of biosynthetic process (GO:0009891)	1.37E-03	9.42E-05
positive regulation of blood vessel endothelial cell migration (GO:0043536)	71	2.14
positive regulation of cardiac muscle cell differentiation (GO:2000727)	100	4.34
positive regulation of catalytic activity (GO:0043085)	1.89	1.84E-06
positive regulation of cell activation (GO:0050867)	13.4	3.91
positive regulation of cell adhesion (GO:0045785)	5.71	9.59E-03
positive regulation of cell communication (GO:0010647)	1.23	3.25E-05
positive regulation of cell cycle (GO:0045787)	8.49E-03	0.468
positive regulation of cell cycle process (GO:0090068)	0.145	0.402
positive regulation of cell death (GO:0010942)	0.184	1.66E-11
positive regulation of cell differentiation (GO:0045597)	5.58	0.0668
positive regulation of cell division (GO:0051781)	100	1.62
positive regulation of cell growth (GO:0030307)	10.5	1.62
positive regulation of cell migration (GO:0030335)	100	0.0247
positive regulation of cell morphogenesis involved in differentiation (GO:0010770)	51.9	2.6
positive regulation of cell motility (GO:2000147)	100	0.0702
positive regulation of cell population proliferation (GO:0008284)	0.0133	1.56E-05
positive regulation of cell size (GO:0045793)	12.6	1.12
positive regulation of cell-cell adhesion (GO:0022409)	43.4	3.57
positive regulation of cell-matrix adhesion (GO:0001954)	100	2.15
positive regulation of cell-substrate adhesion (GO:0010811)	34.7	0.519
positive regulation of cell-substrate junction organization (GO:0150117)	100	2.98

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
positive regulation of cellular amide metabolic process (GO:0034250)	0.991	0.0719
positive regulation of cellular biosynthetic process (GO:0031328)	0.0405	5.64E-04
positive regulation of cellular component biogenesis (GO:0044089)	1.75	15.5
positive regulation of cellular component movement (GO:0051272)	100	0.0528
positive regulation of cellular component organization (GO:0051130)	8.35E-04	0.0139
positive regulation of cellular metabolic process (GO:0031325)	2.71E-05	9.37E-09
positive regulation of cellular process (GO:0048522)	6.45E-06	1.92E-14
positive regulation of cellular protein localization (GO:1903829)	0.036	0.414
positive regulation of cellular protein metabolic process (GO:0032270)	0.03	4.99E-10
positive regulation of chromosome organization (GO:2001252)	0.0217	0.228
positive regulation of collagen biosynthetic process (GO:0032967)	6.42	0.375
positive regulation of collagen metabolic process (GO:0010714)	6.73	0.44
positive regulation of cytokine production (GO:0001819)	0.0464	0.803
positive regulation of defense response (GO:0031349)	28.3	1.17
positive regulation of developmental process (GO:0051094)	3.63	0.1
positive regulation of DNA binding (GO:0043388)	73.8	0.719
positive regulation of DNA biosynthetic process (GO:2000573)	6.37E-03	5.34
positive regulation of DNA metabolic process (GO:0051054)	6.58E-03	0.386
positive regulation of DNA replication (GO:0045740)	2.71	8.75
positive regulation of DNA-binding transcription factor activity (GO:0051091)	8.83	1.11
positive regulation of endothelial cell migration (GO:0010595)	100	0.12
positive regulation of endothelial cell proliferation (GO:0001938)	51	0.761
positive regulation of epithelial cell migration (GO:0010634)	100	4.19E-03
positive regulation of epithelial cell proliferation (GO:0050679)	39.2	0.0988
positive regulation of epithelial to mesenchymal transition (GO:0010718)	58.8	0.989
positive regulation of ERK1 and ERK2 cascade (GO:0070374)	21.5	2.98
positive regulation of establishment of protein localization (GO:1904951)	0.171	0.187
positive regulation of establishment of protein localization to telomere (GO:1904851)	1.03	31.4
positive regulation of fibroblast proliferation (GO:0048146)	0.846	0.358
positive regulation of focal adhesion assembly (GO:0051894)	100	1.91
positive regulation of gene expression (GO:0010628)	3.88E-03	8.52E-05
positive regulation of gene expression, epigenetic (GO:0045815)	0.194	2.83
positive regulation of hemopoiesis (GO:1903708)	75.2	0.126
positive regulation of hydrolase activity (GO:0051345)	11.1	1.22

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
positive regulation of I-kappaB kinase/NF-kappaB signaling (GO:0043123)	35.8	3.47
positive regulation of immune effector process (GO:0002699)	3.72	12.3
positive regulation of immune response (GO:0050778)	9.93	1.11
positive regulation of immune system process (GO:0002684)	10.8	0.0634
positive regulation of inflammatory response (GO:0050729)	46.9	1.53
positive regulation of integrin-mediated signaling pathway (GO:2001046)	60.3	3.28
positive regulation of interleukin-6 production (GO:0032755)	5.57	2.27
positive regulation of intracellular protein transport (GO:0090316)	14.7	3.04
positive regulation of intracellular signal transduction (GO:1902533)	13.5	5.02E-04
positive regulation of intracellular transport (GO:0032388)	4.07	2.46
positive regulation of kinase activity (GO:0033674)	62.9	7.45E-03
positive regulation of leukocyte cell-cell adhesion (GO:1903039)	28.7	3.15
positive regulation of leukocyte differentiation (GO:1902107)	99.3	0.978
positive regulation of leukocyte mediated immunity (GO:0002705)	5.53	2.7
positive regulation of leukocyte proliferation (GO:0070665)	21.7	4.39
positive regulation of locomotion (GO:0040017)	100	0.0891
positive regulation of lymphocyte differentiation (GO:0045621)	100	4.16
positive regulation of lymphocyte mediated immunity (GO:0002708)	7.7	4.15
positive regulation of macromolecule biosynthetic process (GO:0010557)	0.0137	6.16E-03
positive regulation of macromolecule metabolic process (GO:0010604)	1.21E-04	7.39E-09
positive regulation of MAP kinase activity (GO:0043406)	100	1.91
positive regulation of MAPK cascade (GO:0043410)	95.7	9.55E-03
positive regulation of mesenchymal cell proliferation (GO:0002053)	100	1.91
positive regulation of metabolic process (GO:0009893)	1.82E-06	1.03E-08
positive regulation of mitotic cell cycle (GO:0045931)	0.138	2.41
positive regulation of mitotic nuclear division (GO:0045840)	6.74	0.716
positive regulation of molecular function (GO:0044093)	0.0529	1.77E-08
positive regulation of morphogenesis of an epithelium (GO:1905332)	100	1.44
positive regulation of multicellular organismal process (GO:0051240)	3.21	0.298
positive regulation of myeloid cell differentiation (GO:0045639)	55.9	0.301
positive regulation of neural precursor cell proliferation (GO:2000179)	5.55	1.86
positive regulation of neuroblast proliferation (GO:0002052)	0.576	6.96
positive regulation of neuron apoptotic process (GO:0043525)	100	0.652
positive regulation of neuron death (GO:1901216)	55.8	0.0209

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
positive regulation of NF-kappaB transcription factor activity (GO:0051092)	27.1	3.15
positive regulation of nitric oxide biosynthetic process (GO:0045429)	55.4	2.75
positive regulation of nitric oxide metabolic process (GO:1904407)	55.2	2.99
positive regulation of nitrogen compound metabolic process (GO:0051173)	5.52E-03	7.63E-08
positive regulation of nuclear division (GO:0051785)	11	2.09
positive regulation of nucleic acid-templated transcription (GO:1903508)	9.6	2.07
positive regulation of nucleobase-containing compound metabolic process (GO:0045935)	0.26	0.591
positive regulation of nucleocytoplasmic transport (GO:0046824)	0.359	1.62
positive regulation of organelle organization (GO:0010638)	5.50E-03	6.69E-04
positive regulation of ossification (GO:0045778)	100	1.81
positive regulation of osteoblast differentiation (GO:0045669)	100	2.84
positive regulation of phosphate metabolic process (GO:0045937)	0.445	1.41E-03
positive regulation of phosphorus metabolic process (GO:0010562)	0.447	1.41E-03
positive regulation of phosphorylation (GO:0042327)	0.3	2.58E-03
positive regulation of programmed cell death (GO:0043068)	1.48	8.28E-11
positive regulation of prostaglandin biosynthetic process (GO:0031394)	100	3.28
positive regulation of protein import (GO:1904591)	3.67	31.8
positive regulation of protein import into nucleus (GO:0042307)	3.08	28
positive regulation of protein kinase activity (GO:0045860)	38	2.63E-03
positive regulation of protein kinase B signaling (GO:0051897)	11	4.1
positive regulation of protein localization to Cajal body (GO:1904871)	1.22	35.1
positive regulation of protein localization to chromosome, telomeric region (GO:1904816)	1.67	38.8
positive regulation of protein localization to nucleus (GO:1900182)	2.28E-03	3.97
positive regulation of protein metabolic process (GO:0051247)	0.0327	6.58E-10
positive regulation of protein modification process (GO:0031401)	3.31	3.13E-04
positive regulation of protein phosphorylation (GO:0001934)	1.8	2.13E-04
positive regulation of protein serine/threonine kinase activity (GO:0071902)	100	0.724
positive regulation of protein transport (GO:0051222)	3.3	0.467
positive regulation of reactive oxygen species biosynthetic process (GO:1903428)	24.9	2
positive regulation of reactive oxygen species metabolic process (GO:2000379)	57.3	0.372
positive regulation of receptor clustering (GO:1903911)	59.5	3.27
positive regulation of response to external stimulus (GO:0032103)	14.7	0.41

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
positive regulation of response to stimulus (GO:0048584)	0.494	2.27E-06
positive regulation of RNA biosynthetic process (GO:1902680)	9.63	2.08
positive regulation of RNA metabolic process (GO:0051254)	6.26	2.59
positive regulation of signal transduction (GO:0009967)	6.75	1.54E-05
positive regulation of signaling (GO:0023056)	1.25	3.52E-05
positive regulation of smooth muscle cell proliferation (GO:0048661)	48.8	0.154
positive regulation of stem cell proliferation (GO:2000648)	0.436	10
positive regulation of substrate adhesion-dependent cell spreading (GO:1900026)	50.7	0.0754
positive regulation of T cell activation (GO:0050870)	23.7	1.72
positive regulation of T cell mediated immunity (GO:0002711)	4.54	4.78
positive regulation of tau-protein kinase activity (GO:1902949)	7.49	0.346
positive regulation of telomerase activity (GO:0051973)	1.88	19.5
positive regulation of telomerase RNA localization to Cajal body (GO:1904874)	2.19	45.1
positive regulation of telomere maintenance (GO:0032206)	0.736	15
positive regulation of telomere maintenance via telomerase (GO:0032212)	1.63	44.6
positive regulation of telomere maintenance via telomere lengthening (GO:1904358)	2.04	46.7
positive regulation of transcription by RNA polymerase II (GO:0045944)	21.6	4.4
positive regulation of transcription, DNA-templated (GO:0045893)	5.9	3.26
positive regulation of transferase activity (GO:0051347)	33.1	6.77E-04
positive regulation of translation (GO:0045727)	0.33	0.0333
positive regulation of transport (GO:0051050)	0.137	4.70E-04
positive regulation of type I interferon production (GO:0032481)	0.66	21.2
positive regulation of vasculature development (GO:1904018)	100	0.218
positive regulation of viral entry into host cell (GO:0046598)	100	1.11
positive regulation of viral life cycle (GO:1903902)	28.8	0.218
positive regulation of viral process (GO:0048524)	27.7	0.162
posttranscriptional regulation of gene expression (GO:0010608)	0.109	3.02E-04
post-translational protein modification (GO:0043687)	46.8	0.0596
primary metabolic process (GO:0044238)	8.62	2.11E-08
programmed cell death (GO:0012501)	1.23	2.29E-03
protein folding (GO:0006457)	1.23	9.75E-06
protein insertion into mitochondrial outer membrane (GO:0045040)	4.19	59.9
protein localization (GO:0008104)	0.157	1.59E-06

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
protein localization to endoplasmic reticulum (GO:0070972)	100	1.51E-09
protein localization to membrane (GO:0072657)	100	6.93E-04
protein localization to organelle (GO:0033365)	7.13	3.15E-06
protein metabolic process (GO:0019538)	20.9	2.10E-04
protein refolding (GO:0042026)	3.52	4.33
protein stabilization (GO:0050821)	5.57	6.42E-03
protein targeting (GO:0006605)	92.7	4.24E-08
protein targeting to ER (GO:0045047)	100	1.09E-10
protein targeting to membrane (GO:0006612)	100	8.61E-07
protein transport (GO:0015031)	18.9	1.98E-03
protein-containing complex assembly (GO:0065003)	3.50E-04	1.43E-06
protein-containing complex disassembly (GO:0032984)	89.9	2.4
protein-containing complex subunit organization (GO:0043933)	6.19E-04	2.78E-08
protein-DNA complex assembly (GO:0065004)	0.682	1.31E-03
protein-DNA complex subunit organization (GO:0071824)	0.55	4.68E-04
purinergic nucleotide receptor signaling pathway (GO:0035590)	100	3.79
regeneration (GO:0031099)	55.6	3.15
regulated exocytosis (GO:0045055)	4.32E-05	1.76E-12
regulation of actin cytoskeleton organization (GO:0032956)	12.4	0.0666
regulation of actin cytoskeleton reorganization (GO:2000249)	100	0.0989
regulation of actin filament bundle assembly (GO:0032231)	57.2	3.56
regulation of actin filament organization (GO:0110053)	23.6	0.557
regulation of actin filament-based process (GO:0032970)	4.15	0.0572
regulation of adaptive immune response (GO:0002819)	27.4	3.65
regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains (GO:0002822)	22.9	4.93
regulation of anatomical structure morphogenesis (GO:0022603)	72.1	0.815
regulation of angiogenesis (GO:0045765)	100	0.261
regulation of apoptotic process (GO:0042981)	0.216	1.75E-16
regulation of apoptotic signaling pathway (GO:2001233)	4.22	5.44E-05
regulation of ATPase activity (GO:0043462)	3.34	42
regulation of binding (GO:0051098)	2.53	7.53E-04
regulation of biological process (GO:0050789)	3.44E-05	6.62E-14
regulation of biological quality (GO:0065008)	5.08E-03	1.75E-08
regulation of biosynthetic process (GO:0009889)	0.0346	5.22E-06

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
regulation of blood coagulation, extrinsic pathway (GO:2000263)	4.17	12.3
regulation of body fluid levels (GO:0050878)	2.7	0.589
regulation of bone mineralization (GO:0030500)	100	2.52
regulation of branching involved in lung morphogenesis (GO:0061046)	54.5	2.52
regulation of catabolic process (GO:0009894)	3.92	0.038
regulation of catalytic activity (GO:0050790)	0.269	3.25E-09
regulation of cell activation (GO:0050865)	7.16	0.0755
regulation of cell adhesion (GO:0030155)	14.3	2.85E-04
regulation of cell aging (GO:0090342)	0.739	38.1
regulation of cell communication (GO:0010646)	1.74	1.57E-08
regulation of cell cycle (GO:0051726)	3.08E-05	7.54E-04
regulation of cell cycle G2/M phase transition (GO:1902749)	23.6	1.66
regulation of cell cycle phase transition (GO:1901987)	0.459	0.287
regulation of cell cycle process (GO:0010564)	2.11E-05	6.64E-03
regulation of cell death (GO:0010941)	6.80E-03	3.04E-17
regulation of cell differentiation (GO:0045595)	2.79	4.87E-03
regulation of cell division (GO:0051302)	28.2	0.215
regulation of cell killing (GO:0031341)	1.71	3.37
regulation of cell migration (GO:0030334)	58.4	2.62E-03
regulation of cell motility (GO:2000145)	51.9	1.41E-03
regulation of cell population proliferation (GO:0042127)	4.48E-04	9.34E-09
regulation of cell-cell adhesion (GO:0022407)	12.2	9.34E-03
regulation of cell-matrix adhesion (GO:0001952)	100	3.27
regulation of cell-substrate adhesion (GO:0010810)	82.9	1.61
regulation of cellular amide metabolic process (GO:0034248)	0.969	0.0308
regulation of cellular biosynthetic process (GO:0031326)	0.137	5.77E-06
regulation of cellular catabolic process (GO:0031329)	1.56	0.0333
regulation of cellular component biogenesis (GO:0044087)	0.071	2.07
regulation of cellular component movement (GO:0051270)	55.3	2.19E-03
regulation of cellular component organization (GO:0051128)	2.04E-07	3.30E-07
regulation of cellular localization (GO:0060341)	6.59E-03	0.0219
regulation of cellular macromolecule biosynthetic process (GO:2000112)	0.0382	3.04E-05
regulation of cellular metabolic process (GO:0031323)	2.66E-05	1.11E-10
regulation of cellular process (GO:0050794)	7.82E-07	8.25E-10
regulation of cellular protein catabolic process (GO:1903362)	100	0.666

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
regulation of cellular protein localization (GO:1903827)	3.09E-04	5.26E-04
regulation of cellular protein metabolic process (GO:0032268)	3.08E-04	2.18E-13
regulation of cellular response to growth factor stimulus (GO:0090287)	81.4	4.61
regulation of cellular response to stress (GO:0080135)	0.325	0.195
regulation of cellular response to transforming growth factor beta stimulus (GO:1903844)	100	1.21
regulation of cellular senescence (GO:2000772)	3.08	28
regulation of chromatin organization (GO:1902275)	0.879	7.23
regulation of chromatin silencing (GO:0031935)	3.54	19.1
regulation of chromosome organization (GO:0033044)	9.96E-04	0.53
regulation of chromosome segregation (GO:0051983)	0.0112	4.53
regulation of collagen biosynthetic process (GO:0032965)	11.8	1.62
regulation of collagen metabolic process (GO:0010712)	2.88	2.52
regulation of complement-dependent cytotoxicity (GO:1903659)	8.6	4.14
regulation of cyclin-dependent protein kinase activity (GO:1904029)	7.52	3.94
regulation of cysteine-type endopeptidase activity (GO:2000116)	14.2	1.97
regulation of cysteine-type endopeptidase activity involved in apoptotic process (GO:0043281)	9.5	1.97
regulation of cytokine production (GO:0001817)	0.0296	7.53E-05
regulation of cytoskeleton organization (GO:0051493)	0.0375	0.0498
regulation of defense response (GO:0031347)	11.3	3.05
regulation of developmental process (GO:0050793)	0.0268	8.93E-04
regulation of DNA binding (GO:0051101)	14.1	0.194
regulation of DNA biosynthetic process (GO:2000278)	0.0152	13.6
regulation of DNA metabolic process (GO:0051052)	9.73E-05	0.7
regulation of DNA recombination (GO:0000018)	4.13	37.6
regulation of DNA replication (GO:0006275)	2.65	2.28
regulation of DNA-binding transcription factor activity (GO:0051090)	6.44	0.0144
regulation of DNA-templated transcription in response to stress (GO:0043620)	100	2.85
regulation of endopeptidase activity (GO:0052548)	0.806	0.143
regulation of endoribonuclease activity (GO:0060699)	4.15	12.2
regulation of endothelial cell differentiation (GO:0045601)	100	1.13
regulation of endothelial cell migration (GO:0010594)	54.8	0.484
regulation of endothelial cell proliferation (GO:0001936)	15.7	0.825
regulation of epithelial cell differentiation (GO:0030856)	100	4.76
regulation of epithelial cell migration (GO:0010632)	90.7	0.0593

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
regulation of epithelial cell proliferation (GO:0050678)	1.48	0.18
regulation of ERK1 and ERK2 cascade (GO:0070372)	49.1	0.16
regulation of establishment of protein localization (GO:0070201)	0.961	1.40E-04
regulation of establishment of protein localization to chromosome (GO:0070202)	1.44	38.6
regulation of establishment of protein localization to telomere (GO:0070203)	1.22	35.1
regulation of extrinsic apoptotic signaling pathway (GO:2001236)	100	2.81
regulation of fibroblast proliferation (GO:0048145)	0.0169	0.0856
regulation of G2/M transition of mitotic cell cycle (GO:0010389)	40.3	2.06
regulation of gene expression (GO:0010468)	0.0133	2.73E-13
regulation of gene expression, epigenetic (GO:0040029)	0.0879	5.73
regulation of gene silencing (GO:0060968)	2.55	13.5
regulation of generation of precursor metabolites and energy (GO:0043467)	0.807	78.9
regulation of glycoprotein biosynthetic process (GO:0010559)	15.8	2.75
regulation of glycoprotein metabolic process (GO:1903018)	4.13	1.1
regulation of growth (GO:0040008)	19.9	2.85
regulation of heart morphogenesis (GO:2000826)	100	2.53
regulation of hemopoiesis (GO:1903706)	19.3	2.99E-03
regulation of hydrolase activity (GO:0051336)	0.606	0.0178
regulation of immune effector process (GO:0002697)	4.33	3.91
regulation of immune response (GO:0050776)	0.719	0.0602
regulation of immune system process (GO:0002682)	1.82	6.32E-04
regulation of innate immune response (GO:0045088)	2.04	3.13
regulation of integrin-mediated signaling pathway (GO:2001044)	19.1	2.42
regulation of interleukin-6 production (GO:0032675)	1.42	0.976
regulation of intracellular protein transport (GO:0033157)	8.37	0.159
regulation of intracellular signal transduction (GO:1902531)	6.74	2.47E-06
regulation of intracellular transport (GO:0032386)	5.5	0.5
regulation of intrinsic apoptotic signaling pathway (GO:2001242)	9.31	0.436
regulation of kinase activity (GO:0043549)	28	4.76E-04
regulation of leukocyte activation (GO:0002694)	24	0.115
regulation of leukocyte cell-cell adhesion (GO:1903037)	34.5	0.179
regulation of leukocyte differentiation (GO:1902105)	46.5	0.129
regulation of leukocyte proliferation (GO:0070663)	13.8	0.0436
regulation of localization (GO:0032879)	5.06E-03	9.95E-09

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
regulation of locomotion (GO:0040012)	64.2	5.69E-03
regulation of lymphocyte activation (GO:0051249)	17.3	0.313
regulation of lymphocyte differentiation (GO:0045619)	100	2.53
regulation of lymphocyte proliferation (GO:0050670)	26	0.0485
regulation of macromolecule biosynthetic process (GO:0010556)	0.0316	8.55E-05
regulation of macromolecule metabolic process (GO:0060255)	3.20E-05	2.05E-17
regulation of MAP kinase activity (GO:0043405)	100	2.25
regulation of MAPK cascade (GO:0043408)	100	1.19E-03
regulation of mesenchymal cell proliferation (GO:0010464)	100	1.01
regulation of metabolic process (GO:0019222)	2.09E-05	1.25E-15
regulation of microtubule cytoskeleton organization (GO:0070507)	0.516	55.1
regulation of microtubule-based process (GO:0032886)	2.12	69.8
regulation of mitochondrion organization (GO:0010821)	16.9	1.78
regulation of mitotic cell cycle (GO:0007346)	1.41E-03	9.02E-03
regulation of mitotic cell cycle phase transition (GO:1901990)	0.7	0.227
regulation of mitotic nuclear division (GO:0007088)	0.0388	0.0601
regulation of mitotic sister chromatid segregation (GO:0033047)	2.59	6.3
regulation of mitotic spindle organization (GO:0060236)	0.2	46.2
regulation of molecular function (GO:0065009)	0.0232	4.42E-08
regulation of monocyte differentiation (GO:0045655)	100	4.93
regulation of mononuclear cell proliferation (GO:0032944)	26.5	0.0533
regulation of morphogenesis of an epithelium (GO:1905330)	100	2.63
regulation of mRNA catabolic process (GO:0061013)	0.838	0.0741
regulation of mRNA metabolic process (GO:1903311)	1.62	0.0141
regulation of mRNA stability (GO:0043488)	1.67	0.0219
regulation of multicellular organismal development (GO:2000026)	6.26	0.275
regulation of multicellular organismal process (GO:0051239)	0.0131	4.69E-04
regulation of muscle cell differentiation (GO:0051147)	99.2	2.51
regulation of myeloid cell differentiation (GO:0045637)	5.11	0.0983
regulation of neuroblast proliferation (GO:1902692)	1.64	4.88
regulation of neuron apoptotic process (GO:0043523)	23.5	0.246
regulation of neuron death (GO:1901214)	17.3	0.0816
regulation of nitric oxide biosynthetic process (GO:0045428)	74	2.84
regulation of nitrogen compound metabolic process (GO:0051171)	3.09E-04	2.17E-10
regulation of nuclear division (GO:0051783)	0.112	0.24
regulation of nucleic acid-templated transcription (GO:1903506)	9.28	0.171

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
regulation of nucleobase-containing compound metabolic process (GO:0019219)	0.226	0.0734
regulation of nucleocytoplasmic transport (GO:0046822)	0.127	0.297
regulation of organelle organization (GO:0033043)	8.43E-08	7.10E-07
regulation of ossification (GO:0030278)	40.8	0.0122
regulation of osteoblast differentiation (GO:0045667)	100	0.978
regulation of oxidative stress-induced cell death (GO:1903201)	9.25	4.79
regulation of peptidase activity (GO:0052547)	1.28	0.0736
regulation of peptide secretion (GO:0002791)	100	3.43
regulation of peptide transport (GO:0090087)	21.8	7.89E-04
regulation of phosphate metabolic process (GO:0019220)	0.27	1.73E-06
regulation of phosphorus metabolic process (GO:0051174)	0.27	1.76E-06
regulation of phosphorylation (GO:0042325)	0.182	3.91E-06
regulation of polysome binding (GO:1905696)	2.08	46.6
regulation of primary metabolic process (GO:0080090)	1.29E-04	6.57E-10
regulation of programmed cell death (GO:0043067)	0.0904	5.61E-17
regulation of protein binding (GO:0043393)	82.9	3.49
regulation of protein catabolic process (GO:0042176)	100	0.178
regulation of protein export from nucleus (GO:0046825)	2.89	2.53
regulation of protein kinase activity (GO:0045859)	28.1	2.02E-04
regulation of protein kinase B signaling (GO:0051896)	11.1	2.61
regulation of protein kinase C signaling (GO:0090036)	97.8	4.34
regulation of protein localization (GO:0032880)	1.36E-03	6.83E-06
regulation of protein localization to Cajal body (GO:1904869)	1.21	35.1
regulation of protein localization to chromosome, telomeric region (GO:1904814)	2.18	45.1
regulation of protein localization to nucleus (GO:1900180)	0.0462	2.19
regulation of protein metabolic process (GO:0051246)	6.51E-04	5.36E-14
regulation of protein modification by small protein conjugation or removal (GO:1903320)	30.8	0.111
regulation of protein modification process (GO:0031399)	5.57	3.61E-07
regulation of protein phosphorylation (GO:0001932)	1.29	2.81E-07
regulation of protein secretion (GO:0050708)	97.4	3.48
regulation of protein serine/threonine kinase activity (GO:0071900)	92.4	0.0527
regulation of protein stability (GO:0031647)	12.7	1.24E-03
regulation of protein transport (GO:0051223)	18.6	6.54E-04
regulation of protein ubiquitination (GO:0031396)	44.8	0.197

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
regulation of protein-containing complex assembly (GO:0043254)	7.56	0.756
regulation of proteolysis (GO:0030162)	1.74	3.29E-03
regulation of reactive oxygen species biosynthetic process (GO:1903426)	20.4	2.39
regulation of reactive oxygen species metabolic process (GO:2000377)	5.42	0.502
regulation of receptor clustering (GO:1903909)	79.8	1.72
regulation of receptor-mediated endocytosis (GO:0048259)	100	4.33
regulation of response to biotic stimulus (GO:0002831)	0.604	0.37
regulation of response to external stimulus (GO:0032101)	4.56	0.533
regulation of response to stimulus (GO:0048583)	0.0309	1.30E-08
regulation of response to stress (GO:0080134)	0.0151	2.22E-03
regulation of RNA biosynthetic process (GO:2001141)	9.32	0.21
regulation of RNA metabolic process (GO:0051252)	1.23	0.0364
regulation of RNA stability (GO:0043487)	2.08	0.0373
regulation of secretion (GO:0051046)	74.8	2.79
regulation of signal transduction (GO:0009966)	1.86	9.44E-09
regulation of signaling (GO:0023051)	1.88	1.85E-08
regulation of sister chromatid segregation (GO:0033045)	0.881	4.19
regulation of small molecule metabolic process (GO:0062012)	30.8	1.41
regulation of smooth muscle cell proliferation (GO:0048660)	17.2	0.104
regulation of spindle organization (GO:0090224)	0.3	50.7
regulation of stem cell proliferation (GO:0072091)	2.22	5.35
regulation of substrate adhesion-dependent cell spreading (GO:1900024)	67.5	0.44
regulation of supramolecular fiber organization (GO:1902903)	28.9	0.0436
regulation of symbiotic process (GO:0043903)	12.1	0.0813
regulation of T cell activation (GO:0050863)	20	0.123
regulation of T cell proliferation (GO:0042129)	28	0.204
regulation of tau-protein kinase activity (GO:1902947)	15.7	1.72
regulation of telomerase activity (GO:0051972)	4.56	38.1
regulation of telomerase RNA localization to Cajal body (GO:1904872)	3.17	51.2
regulation of telomere maintenance (GO:0032204)	0.112	9.64
regulation of telomere maintenance via telomerase (GO:0032210)	0.963	40.7
regulation of telomere maintenance via telomere lengthening (GO:1904356)	0.227	25.5
regulation of transcription by RNA polymerase II (GO:0006357)	11.9	0.286
regulation of transcription from RNA polymerase II promoter in response to hypoxia (GO:0061418)	100	2.82

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
regulation of transcription from RNA polymerase II promoter in response to stress (GO:0043618)	100	2.16
regulation of transcription, DNA-templated (GO:0006355)	7.04	0.0752
regulation of transferase activity (GO:0051338)	25.8	2.65E-05
regulation of transforming growth factor beta receptor signaling pathway (GO:0017015)	100	1.1
regulation of translation (GO:0006417)	0.313	0.116
regulation of translational initiation (GO:0006446)	3.64	9.66
regulation of transmembrane receptor protein serine/threonine kinase signaling pathway (GO:0090092)	100	1.11
regulation of transport (GO:0051049)	6.91	1.62E-04
regulation of type I interferon production (GO:0032479)	4.21	21.7
regulation of type I interferon-mediated signaling pathway (GO:0060338)	1.63	44.6
regulation of vascular endothelial growth factor production (GO:0010574)	100	4.42
regulation of vasculature development (GO:1901342)	86.1	0.156
regulation of vesicle-mediated transport (GO:0060627)	100	2.41
regulation of viral entry into host cell (GO:0046596)	100	2.33
regulation of viral genome replication (GO:0045069)	1.59	0.3
regulation of viral life cycle (GO:1903900)	8.09	0.0305
regulation of viral process (GO:0050792)	9.23	0.0368
reproduction (GO:0000003)	1.26	42.3
reproductive process (GO:0022414)	1.23	42.2
reproductive structure development (GO:0048608)	0.922	12.2
reproductive system development (GO:0061458)	0.987	12.3
respiratory tube development (GO:0030323)	100	2.29
response to abiotic stimulus (GO:0009628)	7.16E-03	3.28E-05
response to biotic stimulus (GO:0009607)	0.707	0.246
response to chemical (GO:0042221)	7.47E-05	1.67E-09
response to cocaine (GO:0042220)	4.81	39.2
response to cold (GO:0009409)	4.12	1.1
response to corticosteroid (GO:0031960)	0.179	1.3
response to cytokine (GO:0034097)	6.92E-05	4.33E-10
response to decreased oxygen levels (GO:0036293)	0.681	3.01E-03
response to defenses of other organism (GO:0052173)	73.9	1.11
response to drug (GO:0042493)	0.15	1.7
response to endogenous stimulus (GO:0009719)	0.0542	1.39E-05

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
response to endoplasmic reticulum stress (GO:0034976)	35	0.0827
response to estradiol (GO:0032355)	1.81	3.59
response to external biotic stimulus (GO:0043207)	0.602	0.218
response to external stimulus (GO:0009605)	14	1.6
response to gamma radiation (GO:0010332)	1.1	43.6
response to growth factor (GO:0070848)	36.3	4.46E-03
response to hormone (GO:0009725)	0.13	3.31E-03
response to host (GO:0075136)	73.5	1.11
response to host defenses (GO:0052200)	73.9	1.11
response to hydrogen peroxide (GO:0042542)	3.23	17
response to hypoxia (GO:0001666)	1.56	4.57E-03
response to inorganic substance (GO:0010035)	7.54E-03	2.07
response to interferon-gamma (GO:0034341)	14.7	0.481
response to interleukin-1 (GO:0070555)	19	0.0139
response to interleukin-7 (GO:0098760)	100	0.787
response to iron ion (GO:0010039)	1.51	100
response to lipid (GO:0033993)	0.449	0.0129
response to lipopolysaccharide (GO:0032496)	6.84	1.36
response to mechanical stimulus (GO:0009612)	47	3.85
response to metal ion (GO:0010038)	0.887	28.7
response to mineralocorticoid (GO:0051385)	1.4	15.1
response to molecule of bacterial origin (GO:0002237)	8.57	1.02
response to nitrogen compound (GO:1901698)	0.018	4.06E-03
response to organic cyclic compound (GO:0014070)	0.0512	0.057
response to organic substance (GO:0010033)	3.15E-04	5.00E-16
response to organonitrogen compound (GO:0010243)	0.0378	4.16E-03
response to other organism (GO:0051707)	0.604	0.215
response to oxidative stress (GO:0006979)	9.27E-03	0.0143
response to oxygen levels (GO:0070482)	1.1	9.37E-03
response to oxygen-containing compound (GO:1901700)	0.0239	1.36E-04
response to peptide (GO:1901652)	0.788	1.95E-03
response to peptide hormone (GO:0043434)	0.54	0.0531
response to reactive oxygen species (GO:0000302)	0.123	1.92
response to redox state (GO:0051775)	86	2.42
response to steroid hormone (GO:0048545)	1.47	0.298
response to stimulus (GO:0050896)	8.30E-04	8.55E-10

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
response to stress (GO:0006950)	4.35E-06	5.44E-12
response to temperature stimulus (GO:0009266)	3.44	1.56
response to topologically incorrect protein (GO:0035966)	1.63	1.12E-04
response to transforming growth factor beta (GO:0071559)	100	0.11
response to tumor necrosis factor (GO:0034612)	71.8	2.27
response to type I interferon (GO:0034340)	2.84	2.38
response to unfolded protein (GO:0006986)	0.881	3.06E-04
response to wounding (GO:0009611)	0.0234	8.56E-03
ribonucleoprotein complex assembly (GO:0022618)	17.2	1.98E-04
ribonucleoprotein complex biogenesis (GO:0022613)	8.59	5.12E-07
ribonucleoprotein complex subunit organization (GO:0071826)	19.4	8.53E-05
ribosomal large subunit assembly (GO:0000027)	100	0.115
ribosomal large subunit biogenesis (GO:0042273)	100	1.59E-03
ribosomal subunit export from nucleus (GO:0000054)	2.17	12.8
ribosome assembly (GO:0042255)	34.2	0.0194
ribosome biogenesis (GO:0042254)	63.9	3.03E-04
ribosome localization (GO:0033750)	2.2	12.9
RNA catabolic process (GO:0006401)	35.1	4.81E-11
RNA localization (GO:0006403)	0.314	38.7
RNA metabolic process (GO:0016070)	12.5	2.57E-05
RNA processing (GO:0006396)	73.2	0.0524
RNA splicing, via transesterification reactions (GO:0000375)	5.81	3.25
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile (GO:0000377)	5.56	3.07
RNA transport (GO:0050658)	0.146	23.1
rRNA metabolic process (GO:0016072)	100	2.11E-04
rRNA processing (GO:0006364)	100	3.59E-04
rRNA-containing ribonucleoprotein complex export from nucleus (GO:0071428)	2.52	14.3
salivary gland development (GO:0007431)	100	4.89
salivary gland morphogenesis (GO:0007435)	100	3.3
secretion (GO:0046903)	3.29E-04	1.84E-08
secretion by cell (GO:0032940)	1.20E-03	4.65E-10
sensory perception (GO:0007600)	50.2	0.137
sensory perception of chemical stimulus (GO:0007606)	100	0.119
sensory perception of smell (GO:0007608)	42.7	0.639
signal transduction (GO:0007165)	1.83	4.24E-03

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
signaling (GO:0023052)	3.07	0.0913
skin epidermis development (GO:0098773)	94.3	2.67
somatic cell DNA recombination (GO:0016444)	9.97	4.87
somatic diversification of immune receptors via germline recombination within a single locus (GO:0002562)	10	4.89
spindle organization (GO:0007051)	1.22	20.6
SRP-dependent cotranslational protein targeting to membrane (GO:0006614)	100	4.53E-10
stress granule assembly (GO:0034063)	3.93	21
supramolecular fiber organization (GO:0097435)	0.325	4.42
symbiotic process (GO:0044403)	1.23E-07	9.34E-18
system development (GO:0048731)	0.458	0.0977
T cell activation (GO:0042110)	60.1	2.03
T cell differentiation (GO:0030217)	89.7	3.59
telomerase holoenzyme complex assembly (GO:1905323)	4.16	12.2
tissue development (GO:0009888)	4.56	0.159
tissue homeostasis (GO:0001894)	1.23	78.6
tissue migration (GO:0090130)	100	2.78
tissue morphogenesis (GO:0048729)	100	3.34
trachea development (GO:0060438)	98.4	4.35
trachea formation (GO:0060440)	55.9	2.53
transforming growth factor beta receptor signaling pathway (GO:0007179)	100	0.0892
translation (GO:0006412)	34.2	2.71E-11
translational initiation (GO:0006413)	19.4	3.71E-13
transmembrane receptor protein serine/threonine kinase signaling pathway (GO:0007178)	100	0.319
transport (GO:0006810)	3.84E-03	1.09E-10
trans-synaptic signaling (GO:0099537)	100	4.89
TRIF-dependent toll-like receptor signaling pathway (GO:0035666)	100	3.29
tube development (GO:0035295)	100	0.37
tube morphogenesis (GO:0035239)	70.6	0.0468
type I interferon signaling pathway (GO:0060337)	2.22	1.73
Unclassified (UNCLASSIFIED)	7.25E-03	1.27E-12
vasculature development (GO:0001944)	52.1	0.0128
ventricular system development (GO:0021591)	42.2	4.43
vesicle-mediated transport (GO:0016192)	0.101	8.63E-07
viral entry into host cell (GO:0046718)	55.8	3.07

Table S2.2. GO enrichment analysis using DEGs from supervised analysis and unsupervised DEGs. A positive log (fold change) indicates higher expression in the network subpopulation, while a negative log (fold change) indicates higher expression in the spheroid subpopulation. Highlighted genes were identified exclusively by using phenotypic labels, Continued.

GO biological process complete	FDR % (Supervised Analysis)	FDR % (Unsupervised Analysis)
viral gene expression (GO:0019080)	100	6.04E-11
viral life cycle (GO:0019058)	0.287	0.0148
viral process (GO:0016032)	1.03E-07	1.13E-17
viral transcription (GO:0019083)	100	2.41E-10
viral translation (GO:0019081)	100	3.78
wound healing (GO:0042060)	0.11	0.0354

Chapter 3: Supervised bioinformatics enables guided detection of invasive potential within non-invasive phenotypes

3.1 Introduction

Collective migration is a process where multiple cells coordinate their movement(*17–20*). This contrasts with single cell migration, where cells move individually. Despite having long been observed in events such as morphogenesis and wound healing, only recently has attention been paid on this phenomena with respect to cancer(*16*). Studies have demonstrated the possibility for metastatic seeding of cells in cancer progression, and evidence suggests that this more efficient manner of transport is the dominant means by which metastatic lesions form(*21*, *34*, *48*).

Our lab generated a 3D cancer model capable of producing collectively migrating cancer phenotypes(*21*). In previous publications, we demonstrated that these phenotypes contain gene signatures that are prognostic of patient survival across many different cancer types(*21*). In my work, I more closely dissect the different phenotypes we observe. In the previous chapter, I demonstrated that phenotypes differ in invasive potential, which are associated with stress responses, immune processes, and proliferation. However, clinical implications of these heterogeneous phenotypes have not yet been evaluated. Furthermore, the transcriptomic landscape of our phenotypes displays further intricacies that remain unexplored, such as the intraheterogeneity present within the collectively non-invasive cells.

Here, we use further bioinformatics approaches to evaluate the gene signatures I extracted and attempt to add translational value to our platform. To establish clinical relevance, I utilize our gene signatures to cluster patient data on public databases to determine if patients with profiles associated with our collective phenotypes display biases in survival and disease progression. To

elucidate within-subtype-heterogeneity, we use cell cycle analysis, trajectory analysis, and gene pattern expression analysis to dissect our phenotypes on different biological and temporal scales to establish novel relationships. Our results provide preliminary evidence linking our phenotypes to histopathological subtypes, and computational results point to a 3-state spectrum consisting of a collectively invasive, collectively non-invasive, and an intermediate state. By possessing phenotypic metadata, we are able to comprehend bioinformatic analyses through additional lenses, which adds to the power of these approaches.

3.2 Results

3.2.1 Collective migration signatures bias towards specific histopathological subtypes

Since we previously demonstrated that the collective phenotype as a whole contained signatures that predicted patient outcome(*50*), I first evaluated whether the gene expression signatures that differentiated the collective phenotypes we observed could be prognostic. After collecting breast cancer patient transcriptomic data from TCGA, I filtered for the genes we obtained that were differentially regulated between the collectively invasive and non-invasive phenotypes, stratified the patient population according to the expression of these genes, and evaluated the survival of the top and bottom scoring patients. Since this metagene contained markers of cell migration, I hypothesized that patient survival would segregate based on a signature that contained elements of migratory potential. Surprisingly, I did not find a difference in overall survival, progression-free interval, or disease-specific survival (Figure 3.1). I then explored other clinical metadata that could be correlated with our gene signature. I found that patients with transcriptomes that corresponded more with the collectively invasive cells were more associated

with the ductal carcinoma (IDC) subtype, while patients with transcriptomes that corresponded more with the collectively non-invasive cells were more associated with the luminal carcinoma (ILC) subtype (Figure 3.2).

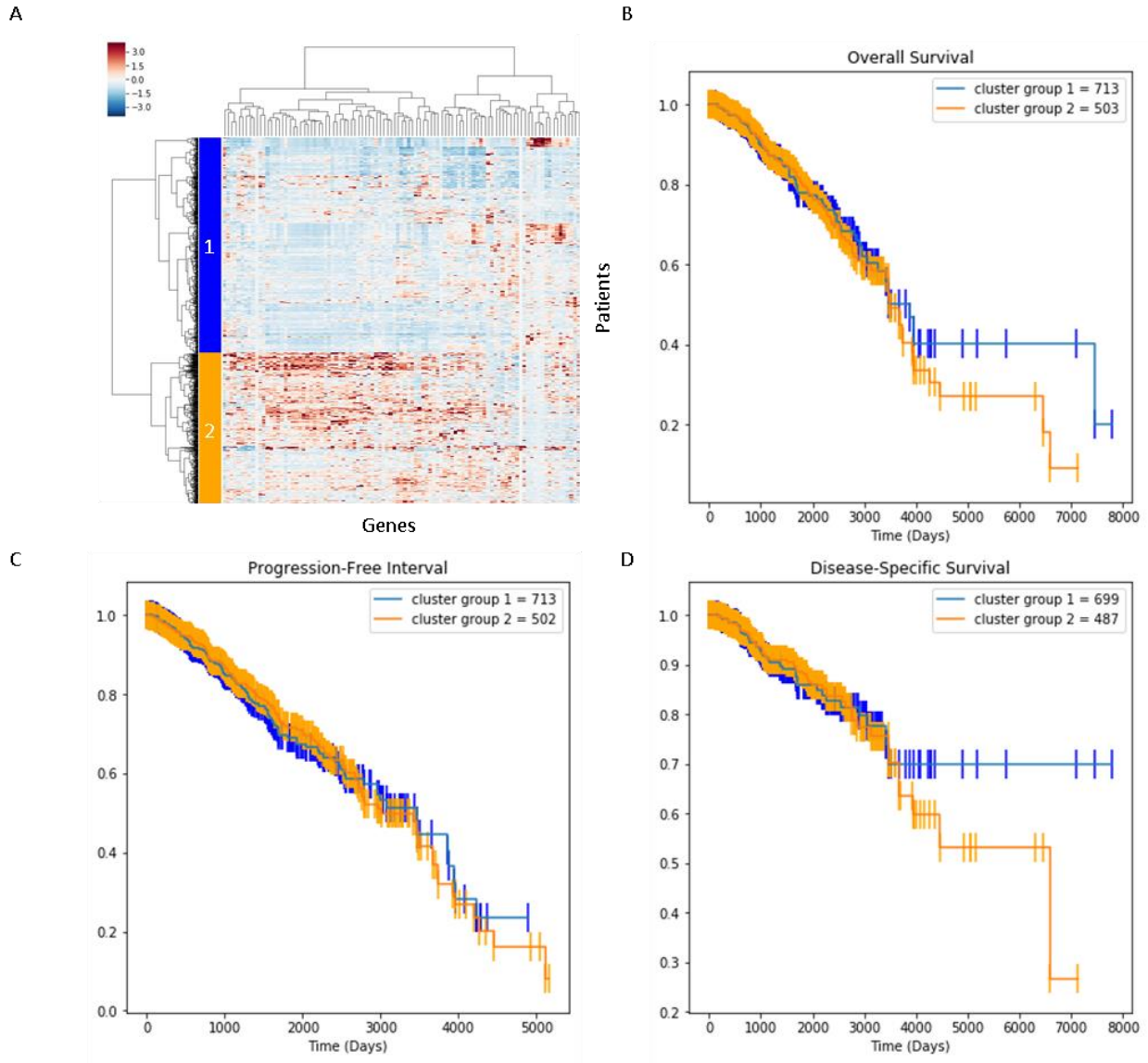


Figure 3.1. Signatures describing collective migration heterogeneity do not predict patient survival. (**A**) Clustermap of the 178 DEGs that describe collective migration heterogeneity applied to breast cancer patient data from TCGA after z-score normalization. Kaplan-Meier plots of patients clustered according to their gene expression according to (**B**) Overall Survival, (**C**) Progression-Free Interval, and **(D)** Disease-Specific Survival. Log-rank tests for all comparisons did not show statistically significant differences.

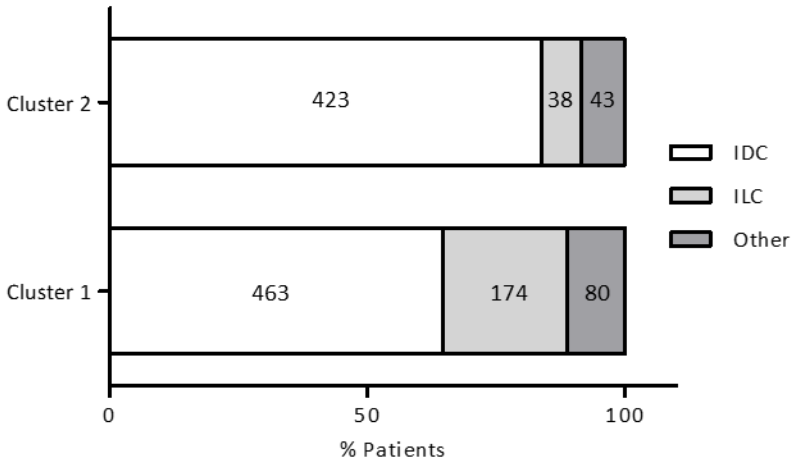


Figure 3.2. Histopathological subtypes of breast cancer patients according to cluster groups. Clinical metadata was used to identify the assigned subtype for each patient in each cluster identified by the clustermap. Patients in cluster 2, which were transcriptionally more similar to the collectively invasive cells, were biased more towards the IDC subtype while patients in cluster 1, which were transcriptionally more similar to the collectively non-invasive cells, were biased more towards the ILC subtype (Fisher's exact test: p-value: 2.49×10^{-16}). Numbers within the column represent the number of patients in each subtype within each cluster.

3.2.2 The invasive phenotype is associated with cancer stem cell markers

Since cells that were labeled as belonging to the spheroid phenotype fell into two distinct transcriptional clusters based on unsupervised analysis, I hypothesized that the spheroid cells may be phenotypically unstable. Clinically, plasticity is a factor that is detrimental to patient survival. To determine the phenotypic stability of MDAs that formed spheroids, we labeled, sorted, and re-embedded spheroid cells as single cells (Fig. 3.3A). This resulted in ~75% of the spheroid cells converting into network cells (Fig. 3.3C). Conversely, sorting and re-embedding invasive network cells as single cells resulted in near 100% invasive network formation (Fig. 3.3B, C). This suggests that spheroid cells retain the capacity to switch into a proliferative and invasive state, but network cells are less capable of reverting into a non-proliferative, non-invasive state. To assess whether one of these states represented a more cancer stem cell-like state, we immunostained for CD24 and CD44. Both phenotypes display a CD44+ profile (Figure 3.3C, D). However, CD24 staining was negative for the network cells (Figure 3.3C) and mixed for the spheroid cells (Figure 3.3D,

E). A positive control was done for CD24 on Nalm6 cells to confirm the absence of CD24 staining (Figure S3.1). When comparing the spheroid population in the smaller lone cluster to the ones in the larger cluster, we found a list of surface markers that could potentially be used to sort the two subpopulations in future experiments (Table S3.1).

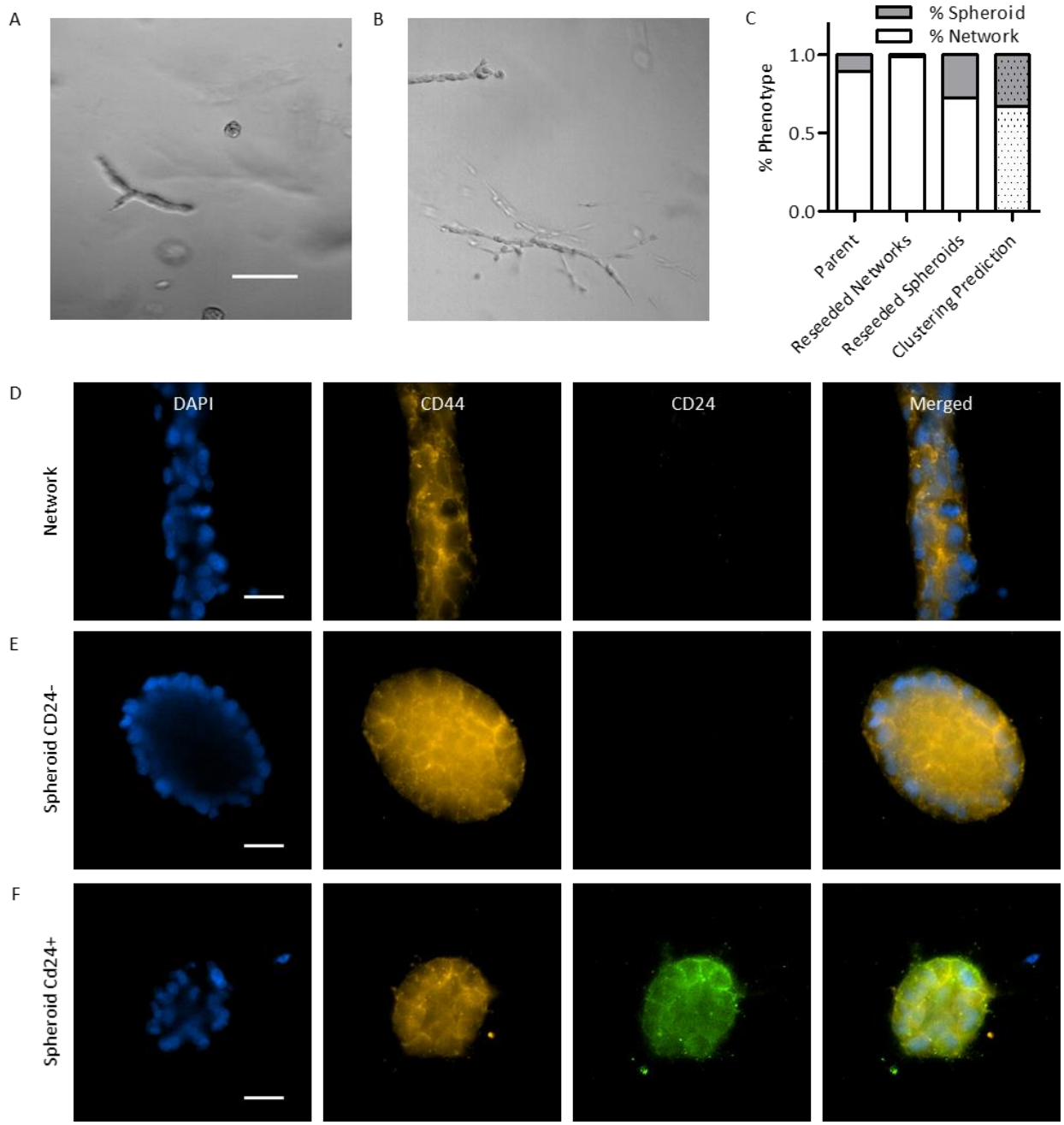


Figure 3.3. Phenotypic sorting enables analysis of phenotype stability. (**A**) Representative brightfield image of reseeded spheroid cells after 7 days of culture in 3D type I collagen. Scale bar 50 µm. (**B**) Representative brightfield image of reseeded network cells after 7 days of culture in 3D type I collagen. (**C**) Quantification of the phenotypes that arise after reseeding from sorted populations. Invasive network cells largely reform network structures, while non-invasive spheroids may either form network or spheroid structures. **(D)** CD44 immunostaining of network cells. **(E)** CD44 staining of spheroid cells that are CD24 negative. **(F)** CD44 staining of spheroid cells that are CD24 positive.

3.2.3 Phenotypically non-invasive cells can occupy a transitory state

As shown before, transcriptomic differences do not fully explain the phenotypic difference we observe *in vitro* (Figure 3.4A-B). To address the heterogeneity present within the non-invasive cells, we looked more closely at the expression of cell cycle markers (Figure 3.4C). In an unsupervised approach, cells in cluster 0 appear to be more actively proliferative (Figure 3.4D). However, applying phenotypic labels, we see that some non-invasive cells, namely those that occupy transcriptomic space near the invasive cells, display intermediate cell cycle activity (Figure 3.4E). Regressing the expression of these markers do not disturb the overall segregation of the cells into two large clusters, with separation between invasive cells (network), non-invasive cells (spheroid), and an intermediate state (spheroid in cluster 0) (Figure 3.4F-H).

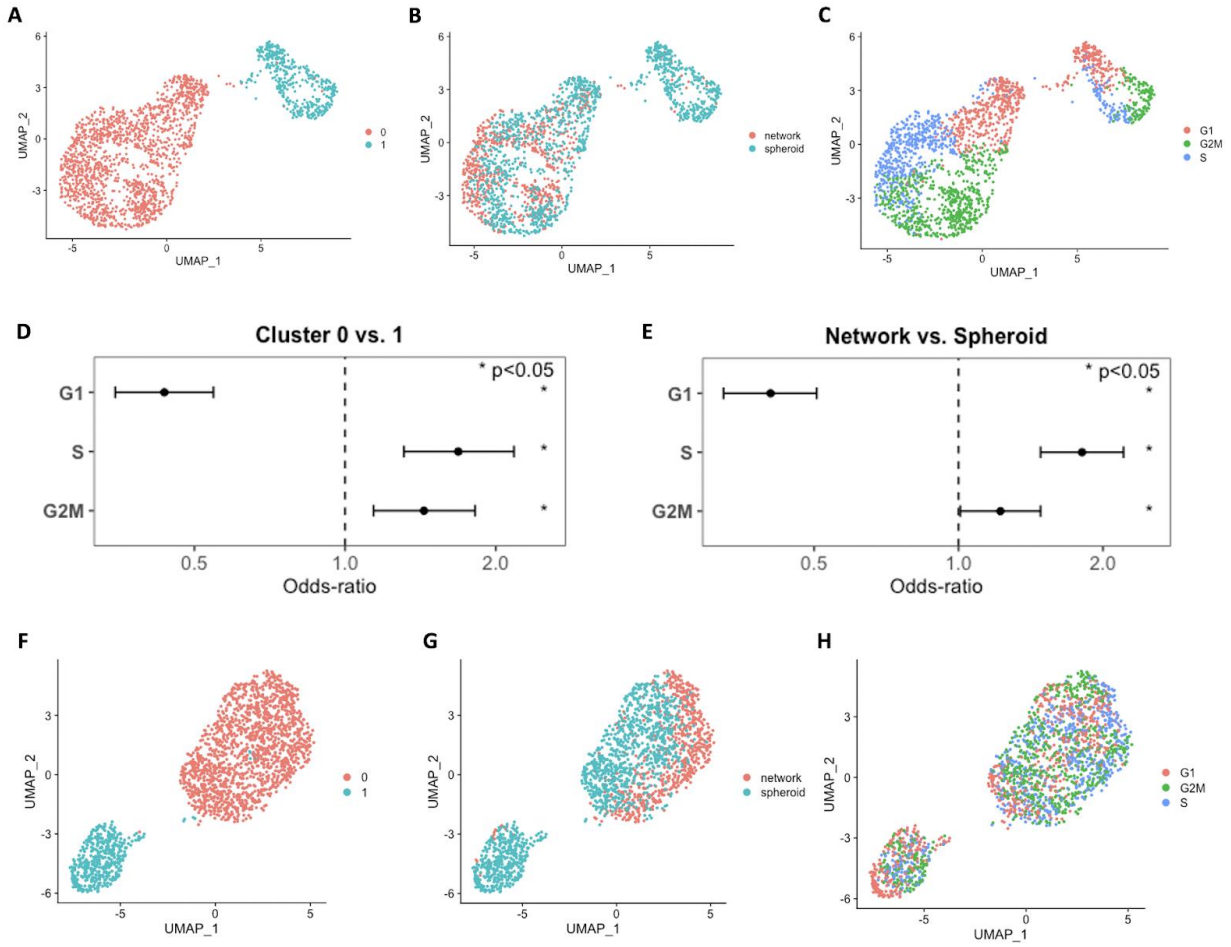


Figure 3.4. Comparative analysis of unsupervised clusters versus phenotypic labels. (A-C) UMAP plots colored by **(A)** unsupervised clusters, **(B)** phenotypic labels, and **(C)** cell cycle phases. **(D-E)** Cell cycle phase distribution in **(D)** (unsupervised clusters) or in **(E)** (network vs. spheroid phenotypically labelled cells). Odds ratios (OR) and 95% confidence intervals of each cell cycle phase (G1, S and G2M) to be prevalent in **(D)** (unsupervised clusters 0 vs. 1) or in **(E)** (network vs. spheroid labelled cells) using Fisher's exact test are shown (*p<0.05). For **(D)** OR>1 means cells in cluster 0 are more prevalent in the indicated cell cycle phase compared to cells in cluster 1, and for **(E)** OR>1 means network cells are more prevalent in the indicated cell cycle phase compared to spheroid cells. **(F-H)** UMAP plots colored by **(F)** unsupervised clusters, **(G)** phenotypic labels, and **(H)** cell cycle phases.

3.2.4 Pseudotime trajectory analysis highlights potential phenotype regulators

In development, cells transition from one state to another. In order to track changes of the cells over time, we analyzed them as a function of pseudotime, which is a measure of how far a cell has moved through biological progress(*51*). Pseudotime progresses along the trajectory of gene expression changes present in the underlying data starting from a root representing the beginning of the biological process. Choosing midpoint of cells in G1 cell cycle phase as the root (Figure 3.5A), we analyzed the cells along the pseudotime trajectory (Figure 3.5B), which displayed the

switching of the transitory spheroid cells to network cells (Figure 3.5C) as they progressed through the cell cycle phases, from G1, to S, to G2M (Figure 3.5A). For this analysis, we used the expression data without regressing out cell cycle effects as the pseudotime trajectory should reflect the natural biological progress cells go through including cell cycle progression. We also demonstrated the expression changes of 3 differentially expressed genes along the pseudotime trajectory: SPANXB1 which is highly expressed in spheroid cells in cluster 1 (Figure 3.5D), S100A4 which displays a higher expression in the transitory spheroid cells compared to the rest of the spheroid or the network cells (Figure 3.5E), and HIST1H1E which has a higher expression in network cells (Figure 3.5F), specifically compared to the transitory spheroid cells.

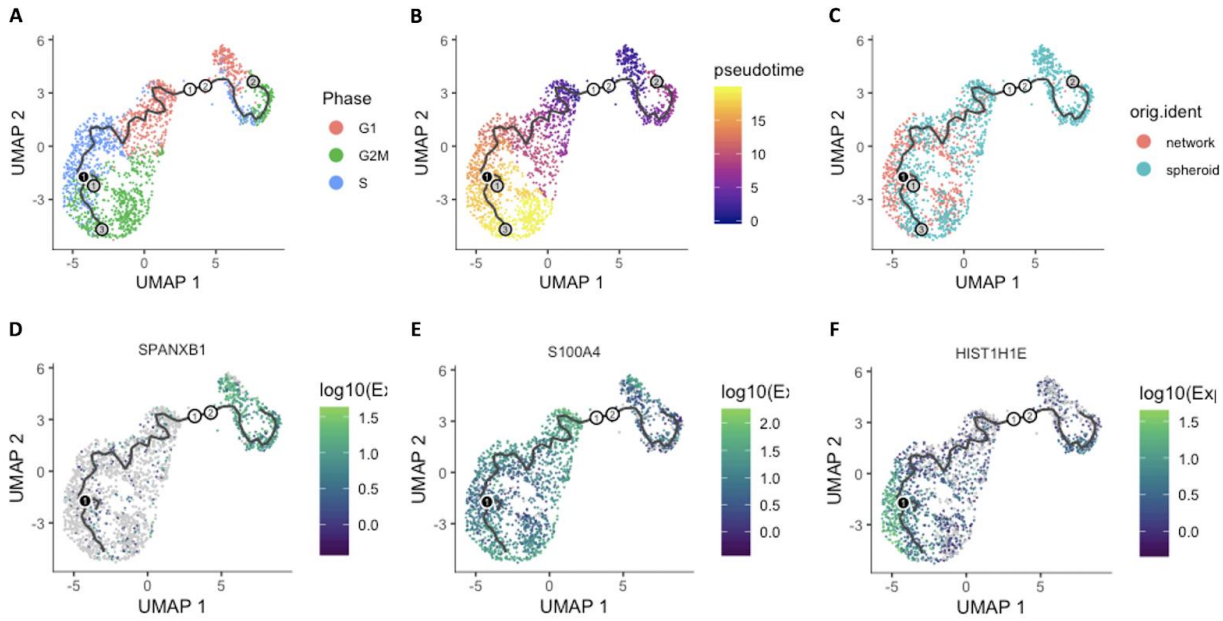


Figure 3.5. Pseudotime trajectory analysis. (A-C) UMAP plots with pseudotime trajectories of cells colored by **(A)** cell cycle phase, **(B)** pseudotime, and **(C)** phenotypically supervised cell groups. **(D-F)** UMAP plots with pseudotime trajectories of cells and the expression of genes **(D)** SPANXB1, **(E)** S100A4, and **(F)** HIST1H1E.

3.2.4 Gene expression pattern analysis identifies potential phenotypic markers

As unsupervised clustering of the expression data did not separate the transcriptionally similar transitory spheroid cells and network cells from each other (Figure 2.2, 3.4), we explored

other methods to see if this differentiation could be achieved through an unsupervised technique as opposed to the supervised phenotypic labeling. We applied a nonnegative matrix factorization algorithm on the expression data to discover the gene expression patterns underlying the functional phenotypes(*33*). Among seven patterns that were discovered, three patterns best explained the expression profiles of spheroid vs. intermediate vs. network phenotypes. Pattern 4 detected spheroid cells (Kruskal-Wallis rank-sum test p-value=9.34e-203) (Figure 3.6A), pattern 7 detected network cells (Kruskal-Wallis rank-sum test p-value=1.96e-141) (Figure 3.6B), and pattern 3 detected intermediate cells (Kruskal-Wallis rank-sum test p-value=2.433659e-115) (Figure 3.6C). A heatmap of pattern weights further demonstrated the enrichment of pattern 4 for spheroid cells, pattern 7 for network cells and pattern 3 for intermediate cells (Figure 3.6D).

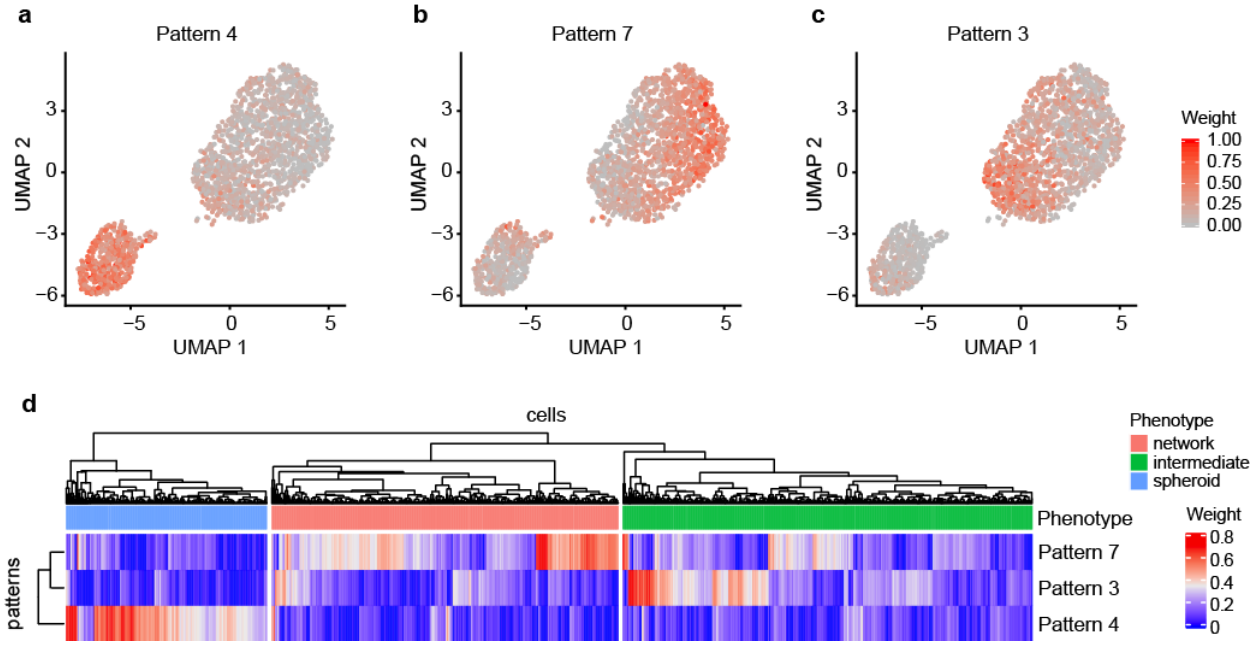


Figure 3.6. Unsupervised detection of gene expression patterns. (a-c) UMAP plots overlaid with pattern weights for each cell for **(a)** pattern 4, **(b)** pattern 7, and **(c)** pattern 3. Cell cycle phase effects are regressed out. **(d)** Heatmap of pattern weights by cells. Hierarchical clustering is performed between rows and between columns within and between phenotypes.

3.3 Discussion

Our evaluation of the clinical relevance of our gene set describing heterogeneity in collective migration suggests that the collective invasion phenotype is not associated with poorer clinical outcomes. Here, my analysis utilized data from TCGA, which only contained bulk sequencing transcriptomic data and may not possess the data resolution required to appropriately assess our single cell sequencing results, since population level signals may drown out single cell signals(*21*). While the preliminary analyses are presented here, future work will focus on comprehensively assessing the clinical relevance of our gene signature. Matching our gene signature to single cell patient data may provide more informative results, as we can better match patient profiles to our phenotypes and perhaps evaluate the percentage of cells associated with particular collective migration phenotypes. Given that the signatures corresponding to collective migration as a whole was prognostic(*53–56*), it is possible that a particular mixture of phenotypes may enhance the metastatic process, and is thus more detrimental to patient health. Though not prognostic, it is interesting to note that the transcriptomic signatures of each phenotype were biased towards particular histopathological subtypes. Our results correspond to that of another study which found that the top pathways that were significantly enriched in ILC were related to immune response(*57*). Further investigation is required to understand the implications of this bias, though perhaps this indicates potential exploitation of immunotherapy in patients with ILC.

I hypothesize that the reason our gene signature is not predictive of patient outcome may stem from the internal heterogeneity present within collectively non-invasive cells. As demonstrated through our omics data, phenotypically non-invasive cells may lie in a transitory state that carries invasive potential. Though we initially hypothesized that patient profiles corresponding to the invasive phenotype should have worse clinical outcomes, perhaps this plastic

phenotype is more detrimental. This phenotype is reminiscent of the phenomena of cancer dormancy, where tumor growth and progression is arrested but can reactivate(*53–56*) . Studies have shown that cancer cells can progress through phases of active carcinogenesis, quiescence, and resurgence. In our cancer model, these cells could potentially be represented by the non-invasive cells that can become invasive upon reseeding. These cells may also correspond to the non-invasive cells that are CD44+/CD24-, which is the profile of classic breast cancer stem cells(*57*). Further validation experiments involving these subgroups are warranted to confirm if our culture model recapitulates more specific aspects of cancer dormancy and resurgence.

Our use of various bioinformatic tools enabled us to more closely inspect our datasets describing heterogeneous collective cancer migration. However, these guided analyses were only possible because we possessed phenotypic metadata. We demonstrate that cell cycle, pseudotime trajectory, and gene pattern expression analysis were able to isolate patterns and signatures that described specific transcriptomic subgroups, but we were only able to assign these patterns to particular phenotypes because of our phenotypic sorting platform. Our methodology illustrates the power behind a supervised, functionally informed computational analysis and presents an opportunity towards more direct approaches at investigating heterogeneity.

3.4 Methods

3.4.1 Cell Culture

MDA-MB-231 cells were a gift from Adam Engler (UCSD Bioengineering) and 4T1 cells were obtained from ATCC (Manassas, VA). All cells were cultured in high glucose Dulbecco's modified Eagle's medium supplemented with 10% (v/v) fetal bovine serum (FBS, Corning, Corning, NY) and 0.1% gentamicin (Gibco Thermofisher, Waltham, MA) and

maintained at 37°C and 5% CO_2 in a humidified environment during culture and imaging. The cells were passaged every 2-3 days. Cells were tested for mycoplasma contamination using the Mycoalert kit (Lonza, Basel, Switzerland).

To generate MDA-MB-231 cells that express Dendra2, we generated viral particles by cloning a Dendra2-Lifeact-7 plasmid (Addgene #54694, Watertown, MA) into a lentiviral vector. We then transfected the plasmid into lentiX293 T cells (Clonetech, Mountain View, CA. Cat #632180) along with packaging expressing plasmid (psPAX2, Addgene #12260) and envelope expressing plasmid (pMD2.G, Addgene #12259). Viral particles were collected at 48 h after transfection and they were purified by filtering through a 0.45 μm filter. MDA-MB-231 cells were then transduced with the viral particles in the presence of polybrene (Allele Biotechnology, San Diego, CA).

3.4.2 3D culture in type I collagen hydrogels

Cells embedded in 3D collagen matrices were prepared by mixing cells suspended in culture medium and 10× reconstitution buffer, 1:1 (v/v). Polyethylene glycol (PEG, Sigma-Aldrich, St. Louis, MO) was diluted in phosphate buffered saline (PBS, Gibco Thermofisher, Waltham, MA) and added to achieve a final concentration of 10 mg/mL. Soluble rat tail type I collagen in acetic acid (Corning, Corning, NY) was added to achieve a final concentration of 2.5 mg/mL. 6.25% of 1 M NaOH (volume of NaOH / volume of type I collagen) was used to normalize pH and the mixture was polymerized at 37 °C.

3.4.3 Phenotypic cell sorting

Collagen gels containing MDA Dendra cells were transferred to a microscope stage top incubator. Collective cell structures were identified using a Nikon TiE fluorescent microscope

(Nikon Instruments Inc., Melville, NY). Regions of interest were outlined using NIS-Elements software, and a Galvo Miniscanner (Nikon Instruments Inc.) was used to control the exposure of 405 nm laser from a Nikon LUnA power source (Nikon Instruments Inc.) to the outlined region to photoconvert the selected cells. 25% laser power with a 300 us dwell time were used to photoconvert the cells. This results in less than 1mJ of energy delivered to each multicellular structure. To ensure fidelity of converting only the desired phenotype, we did not photoconvert overlapping cell structures. The collagen gel was then digested using collagenase for 15 minutes at 37ºC (Sigma-Aldrich) and the cells were resuspended in FACS buffer (1% BSA, 0.5 mM EDTA in PBS). A gel with cells that were not photoconverted was used as a sorting control. Cells were sorted at the stem cell core of Sanford Consortium of Regenerative Medicine (La Jolla, CA) using a BD Influx cell sorter (BD, Franklin lakes, NJ). The cells from the control gel were used to establish a negative gate, and cells expressing red fluorescence above that gate were collected for re-culture or sequencing.

3.4.4 Single cell sequencing and analysis

RNA extraction and library construction were performed using the Chromium Single Cell 3' v3 kit (10x genomics, Pleasanton, CA). At least 700 cells were extracted for the network and spheroid cells, with one independent experiment each. These two experiments were pooled together for sequencing. The non-sorted MDA cells and the 4T1s were also extracted from one independent experiment each, and these were pooled together for sequencing. RNA integrity was verified using RNA Analysis ScreenTape (Agilent Technologies, La Jolla, CA) before sequencing. The RNA was sequenced on the Illumina HiSeq 4000 at a depth of > 20,000 reads per cell. The cellRanger analysis pipeline was used to construct the human reference genome GRCh38 and align reads. Differential gene expression analysis was performed using Seurat(*46*).

Cell expression data was filtered, log-normalized, and scaled prior to differential expression analysis using the non-parametric Wilcoxon rank sum test. Differentially expressed genes were filtered for those that had an absolute log-fold change of > 0.25 and expressed in at least 10% of either subpopulation with an adjusted $P < 0.05$ (Bonferroni).

3.4.5 Phenotypic analysis

Phenotype quantification for the reseeding experiments were performed in NIS-Elements using at least three independent experiments performed on different days with different cell passages.

3.4.6 Immunofluorescence and cell imaging

For cell imaging after 7 days of culture to visualize collective phenotypes, collagen gels were fixed with 4% PFA for 30 min at room temperature. F-actin was stained using AlexaFluor® 488 Phalloidin (Cell Signaling Technology, Danvers, MA) and the nuclei were counterstained with DAPI. For immunofluorescence staining the gels were incubated with the primary antibody for 24 hrs at 4^{o}C. The antibodies used were anti-CD24 (1:200 dilution, NB120-6586, Novus Biologicals, Littleton, CO), anti-CD44(1:200 dilution, MAB19562, Millipore Sigma).

3.4.7 TCGA data analysis

Breast cancer patient transcriptomic data and corresponding clinical metadata were obtained from the TCGA data portal. Patients were scored with respect to the 178 DEGs that were expressed differentially between the collectively invasive and non-invasive cells. A summation of the z-scores for the expression of each gene, aligned to whether the expression was correlated to the invasive or non-invasive phenotype, was used to score each patient. Survival analysis was

conducted using Kaplan-Meier plots, and the log-rank test was used to determine significance of survival differences between groups.

3.5 Supplementary Material

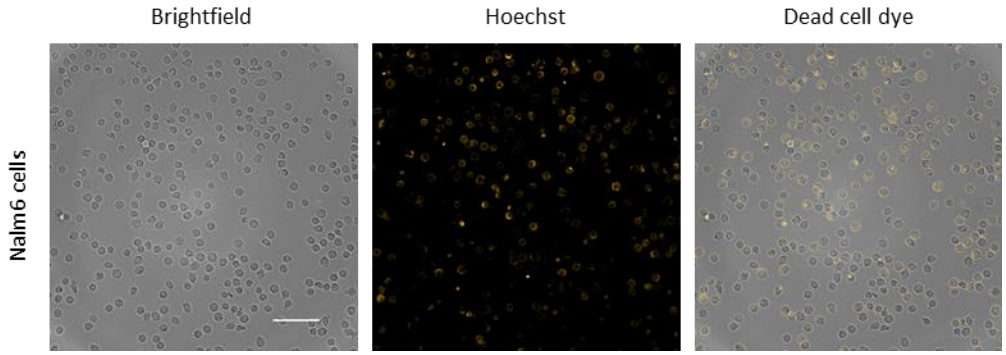


Figure S3.1. CD24 staining of Nalm6 cells as a positive control for the antibody used to stain for CD24. Scale bar 100 µm.

Table S3.1. Surface Markers That Are Differentially Expressed Between the Spheroid Populations.

CD82
NT5E
CD74
PRNP
CD59
CD40
BST2
ITGA2
NRP1
CDCP1
PLAUR
CD9
MCAM
LAMP1
CD63
CD33
BSG
JAG1
ITGA1
CD55
CD320
F3

The contents of this dissertation contain partial reprints of manuscripts being submitted to journals for publication. The dissertation author was the primary investigator and author of all papers. The following chapters contain material that have been submitted for publication as it may appear in iScience, 2020, Chen, Kevin; Ozturk, Kivilcim; Contreras, Ryne L; Simon, Jessica; McCann, Sean; Chen, Wei Ji; Carter, Hannah; Fraley, Stephanie I, Cell Press, 2020. The following chapters also contain material currently being prepared for submission for publication. Chen, Kevin; Ozturk, Kivilcim; Liefeld, Ted; Reich, Michael; Mesirov, Jill; Carter, Hannah; Fraley, Stephanie I.

Bibliography

1. G. H. Heppner, Tumor heterogeneity. *Cancer Res.* **44**, 2259–2265 (1984).

2. D. R. Welch, Tumor heterogeneity--A "contemporary concept" founded on historical insights and predictions. *Cancer Res.* **76**, 4–6 (2016).

3. N. McGranahan, C. Swanton, Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell*. **28**, 141 (2015).

4. D. F. Quail, J. A. Joyce, Microenvironmental regulation of tumor progression and metastasis. *Nat. Med.* **19**, 1423–1437 (2013).

5. P. Beri, A. Popravko, B. Yeoman, A. Kumar, K. Chen, E. Hodzic, A. Chiang, A. Banisadr, J. K. Placone, H. Carter, S. I. Fraley, P. Katira, A. J. Engler, Cell Adhesiveness Serves as a Biophysical Marker for Metastatic Potential. *Cancer Res.* **80**, 901–911 (2020).

6. P. Dalerba, T. Kalisky, D. Sahoo, P. S. Rajendran, M. E. Rothenberg, A. A. Leyrat, S. Sim, J. Okamoto, D. M. Johnston, D. Qian, M. Zabala, J. Bueno, N. F. Neff, J. Wang, A. A. Shelton, B. Visser, S. Hisamori, Y. Shimono, M. van de Wetering, H. Clevers, M. F. Clarke, S. R. Quake, Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.* **29**, 1120–1127 (2011).

7. I. Tirosh, B. Izar, S. M. Prakadan, M. H. Wadsworth 2nd, D. Treacy, J. J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, M. Fallahi-Sichani, K. Dutton-Regester, J.-R. Lin, O. Cohen, P. Shah, D. Lu, A. S. Genshaft, T. K. Hughes, C. G. K. Ziegler, S. W. Kazer, A. Gaillard, K. E. Kolb, A.-C. Villani, C. M. Johannessen, A. Y. Andreev, E. M. Van Allen, M. Bertagnolli, P. K. Sorger, R. J. Sullivan, K. T. Flaherty, D. T. Frederick, J. Jané-Valbuena, C. H. Yoon, O. Rozenblatt-Rosen, A. K. Shalek, A. Regev, L. A. Garraway, Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. **352**, 189–196 (2016).

8. W. Chung, H. H. Eum, H.-O. Lee, K.-M. Lee, H.-B. Lee, K.-T. Kim, H. S. Ryu, S. Kim, J. E. Lee, Y. H. Park, Z. Kan, W. Han, W.-Y. Park, Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* **8**, 15081 (2017).

9. S. Cheng, Y. Pei, L. He, G. Peng, B. Reinius, P. P. L. Tam, N. Jing, Q. Deng, Single-cell RNA-seq reveals cellular heterogeneity of pluripotency transition and X chromosome dynamics during early mouse development. *Cell Rep.* **26**, 2593-2607.e3 (2019).

10. E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, S. A. McCarroll, Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*. **161**, 1202–1214 (2015).

11. C. Xu, Z. Su, Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*. **31**, 1974–1980 (2015).

12. J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, E.-A. D. Amir, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, R. Finck, A. L. Gedman, I. Radtke, J. R. Downing, D. Pe'er, G. P. Nolan, Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*. **162**, 184–197 (2015).

13. R. S. Ahn, K. Taravati, K. Lai, K. M. Lee, J. Nititham, R. Gupta, D. S. Chang, S. T. Arron, M. Rosenblum, W. Liao, Transcriptional landscape of epithelial and immune cell populations revealed through FACS-seq of healthy human skin. *Sci. Rep.* **7** (2017), doi:10.1038/s41598-017-01468-y.

14. A. K. Mitra, U. K. Mukherjee, T. Harding, J. S. Jang, H. Stessman, Y. Li, A. Abyzov, J. Jen, S. Kumar, V. Rajkumar, B. Van Ness, Single-cell analysis of targeted transcriptome predicts drug sensitivity of single cells within human myeloma tumors. *Leukemia*. **30**, 1094–1102 (2016).

15. A. Haeger, K. Wolf, M. M. Zegers, P. Friedl, Collective cell migration: guidance principles and hierarchies. *Trends Cell Biol.* **25**, 556–566 (2015).

16. P. Friedl, D. Gilmour, Collective cell migration in morphogenesis, regeneration and cancer. *Nat. Rev. Mol. Cell Biol.* **10**, 445–457 (2009).

17. K. J. Cheung, A. J. Ewald, A collective route to metastasis: Seeding by tumor cell clusters. *Science*. **352**, 167–169 (2016).

18. D. P. Tabassum, K. Polyak, Tumorigenesis: it takes a village. *Nature Reviews Cancer*. **15** (2015), pp. 473–483.

19. N. Aceto, A. Bardia, D. T. Miyamoto, M. C. Donaldson, B. S. Wittner, J. A. Spencer, M. Yu, A. Pely, A. Engstrom, H. Zhu, B. W. Brannigan, R. Kapur, S. L. Stott, T. Shioda, S. Ramaswamy, D. T. Ting, C. P. Lin, M. Toner, D. A. Haber, S. Maheswaran, Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell*. **158**, 1110–1122 (2014).

20. K. J. Cheung, E. Gabrielson, Z. Werb, A. J. Ewald, Collective invasion in breast cancer requires a conserved basal epithelial program. *Cell*. **155**, 1639–1651 (2013).

21. D. O. Velez, B. Tsui, T. Goshia, C. L. Chute, A. Han, H. Carter, S. I. Fraley, 3D collagen architecture induces a conserved migratory and transcriptional response linked to vasculogenic mimicry. *Nat. Commun.* **8**, 1651 (2017).

22. R. Fisher, L. Pusztai, C. Swanton, Cancer heterogeneity: implications for targeted therapeutics. *Br. J. Cancer*. **108**, 479–485 (2013).

23. P. R. Prasetyanti, J. P. Medema, Intra-tumor heterogeneity from a cancer stem cell perspective. *Mol. Cancer*. **16** (2017), doi:10.1186/s12943-017-0600-4.

24. L. González-Silva, L. Quevedo, I. Varela, Tumor functional heterogeneity unraveled by scRNA-seq technologies. *Trends Cancer*. **6**, 13–19 (2020).

25. A. Davis, R. Gao, N. Navin, Tumor evolution: Linear, branching, neutral or punctuated? *Biochim. Biophys. Acta Rev. Cancer*. **1867**, 151–161 (2017).

26. V. Y. Kiselev, T. S. Andrews, M. Hemberg, Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20**, 273–282 (2019).

27. L. Chappell, A. J. C. Russell, T. Voet, Single-Cell (Multi)omics Technologies. *Annual Review of Genomics and Human Genetics*. **19** (2018), pp. 15–41.

28. A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, A. Regev, Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*. **167**, 1853-1866.e17 (2016).

29. D. A. Jaitin, A. Weiner, I. Yofe, D. Lara-Astiaso, H. Keren-Shaul, E. David, T. M. Salame, A. Tanay, A. van Oudenaarden, I. Amit, Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell*. **167**, 1883-1896.e15 (2016).

30. C. R. Cadwell, A. Palasantza, X. Jiang, P. Berens, Q. Deng, M. Yilmaz, J. Reimer, S. Shen, M. Bethge, K. F. Tolias, R. Sandberg, A. S. Tolias, Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nat. Biotechnol.* **34**, 199–203 (2016).

31. L. Kester, A. van Oudenaarden, Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell*. **23**, 166–179 (2018).

32. E. Lein, L. E. Borm, S. Linnarsson, The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science*. **358**, 64–69 (2017).

33. J. Konen, E. Summerbell, B. Dwivedi, K. Galior, Y. Hou, L. Rusnak, A. Chen, J. Saltz, W. Zhou, L. H. Boise, P. Vertino, L. Cooper, K. Salaita, J. Kowalski, A. I. Marcus, Image-guided genomics of phenotypically heterogeneous populations reveals vascular signalling during symbiotic collective cancer invasion. *Nat. Commun.* **8**, 15078 (2017).

34. S. K. Ranamukhaarachchi, R. N. Modi, A. Han, D. O. Velez, A. Kumar, A. J. Engler, S. I. Fraley, Macromolecular crowding tunes 3D collagen architecture and cell morphogenesis. *Biomater. Sci.* **7**, 618–633 (2019).

35. A. J. Grimbergen, J. Siebring, A. Solopova, O. P. Kuipers, Microbial bet-hedging: the power of being different. *Curr. Opin. Microbiol.* **25**, 67–72 (2015).

36. K. Hinohara, H.-J. Wu, Sébastien Vigneau, T. O. McDonald, K. J. Igarashi, K. N. Yamamoto, T. Madsen, A. Fassl, S. B. Egri, M. Papanastasiou, L. Ding, G. Peluffo, O. Cohen, S. C. Kales, M. Lal-Nag, G. Rai, D. J. Maloney, A. Jadhav, A. Simeonov, N. Wagle, M. Brown, A. Meissner, P. Sicinski, J. D. Jaffe, R. Jeselsohn, A. A. Gimelbrant, F. Michor, K. Polyak, KDM5 histone demethylase activity links cellular transcriptomic heterogeneity to therapeutic resistance. *Cancer Cell*. **35**, 330–332 (2019).

37. M. Duan, J. Hao, S. Cui, D. L. Worthley, S. Zhang, Z. Wang, J. Shi, L. Liu, X. Wang, A. Ke, Y. Cao, R. Xi, X. Zhang, J. Zhou, J. Fan, C. Li, Q. Gao, Diverse modes of clonal evolution in HBV-related hepatocellular carcinoma revealed by single-cell genome sequencing. *Cell Res.* **28**, 359–373 (2018).

38. W. Zhang, A. Bojorquez-Gomez, D. O. Velez, G. Xu, K. S. Sanchez, J. P. Shen, K. Chen, K. Licon, C. Melton, K. M. Olson, M. K. Yu, J. K. Huang, H. Carter, E. K. Farley, M. Snyder, S. I. Fraley, J. F. Kreisberg, T. Ideker, A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet.* **50**, 613–620 (2018).

39. Y. H. Choi, J. K. Kim, Dissecting cellular heterogeneity using single-cell RNA sequencing. *Mol. Cells*. **42**, 189–199 (2019).

40. H. Wang, S. Lacoche, L. Huang, B. Xue, S. K. Muthuswamy, Rotational motion during three-dimensional morphogenesis of mammary epithelial acini relates to laminin matrix assembly. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 163–168 (2013).

41. C. M. Williams, A. J. Engler, R. D. Slone, L. L. Galante, J. E. Schwarzbauer, Fibronectin expression modulates mammary epithelial cell proliferation during acinar differentiation. *Cancer Res.* **68**, 3185–3192 (2008).

42. B. A. Weaver, How Taxol/paclitaxel kills cancer cells. *Mol. Biol. Cell*. **25**, 2677–2681 (2014).

43. A. Tanay, A. Regev, Scaling single-cell genomics from phenomenology to mechanism. *Nature*. **541**, 331–338 (2017).

44. S. V. Sharma, D. Y. Lee, B. Li, M. P. Quinlan, F. Takahashi, S. Maheswaran, U. McDermott, N. Azizian, L. Zou, M. A. Fischbach, K.-K. Wong, K. Brandstetter, B. Wittner, S. Ramaswamy, M. Classon, J. Settleman, A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell*. **141**, 69–80 (2010).

45. S. M. Shaffer, M. C. Dunagin, S. R. Torborg, E. A. Torre, B. Emert, C. Krepler, M. Beqiri, K. Sproesser, P. A. Brafford, M. Xiao, E. Eggan, I. N. Anastopoulos, C. A. Vargas-Garcia, A. Singh, K. L. Nathanson, M. Herlyn, A. Raj, Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*. **546**, 431–435 (2017).

46. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck 3rd, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data. *Cell*. **177**, 1888-1902.e21 (2019).

47. H. Mi, A. Muruganujan, D. Ebert, X. Huang, P. D. Thomas, PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).

48. D. O. Velez, S. K. Ranamukhaarachchi, A. Kumar, R. N. Modi, E. W. Lim, A. J. Engler, C. M. Metallo, S. I. Fraley, 3D collagen architecture regulates cell adhesion through degradability, thereby controlling metabolic and oxidative stress. *Integr. Biol. (Camb.)*. **11**, 221–234 (2019).

49. J. Cao, M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers, C. Trapnell, J. Shendure, The single-cell transcriptional landscape of mammalian organogenesis. *Nature*. **566**, 496–502 (2019).

50. E. J. Fertig, J. Ding, A. V. Favorov, G. Parmigiani, M. F. Ochs, CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. *Bioinformatics*. **26**, 2792–2793 (2010).

51. L. N. Handly, J. Yao, R. Wollman, Signal transduction at the single-cell level: Approaches to study the dynamic nature of signaling networks. *J. Mol. Biol.* **428**, 3669–3682 (2016).

52. T. Du, L. Zhu, K. M. Levine, N. Tasdemir, A. V. Lee, D. A. A. Vignali, B. Van Houten, G. C. Tseng, S. Oesterreich, Invasive lobular and ductal breast carcinoma differ in immune response, protein translation efficiency and metabolism. *Sci. Rep.* **8** (2018), doi:10.1038/s41598-018-25357-0.

53. C. M. Neophytou, T.-C. Kyriakou, P. Papageorgis, Mechanisms of metastatic tumor dormancy and implications for cancer therapy. *Int. J. Mol. Sci.* **20**, 6158 (2019).

54. T. G. Phan, P. I. Croucher, The dormant cancer cell life cycle. *Nat. Rev. Cancer*. **20**, 398–411 (2020).

55. H. Endo, M. Inoue, Dormancy in cancer. *Cancer Sci.* **110**, 474–480 (2019).

56. A. Boire, S. B. Coffelt, S. A. Quezada, M. G. Vander Heiden, A. T. Weeraratna, Tumour dormancy and reawakening: Opportunities and challenges. *Trends Cancer*. **5**, 762–765 (2019).

57. M. Al-Hajj, M. S. Wicha, A. Benito-Hernandez, S. J. Morrison, M. F. Clarke, Prospective identification of tumorigenic breast cancer cells. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3983–3988 (2003).