UNIVERSITY OF CALIFORNIA,
IRVINE

Finding signals in the noise: Elucidating the many sources of heterogeneity in breast cancer metastasis using single-cell 'omics

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical, Computational, and Systems Biology

by

Kerrigan Blake

Dissertation Committee:
Assistant Professor Devon A. Lawson, Chair
Professor Kim N. Green
Professor Marian L. Waterman
Associate Professor Olivier Cinquin
Assistant Professor Timothy L. Downing

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Devon Lawson, for your continuous support and advice throughout this process. You have taken setbacks in stride and celebrated every accomplishment of yourself and those around you. The positive scientific environment that you fostered made it very easy to learn and grow and you made my PhD an enviable experience. I would be remiss to not also include Dr. Kai Kessenbrock for his contributions to my education and lab experience in the form of both scientific advice and lots of well-timed jokes. Together, Devon and Kai have served as leaders of a science team I'm proud to be a part of and a pair that constantly remind me to be ambitious.

To the rest of the lab, thank you for your constant help and excellent company. Of particular mention: Katrina, without your hard work and dedication, the microglia study would not exist, and I am forever grateful to you for letting me join you on this amazing project. And to computational corner, Ryan, Nick, and honorary members Kevin & Quy, I suspect I would have been more productive sitting somewhere else, but your particular style of peanut gallery sarcasm mixed with genuine scientific and emotional support made work fun.

To my committee, while I did not get to meet with you nearly enough, your attentiveness and advice on these projects was vital to our understanding and presentation of the material within them. You are all scientists that I deeply respect and wish to emulate.

And to the rest of CCBS, both administrative and academic, thank you for caring about and listening to graduate students like myself – you softened every bump in the road and made me so comfortable that I almost didn't leave. To my MCSB classmates, thank you for all the laughs, debates, homework sessions, and vacations (and an extra thank you to Lara, Matt, Tessa, and Joe, the friends I spent much of my time with for the last six and a half years).

Another thanks to all of my other friends and family who have supported me through this journey and beyond. A special mention for my father, Terry Blake, for instilling in me an appreciation for science and mathematics very early in life that you have only helped grow over the years by engaging in discussions of genetics, psychology, and medicine with me each week.

Finally, I want to thank Springer Nature for permission to include figures and text for Chapter 2, originally published in *Nature Communications,* and Chapter 3, originally published in *Nature Cell Biology,* as part of my dissertation. I also want to thank Dustin Maurer for his contributions to the forward selection, logistic regression method in Chapter 3. And I want to again thank Katrina Taylor Evans for her permission to use our unpublished, co-authored manuscript as Chapter 4 of my dissertation. Financial support was provided by the National Institute of Health - NIHT32EB009418, the UCI Public Impact Fellowship, and CCBS in the form of an opportunity award that became published work in Chapter 3.

VITA
Kerrigan Blake

EDUCATION:

Ph.D. in Mathematical, Computational, and Systems Biology 2014 – 2021
University of California, Irvine

B.S. in Mathematics, Magna Cum Laude 2010 – 2014
University of Kansas

RESEARCH EXPERIENCE:

Graduate Student Researcher: Devon Lawson Lab Apr. 2017 – Present
University of California, Irvine
Projects:

- Characterizing the role of microglia in tumor progression and immune responses during breast cancer brain metastasis using single-cell RNA-seq.
- Elucidating genetic heterogeneity in breast cancer metastasis using bulk and single-cell whole exome sequencing of patient-derived xenografts.
- Identification of immune cell heterogeneity in healthy human breast tissue for the Human Cell Atlas project using single-cell RNA-seq. (*manuscript in preparation*)

Graduate Student Researcher: Scott Atwood Lab June 2015 – Apr. 2017
University of California, Irvine
Projects:

- Generating biophysical models of how the actin cytoskeleton influences cell signaling through primary cilia length control using ordinary and partial differential equations in Matlab.
- Establishing the role of phosphorylation on zinc finger transcription factor DNA binding using in vitro techniques including molecular cloning, gel shift assays, and cell culture transfections.

Undergraduate Student Researcher: Eric Deeds Lab Oct. 2012 – May 2014
University of Kansas
Projects:

- Modeling the dependence of tumor growth rate and size on metabolic resources and oxygen diffusion using an ordinary differential equation model in C++.

Undergraduate Student Researcher: John Karanicolas Lab Jan. 2011 – Oct 2012
University of Kansas
Projects:

- Generating computational space-filling amino acid mutations using Rosetta in Mcl-1 to generate an improved crystal structure.
- Finding substrate binding pockets to stabilize ricin vaccines using Rosetta.

PUBLICATIONS:

Davis, R.T., **Blake, K.**, Ma, D. et al. Transcriptional diversity and bioenergetic shift in human breast cancer metastasis revealed by single-cell RNA sequencing. Nat Cell Biol 22, 310–320 (2020). https://doi.org/10.1038/s41556-020-0477-0

Nguyen, Q.H.*, Pervolarakis, N.*, **Blake, K.** et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. Nat Commun 9, 2028 (2018). https://doi.org/10.1038/s41467-018-04334-1

Ma D*, Hernandez GA*, Lefebvre AEYT, Alshetaiwi H, **Blake, K**, Dave KR, et al. Patient-derived xenograft culture-transplant system for investigation of human breast cancer metastasis. bioRxiv 2020.06.25.172056; doi: https://doi.org/10.1101/2020.06.25.172056

* denotes authors of equal contribution

GRANTS & FELLOWSHIPS:

NIH T32 Training Grant in Systems Biology (EB009418) Apr. 2015 – Apr. 2017
University of California, Irvine
Amount: Full PhD funding for two-years

Public Impact Fellowship Dec. 2019 – Dec. 2020
University of California, Irvine
Project Title: "Characterizing the response of microglia to breast cancer brain metastasis."
Amount: $1,000

AAU ALL-STEM Graduate Student Fellowship Sept. 2018 – Sept. 2019
University of California, Irvine
Project Title: "Effects of quantitative education on student attitudes and performance in genetics"
Amount: $500

COMPASS Outreach Grant Nov. 2017
American Society for Cell Biology
Project Title: "**Ask a Scientist Night** & Middle School Essay Content"
Amount: $210

Center for Complex Biological Systems Opportunity Award Aug. 2017 – Mar. 2018
University of California, Irvine
Project Title: "Investigating gene regulatory networks in metastasis"
Amount: $10,000

Center for Complex Biological Systems Opportunity Award Aug. 2017 – Mar. 2018
University of California, Irvine
Project Title: "Characterizing metastatic potential by integrated analysis of bulk and single-cell DNA Sequencing"
Amount: $10,000

Center for Complex Biological Systems Opportunity Award Aug. 2016 – Mar. 2017
University of California, Irvine
Project Title: "Gli phospohrylation and Sufu involvement in drug resistant Basal Cell Carcinoma"
Amount: $10,000

LEADERSHIP:

MCSB PhD Program Executive Committee Student Representative Sept. 2015 – Nov. 2020
University of California, Irvine
Overview: I was appointed by the MSCB PhD program director to serve as the sole student representative on the PhD program's executive committee which meets at least once a year to discuss student concerns and any curriculum or administrative changes.

Founder & Organizer of Biophysics and Systems Biology Seminar Series Sept. 2016 – Jan. 2019
University of California, Irvine
Overview: I started a series of research in progress (RIP) talks to give a forum for graduate students and postdocs to share their work with an interdisciplinary audience. The co-creator and I also worked with Dr. Jun Allard to get funding to expand the talk series to include student and faculty invited speakers. Our regular duties included speaker recruitment, scheduling, and maintaining our website for the first few years after its induction.

Mentor for UCI Cancer Research Institute Youth Science Fellowship Program July 2018
University of California, Irvine
Overview: I designed a small research project for the six-week intensive summer program for high schoolers that used a logistic regression model to predict patient outcomes from microarray datasets using gene signatures identified from single-cell RNA-sequencing. My mentee was able to present this work with a poster at American Society for Cell Biology (2018) and a talk at Western Medical Research Conference (2019) and returned to work with me on another project I designed in summer 2019, which can be found at https://github.com/jinwooee415/queryPathcards.

TEACHING EXPERIENCE:

Introduction to the High Performance Computing Cluster (HPC): Incoming Graduate Student Bootcamp Sept. 2019
University of California, Irvine
Overview: Designed & instructed 30-minute lecture on how to best use the bash and the HPC for parallelizing research tasks.

Cancer Systems Biology Short Course: Computational Methods for Single-Cell RNA-Sequencing Analysis Jan. 2019, May 2018
University of California, Irvine
Overview: Designed & instructed a four hour lecture and activity set on how to analyze 10X single-cell RNA-seq datasets from FASTQ to finalized clustering and GO term analysis.

Cancer Systems Biology Short Course: Duplex Sequencing Analysis Jan. 2019
University of California, Irvine
Overview: Assisted in the design and instruction of a four hour activity set on how to align with bwa, call and annotate SNPs with GATK/Annovar, and visualize duplex-sequencing genomics data with IGV.

Task Arrays on the High Performance Computing Cluster (HPC): Bioinformatics Support Group
University of California, Irvine Mar. 2018
Overview: Designed & instructed an hour lesson on how to organize scripts and use task arrays to parallelize jobs on the HPC.

Reporting Graduate Teaching Assistant – Genetics Sept. 2018 – Dec. 2018
University of California, Irvine
Overview: All of the duties of a regular TA in addition to duties including handling student exam concerns and assisting with the design and data analysis of a research study on student attitudes towards quantitative learning with Dr. Zeba Wunderlich for which I received the AAU ALL-STEM Graduate Student Fellowship.

Head Graduate Teaching Assistant – Genetics Sept. 2017 – Dec. 2017
University of California, Irvine
Overview: All of the duties of a regular TA in addition to duties including weekly meeting organization (3 instructors/8 TAs), handing student exam concerns, and assisting with special projects (online test bank construction and generation of custom scripts to create three-point cross and Hfr problem sets).

PRESENTATIONS:

Biophysics and Systems Biology Seminar Series: Speaker May. 2019
Title: *Single-cell analysis of the brain microenvironment in breast cancer brain metastasis*

Southern California Systems Biology Conference: Speaker Feb. 2019
Title: *Single-Cell analysis of the brain microenvironment in breast cancer brain metastasis*

Chan-Zuckerburg Initiative West Coast Retreat: Poster June 2018
Title: *Single-Cell analysis of the brain microenvironment in breast cancer brain metastasis*

Biophysics and Systems Biology Seminar Series: Speaker Apr. 2018
Title: *Identifying genetic patterns in breast cancer metastasis with single-cell sequencing*

Center for Complex Biological Systems Retreat: Speaker Mar. 2018
Title: *Investigating gene regulatory networks in metastasis*

Center for Complex Biological Systems Retreat: Speaker Mar. 2018
Title: *Characterizing metastatic potential with integrated analysis of bulk and single cell whole exome sequencing*

Northern California Computational Biology Symposium: Speaker Oct. 2017
Title: *Characterizing metastatic potential with integrated analysis of bulk and single cell whole exome sequencing*

Center for Complex Biological Systems Retreat: Speaker Mar. 2017
Title: *Gli phosphorylation and Sufu involvement in drug resistant Basal Cell Carcinoma*

UCI Skin Club: Speaker Mar. 2017
Title: *The biophysical roles of actin in mammalian primary cilia length regulation*

ABSTRACT OF THE DISSERTATION

Finding signals in the noise: Elucidating the many sources of heterogeneity in breast cancer metastasis using single-cell 'omics

by
Kerrigan Blake
Doctor of Philosophy in Mathematical, Computational, and Systems Biology
University of California, Irvine, 2021
Assistant Professor Devon A. Lawson, Chair

Despite improvement in screening and early detection, breast cancer remains the second leading cause of cancer related deaths among women according to the American Cancer Society. These deaths can be almost entirely attributed to metastatic disease, which is far more difficult to treat than local disease. It is therefore critical to gain a deeper understanding of what drives breast cancer metastasis and how the cells that surround metastatic tumors, known as the metastatic niche, respond to breast cancer cells to facilitate or inhibit their outgrowth. In this work, we use single-cell RNA-sequencing (scRNA-seq) and custom analytical pipelines to characterize breast cancer metastasis from multiple angles. We start by looking at pre-neoplastic breast epithelial cells and investigate the conserved lineage relationships between each epithelial cell state using a pseudotime analysis pipeline. We next look at matched primary tumor and metastatic cells from triple-negative breast cancer patient-derived xenograft models and identify biomarkers of micrometastasis (very small, early stage metastatic tumors) using generalized linear models which we validate are prognostically useful for relapse-free survival. And finally, we characterize the response heterogeneity of microglia, the brain resident macrophage, to breast cancer brain metastasis, demonstrate that these responses are conserved in human microglia, and show that microglia facilitate tumor regression using a genetic depletion model. In all of these projects,

we find that our newly defined cell states and tissue heterogeneity can be generalized across patients and models, suggesting that much of the noise seemingly inherent to breast cancer metastasis can be understood by asking the right questions. Further, by observing these systems at the single-cell level, we demonstrate the plasticity of breast epithelial cells in homeostasis, identify novel biomarkers and patient stratification opportunities for early breast cancer metastasis detection, and suggest new therapeutic routes for breast cancer brain metastasis patients.

CHAPTER 1: Applications of scRNA-seq in studies of the breast during health & disease

1.1 Open questions in mammary & breast biology

The mature mammary epithelium is made up of two major cell types, basal and luminal cells. Luminal cells are the secretory, milk-producing cells of the breast and basal and myoepithelial cells surround the luminal cells and attach to the basement membrane [1]. Mammary epithelial cells together form ducts extending from the nipple region that end in milk-containing alveoli which grow out during pregnancy and involve post-lactation [2]. The human breast epithelium has homologous cell types; however, the anatomical structure is more complicated, consisting of ducts that branch into 17-30 lobules, each of which contains multiple alveoli [3]. Within the luminal compartment, progenitor, hormone receptor (HR) positive, and HR negative cell states have been established, though the number of progenitors, their positions in the differentiation hierarchy, and their anatomical restrictions are still debated [1,4]. The basal compartment can also be further separated into mature myoepithelial cells, basal progenitors, as well as a small number of mammary stem cells (MaSCs), a bipotent progenitor state capable of recapitulating the entire mammary gland from a single cell upon transplantation into a cleared mammary fat pad [1,4]. As with luminal cells, the basal hierarchy is the subject of debate, especially its connection to the luminal lineage through MaSCs, which some believe is a state that is no longer present in adult basal cells during homeostasis, but can be induced by systemic perturbations [1,4].

Given the multitude of cell types and states in the epithelial hierarchy, it is perhaps unsurprising that breast cancer, derived from the breast epithelium, also has substantial heterogeneity [1]. Breast cancer subtypes are clinically stratified based on the protein-level expression of hormone receptors (estrogen receptor (ER) and progesterone receptor (PR)), HER2, and the proliferation marker KI67 [5]. The expression of these four proteins can suggest therapeutic regimens and secondarily predict the 'intrinsic' subtype of a given tumor. 'Intrinsic' subtypes refer to tumors with conserved gene expression profiles which correspond to different patient outcomes and metastatic proclivities [5–8]. Common subtypes include luminal A ([ER+|PR+] HER2-KI67-), luminal B ([ER+|PR+] HER2-KI67+), Her2-enriched ([ER-PR-] HER2+), and basal-like ([ER-PR-] HER2-) [5–8]. It is thought that these subtypes arise in breast cancer because oncogenic transformation can occur in distinct epithelial cell types and can be driven by different sets of mutations [1,5,6,9,10]. With the exception of the rare and particularly aggressive triple-negative ([ER-PR-] HER2-) subtype Claudin-low, hypothesized to arise from the MaSC [11], breast cancer is predicted to occur in luminal cells in various stages of differentiation. Basal-like breast cancer, the most common triple-negative subtype, is thought to derive from luminal progenitors while the hormone-receptor expressing luminal A and luminal B types are thought to come from more differentiated, HR+ luminal cells [1,6,9,10,12]. The Her2-enriched subtype appears to be primarily driven by a genomic amplification of HER2, so it may not have a single cell type of origin, though a luminal cell state only induced after a full term pregnancy (e.g. parity induced) has been proposed as a possible origin based on studies in mouse models [12,13].

More granular tumor subtyping schemas have also identified gene expression profiles suggestive of increased immune cell infiltration and immune responses in subsets of the

aforementioned intrinsic subtypes [14,15]. Microenvironmental contributions from immune and stromal cells to tumor gene expression profiles have mostly been treated as separate from tumor intrinsic gene expression and genetic alterations [16]. However, the microenvironment can substantially contribute to tumor heterogeneity through the 'seed and soil' hypothesis of cancer metastasis, which states that breast cancer cells can be thought of as 'seeds' that spread throughout the body, but only grow out in appropriately permissive 'soil' [17]. Each tumor subtype in breast cancer has its own 'organotropism', referring to distal organs to which it will commonly metastasize, supporting this idea of permissive and non-permissive microenvironments [18]. Additionally, these metastatic microenvironments appear able to adapt, or select for breast tumors of different subtypes than what was seen in the bulk of the primary tumor, a process referred to as 'metastatic switching' [19]. This switching almost always gives rise to a metastatic tumor with a more clinically 'aggressive' intrinsic subtype than what was seen in the primary tumor and promotes intra-patient breast cancer heterogeneity.

The relationship between the microenvironment and tumor subtypes in breast cancer warrants additional investigation, especially since the 'omics methods used to define breast cancer subtypes were not equipped to deconvolve the phenotypic contributions of healthy and neoplastic tissues. Further, while intertumoral heterogeneity, or patient to patient variation, is already well-characterized in breast cancer, the extent to which intratumoral heterogeneity influences breast cancer outcomes has lagged behind. However, the relatively recent advent of single-cell RNA sequencing (scRNA-seq) has hinted at the answers to these outstanding questions. Unlike prior methodologies, scRNA-seq can unbiasedly separate cell types and states and distinguish signals from both within and

between the microenvironmental cells and tumor cells. It can also reveal both discrete and continuous phenotypic changes, offering a more comprehensive view of the tumor ecosystem and intratumoral heterogeneity. In this review, we will discuss both the advantages and limitations of scRNA-seq for answering questions in breast biology, and examine how the single-cell studies published so far have enhanced our understanding of cellular states and plasticity in breast homeostasis and cancer.

1.2 Single-cell RNA-seq methodologies

1.2.1 Single-cell library construction considerations

Bulk RNA-seq and microarray analysis have been useful for unbiased studies of the mammary gland and breast cancer, allowing us to characterize and compare sorted cell types and establish many of the current tumor subtyping schemas [8,14,20]. Using bulk sequencing, cell types must be identified *a priori* by surface markers or transgenic reporter expression and subsequently separated using techniques like fluorescence-activated cell sorting (FACS). Using scRNA-seq, cell types can be separated computationally and identified on the basis of established gene expression patterns. Therefore, it does not require marker-based sorting, though this step can still be used for cell type enrichment if one chooses. However, scRNA-seq requires the separation of viable single cells from whole tissues which was not necessary in bulk methods, and technologies for scRNA-seq library construction can be broadly distinguished by how they capture these cells. Three common types of cell capture are microfluidic circuits (e.g. Fluidigm C1) [21], plate based capture (e.g. Smart-seq2) [22], and droplet based capture (e.g. 10X Genomics) [23]. In addition to cell capture, technologies can differ on whether they sequence full length RNA molecules or only the 3' or 5' ends of RNA

fragments [24]. Generally, microfluidic circuit and plate based methods are part of protocols that allow for better gene capture while droplet based methods are more scalable in terms of cell numbers. With all of the available technologies, scRNA-seq is prone to gene "dropout", referring to expressed but uncaptured transcripts in each cell [25]. As a result of dropout, single-cell library count matrices contain mostly zeros and lowly expressed genes like transcription factors can be too sparsely captured to interpret. Despite this, the captured genes are generally sufficient to discriminate cell types and states, allowing for the removal of untargeted cell populations in downstream analysis which was not possible using bulk sequencing methods. Overall, scRNA-seq improves upon the accuracy of cell type identification in transcriptomic studies but has low transcriptomic resolution compared to bulk methods due to dropout.

ScRNA-seq results in much "cleaner" transcriptomic libraries than bulk, but it is important to mention that scRNA-seq libraries still contain contaminants that can influence data interpretations. Minor background reads from nearby lysed cells or incomplete washes are almost always sequenced with each cell and therefore one should ensure that comparison groups and sequencing batches are non-overlapping. If this is not possible, batch correction algorithms have been developed specifically for scRNA-seq data to mitigate their effects [26]. A related issue that is unique to scRNA-seq libraries is "doublets", which refer to two cells sequenced under a single identifier. This can generate populations of "cells" with misleadingly interesting properties. Doublets represent a few percent of total capture for most scRNA-seq technologies and are unlikely to be completely avoided [23]. Doublet detection using biological marker knowledge in combination with a high unique gene number cutoff (since doublets tend to express more unique genes than any given single cell) seems to work

well, and more sophisticated algorithms have been developed for doublet detection when orthogonal validation is warranted [27–29]. Finally, non-biological or uninteresting transcriptomic variation is often present in single-cell libraries due to cell cycle states or stress associated signatures. To go from intact tissues to single cells, long manual and enzymatic digestion protocols can be necessary, especially in tissues like the breast which contains a dense extracellular matrix. These prolonged digestions can result in the expression of stress or early apoptosis genes in the separated cells. Similarly, scRNA-seq studies that perform FACS to enrich for live cells or specific subpopulations of interest sometimes find a stress associated profile consisting of intermediate early genes in a portion of their libraries. Studies of digestion and sorting induced phenotypes are starting to emerge [30–32], though the full range of tissue processing and sorting-associated states is not yet defined and it remains difficult to distinguish them from related *in vivo* phenotypes. All of the aforementioned technical issues can be mitigated by careful experimental design, but not fully removed without the aid of computational tools. Therefore, while scRNA-seq is a mostly unbiased capture methodology, the data cannot and should not be analyzed unbiasedly; rather, biological knowledge of the cell types and states of interest must be utilized to overcome technical limitations and identify relevant findings.

1.2.2 Overview of scRNA-seq computational analyses

ScRNA-seq data analysis often begins by performing clustering and dimensionality reduction (DR) to identify cell populations with distinct transcriptome profiles. In bulk RNA-seq, DR is performed primarily with principal component analysis (PCA) or multidimensional scaling (MDS) to compare biological replicates and conditions, and

transcriptome similarities are directly associated with the distances between libraries in the plot. Due to dropout and highly distinct subpopulations of cells found in single-cell datasets, scRNA-seq is better visualized using non-linear DR techniques, though linear DR techniques perform equally well for downstream statistical analyses [33]. Non-linear DR techniques reduce computational outlier effects compared to linear DRs and result in visualizations where groups of cells are transcriptionally similar but the amount of space between populations is not necessarily proportional to their transcriptional differences. The most frequently used techniques for non-linear DR in scRNA-seq are t-distributed stochastic neighbor embedding (tSNE) [34] and uniform manifold approximation and projection (UMAP) [35]. After or in conjunction with DR, clustering is performed to increase the granularity of population identification and label cells for differential expression testing. Graph-based, Louvain clustering methods have been shown to work particularly well for scRNA-seq datasets [24,36], and this method can be performed on either the non-reduced cell coordinates (for example, principal components or full gene expression profiles) or the cell positions in the DR space. The advantage of performing clustering on non-reduced cell coordinates is that it offers orthogonal validation for populations structures since the labels should align very well to the reduced cell positions in tSNE or UMAP space, supporting that the chosen DR captured relevant differences between cells. Overall, it is best to think of DR as a visualization tool and clustering as a statistical analysis tool which together reveal transcriptionally distinct cell populations for labeling and hypothesis generation.

Once populations are identified computationally, one must use biological expertise to determine the cell types and cell states represented by each cluster. Marker gene analysis is a type of differential expression analysis that looks at genes specifically upregulated in one

cluster compared to all of the other clusters. Using marker genes, one can label cell types and phenotypes through comparison to known cell type markers and gene ontology terms [37]. The phenotype labels can be refined by performing "gene scoring" on individual cells with bulk transcriptome profiles from prior cell type studies or curated gene lists for cell states. Briefly, gene scoring techniques seek to quantify the enrichment of an input transcriptome profile in aggregate to determine whether it is specifically expressed in a given cell or population of cells. Many algorithms exist for this purpose, but a popular heuristic was developed by Tirosh & Izar et al [38] which takes as input a list of genes assumed to be regulated unidirectionally (either up or down) and compares the expression of genes in this gene set to a random background set of genes with similar expression in the bulk of the data. As a result, a positive score means that the gene set is "upregulated" in a given cell and a negative score means that the gene set is "downregulated", though the absolute value of the scores do not have a clear interpretation since they are normalized within the dataset being investigated and are impacted by the algorithm parameters. It is also common to see gene scoring techniques developed for bulk sequencing data (e.g. GSEA [39]) used on scRNA-seq data, but this should be done with caution since the statistical assumptions of these algorithms do not necessarily hold in sparse, outlier heavy single-cell data.

After cell type and cell state identification, one can ask how certain cell types move through a differentiation or activation process. Since these processes are often conceptualized as continuous rather than discrete behaviors (i.e. a cell can be 'more stem like' or 'more activated' rather than existing in a binary 'stem' or 'activated' state), the algorithms for investigating them also implement these assumptions. This analysis is broadly referred to as 'pseudotemporal' analysis since it can separate individual cells moving

through a temporally dependent process by their differentiation 'time' (i.e. phenotypic state) rather than relying on the timepoint at which the data was collected. This can be done separately from the original DR, though PCA and UMAP capture global cell type similarities well enough to be used as part of pseudotemporal algorithms. Other commonly used DR bases for pseudotime include diffusion maps [40] and reversed graph embedding (e.g. DDRTree in Monocle 2) [41]. After placing the cells in two or three dimensions with a globally sensitive DR algorithm, a 'timecourse' is added to the DR as a path through the plot with a start and end point defined by the user. There are a few different methods for drawing this path, most of which rely on local cell density and minimum complexity assumptions (i.e. avoid drawing unnecessary detours between groups of cells) to determine feasible trajectories [24,42]. A relatively new addition to these analyses is RNA velocity, which seeks to establish the direction and speed of cell movements through a temporal process based on the ratios of spliced and unspliced transcripts [43,44]. This model assumes that a higher than expected ratio of unspliced, or nascent transcripts to spliced transcripts from a particular state suggests a movement towards that state, while the opposite indicates a movement away, or a repression of that state. These assumptions allow the algorithm to create a vector field of cell movements on top of an existing single-cell trajectory which can guide the interpretation of how cells move across a proposed path. Additionally, it can serve as orthogonal validation of hypothesized start, end, or transition states in a trajectory since it relies on nascent RNA capture that was not accounted for in the original DR or trajectory inference.

The computational methods we have discussed so far are frequently used algorithms for analyzing scRNA-seq datasets in the field of breast and breast cancer, but we have by no

means given a complete overview of the possible analysis techniques. For those interested in more detail on these methods or alternative options, Leucken and Theis, 2019 provides an excellent tutorial of computational methods and how they're used in scRNA-seq without field specificity [24]. For the rest of this review, we will occasionally refer back to these methods and their caveats as we discuss breast and mammary related scRNA-seq findings. As with other profiling methods, scRNA-seq is primarily used to generate hypotheses and the data is open to reinterpretation as new biological understandings and computational methods become available. **Table 1** provides a list of the peer-reviewed studies we will cover in the subsequent sections, filtered to contain only *in vivo* derived libraries, and includes the location of the associated datasets to facilitate their reuse, reanalysis, and meta-analysis for novel hypothesis generation.

Paper	Organism	Cell types sequenced	Microfluidic-based	Plate-based	Drop-seq	ScRNA-seq data locations
Nguyen & Pervolarakis et al, 2018	Human	Breast epithelia	Fluidigm C1		10x	GSE113197
Sun et al, 2018	Mouse (FVB)	Mammary epithelia	Fluidigm C1			Table S1
Giraddi et al, 2018	Mouse (C57BL/6)	Mammary epithelia	Fluidigm C1		10x	SAMN07138894, GSE111113
Bach et al, 2017	Mouse (C57BL/6)	Mammary epithelia			10x	GSE106273
Pal, Chen, & Vaillant et al, 2017	Mouse (FVB)	Mammary epithelia	Fluidigm C1		10x	GSE98131, GSE103275
Wuidart & Sifrim et al, 2018	Mouse (C57BL/6)	Mammary epithelia		SmartSeq2		GSE109711
Davis et al, 2020	Human (PDX in NOD/SCID)	Tumor cells		SmartSeq2		GSE123837
Karaayvaz & Cristea et al, 2018	Human	Tumor, stroma, & immune cells		SmartSeq2		GSE118390
Chung & Eum et al, 2017	Human	Tumor, stroma, & immune cells	Fluidigm C1			GSE75688
Bartoschek et al, 2018	Mouse (MMTV-PyMT)	Stromal cells		SmartSeq2		GSE111229

Kieffer & Hocine et al, 2020	Human	Stromal cells			10x	EGAS00001004030
Azizi, Carr, Plitas, & Cornish et al, 2018	Human	Immune cells			inDrop/10x	GSE114727, GSE114725, GSE114725
Savas & Virassamy et al, 2018	Human	Immune - T cells			10x	GSE110686
Alshetaiwi et al, 2020	Mouse (MMTV-PyMT)	Immune - MDSCs			10x	GSE139125

Table 1: ScRNA-seq papers from *in vivo* studies focused on the mammary gland, breast, and breast cancer. Data accessions beginning with S can be accessed through the short read archive (SRA), those beginning with G can be accessed through the gene expression omnibus (GEO), and those beginning with E can be accessed through the European Genome-Phenome Archive. MDSC = myeloid derived suppressor cell.

1.3 Biological insights derived from scRNA-seq studies

1.3.1 The healthy mammary and breast epithelium

Single-cell studies in the mammary gland so far have been able to unbiasedly assay the epithelial cells from embryonic tissue, pre-puberty, post-puberty, pregnancy, and post-pregnancy involution [45–49]. Notably, all single-cell datasets showed one major differentiated basal population and two major differentiated luminal populations (HR+ and HR-) in the homeostatic adult mammary gland, as well as the adult breast, which matches well with the expectations from flow cytometry studies [45–51]. However, there was less consensus on the number and markers of progenitor populations. Some groups argue that their data did not suggest the persistence of a bipotent progenitor resembling a MaSC in the adult mammary gland [45,46,49] and others proposed CDH5 [48] or luminal-progenitor/basal associated co-expression [47] as data-supported markers. Sun et al demonstrated that their hypothesized MaSC cluster, marked by *Procr, Cldn5, Pecam1,* and other vascular-related genes, could be sorted using CDH5 and that CDH5+ basal cells had a greater ability to reconstruct the mammary gland in transplantation assays than CDH5- basal cells [48]. Pal, Chen, & Vaillant et

al did not demonstrate biopotency in their luminal-progenitor marked basal cells, but they did validate a novel marker of luminal stemness [47]. This group noticed a small number of CD55+ epithelial cells that moved from the basal compartment in pre-puberty to the luminal compartment in the adult mammary gland, which they hypothesized represented an intermediate luminal progenitor [47]. They demonstrated that these CD14+CD55+ luminal cells had a higher colony forming capacity in Matrigel than CD14- or CD55- luminal cells, supportive of increased progenitor capabilities [47]. Embryonic MaSCs, which require less experimental validation since they are guaranteed to have stem-like phenotypes due to their differentiation timing, have also been included in scRNA-seq studies of the epithelial hierarchy. Interestingly, while gene scoring analyses supported a mixed basal/luminal lineage phenotype for embryonic MaSCs, neither study found any cells from the adult mammary gland transcriptionally similar enough to MaSCs to cluster with them [45,49].

Unfortunately, scRNA-seq did not readily end any debates on progenitor populations in the mammary epithelium, pseudotime analyses have revealed interesting cell state dynamics. Bach et al focused their study on luminal cells in the pre- and post-lactation mammary gland [46]. In this study, they find a smooth transcriptional bifurcation between luminal progenitors and HR+ or HR- differentiated luminal cells, and argue that their diffusion map does not support a connection between the basal and luminal lineages [46]. Interestingly, they also found that luminal cells appear to 'remember' their secretory alveolar state after pregnancy, since the HR+ luminal cells from post-parous mice skew towards the nulliparous HR- luminal cells rather than the nulliparous HR+ luminal cells when projected onto the same diffusion map [46]. In the human breast, a preprint from Murrow et al also found that parity resulted in decreased hormone-responsiveness in HR+ luminal

cells compared to nulliparous samples based on gene expression profiles downstream of the ER and PR pathways [51]. In terms of the differention hierarchy in human breast, Nguyen & Pervolorakis et al similarly found a smooth bifurcation between luminal cell progenitors and HR+ and HR- differentiated states in healthy reduction mammoplasty samples using Monocle2 [50]. This plasticity and apparent memory of luminal cells seen in both species may be an important source of intertumoral breast cancer heterogeneity and could also help to explain parity-associated differences in the incidence of breast cancer development across subtypes [46,51,52].

1.3.2 Inter- and intra-tumoral heterogeneity in breast cancer

Before the advent of scRNA-seq, it was hypothesized that luminal differentiation states corresponded to distinct tumor types with aggressive tumor subtypes arising from luminal progenitors (e.g. basal-like) and less aggressive tumor types arising from differentiated HR+ luminal cells (e.g. luminal A and luminal B). Single-cell studies to date have supported these phenotypic similarities between breast epithelial cell types and tumor subtypes using gene scoring analyses on healthy epithelial cells for tumor signatures as well as on tumor cells for healthy epithelial signatures [50,53,54]. These gene scoring analyses also revealed potential subtype mixing, with each tumor having a few cells classified as a different subtype than the bulk of the tumor [53,54]. Subtype mixing is an interesting and potentially relevant finding but it is important to remember that the PAM50 [7], METABRIC [14], and TNBC subtypes [20] were all developed on bulk datasets and may be sensitive to gene dropout. Chung & Eum et al, who profiled luminal A, luminal B, Her2-enriched, and triple-negative tumor cells, showed that the expression of subtype-specific markers and related pathways in

single-cells align very well with their matched bulk RNA-seq and the expected tumor subtypes from pathology [54]. However, a few cells from each tumor lack the expression of key markers for any subtype, likely indicative of high dropout [54]. Unfortunately, it is not yet clear whether the computational methods used for gene expression based breast cancer subtyping are mislabeling low quality cells as distinct subtypes because of technical artifacts or whether they are picking up on true biological differences and the hypothesis of subtype mixing should be taken lightly.

While subtype identification may not be optimized for scRNA-seq quite yet, single-cell studies have given us a great look into both inter- and intra-tumoral heterogeneity in breast cancer. Intertumoral heterogeneity is very apparent and well supported in single-cell studies, with every group finding that unbiased clustering is primarily driven by cell type in healthy cells and patient differences in malignant tumor cells [54–56]. Interestingly, Karaayvaz & Cristea et al found that genomic copy number variations, seen in both bulk whole exome sequencing as well as inferred from their scRNA-seq data, were highly correlated to patient specific gene expression profiles [55]. Thus, transcriptomic differences seen between different patient's tumors may be largely driven by genomic alterations. On the other hand, intratumoral heterogeneity, or differences between individual cells within a patient tumor, may be more impacted by the local environment. Based on clustering within individual patient tumors, it does appear that subgroups of tumors are dominated by distinct biological behaviors; however, non-linear DR techniques sensitive to global changes do not fully separate these populations, meaning these behaviors may be gradiented or have some overlap with one another [53,56,57]. Davis et al showed that individual patient-derived triple-negative xenograft tumors have clusters with different functional phenotypes based on

marker gene ontology analysis, including increased metabolism (glycolysis or oxidative phosphorylation), inflammatory responses, and extracellular matrix rearrangement [56]. Chung & Eum et al similarly show that individual tumor cells within a patient have heterogeneous expression of common breast cancer associated pathways, including HER2 amplification, PI3K/AKT, and estrogen response, though there were too few tumor cells sequenced to form defined clusters [54]. These subtle intratumoral specializations may be driven by the location of cells within the tumor mass, where cells in hypoxic regions, next to the tumor edge, or near cytokine secreting immune and stromal cells shift their transcriptomes to respond to the relevant stressors. No spatial follow-ups were performed in the scRNA-seq papers discussed so far; however, Jackson & Fischer et al performed imaging mass cytometry on 352 breast cancer patient tumors and this data supports that intratumoral heterogeneity is a spatially segregated phenomenon [58].

1.3.3 Insights into breast cancer metastasis

Metastatic initiation has been difficult to study due to its seemingly stochastic timing and progression, but new insights on this question have been gained from single-cell studies. A major question in metastasis is whether metastatic tumors derive from a rare tumor cell with definable properties or whether successful metastases are the consequence of chance and environmental adaptations. In general, metastatic tumors in breast cancer show more aggressive phenotypes than the primary tumors they arise from. Prior studies using bulk sequencing have shown that metastases from primary tumors with the basal-like subtype almost always seed basal-like metastases and the less aggressive subtypes often seed metastases with a worse prognosis (e.g. luminal B to basal-like) [19]. A version of metastatic

switching was observed at the single-cell level in Chung & Eum et al who found that a tumor with HER2+ER+ cells, classified as luminal B, showed an upregulation of genes associated with the HER2 amplification pathway and a downregulation of estrogen response in lymph node metastases compared to primary tumor cells, indicating priming towards the Her2-enriched phenotype in metastasis [54]. Additionally, using single-cell qPCR, Lawson et al demonstrated that triple-negative patient-derived xenograft micrometastases have basal/MaSC associated gene expression profiles while primary and high burden metastatic tumors had more luminal like gene expression patterns, suggesting that early metastasis may be facilitated by the upregulation of epithelial progenitor characteristics [59]. Perhaps the most important observation has been made in multiple papers; namely, Lawson et al, Davis et al, and Chung & Eum et al all found rare cells in the primary tumor reflecting phenotypes of their associated metastatic tumors, supporting the hypothesis of a non-random metastatic cell of origin [54,56,59].

In addition to supporting hypothesized profiles of metastasis in breast cancer, scRNA-seq has revealed unexpected connections between metabolism and metastatic progression in the triple-negative subtype. Giraddi et al found fetal MaSCs upregulate a large number of metabolism-associated genes compared to epithelial cells from older mice, especially those involved in glycolysis [45]. Using the metabolism associated marker genes for fetal MaSCs, they develop a signature that they find is enriched in basal-like breast tumors compared to other proliferative tumor subtypes and further note is upregulated in metastatic basal-like tumors compared to primary basal-like tumors [45]. In contrast, Davis et al found that micrometastatic cells from triple-negative patient-derived xenografts upregulate oxidative phosphorylation (OxPhos) associated genes and downregulate glycolytic associated genes and that metastasis

can be decreased by blocking OxPhos with Oligomycin [56]. Taken together, these two observations suggest that OxPhos is needed in early metastasis and glycolysis is beneficial in late metastasis. A third metabolic pathway relevant to metastasis was identified in Karaayvaz & Cristea et al, whose unbiased clustering revealed a glycosphingolipid metabolism signature conserved across their six triple-negative patient tumors and further found that this signature was associated with poor survival outcomes in triple-negative tumors from the METABRIC cohort [14,55]. While it remains to be seen how these metabolic pathways function in metastatic progression in triple-negative tumors, it seems clear that these tumors benefit from the expression of multiple metabolic pathways as well as metabolic switching capabilities.

1.3.4 Microenvironmental influences in breast cancer

The tumor microenvironment in the breast consists of healthy epithelial cells, stroma, and immune cells. The immune microenvironment can be further broken down into lymphoid and myeloid cells. Most myeloid cell types in the tumor microenvironment are considered supportive due to their growth-promoting and pro-angiogenic cytokine secretion and most lymphoid cell types are responsible for tumor cell recognition and killing [60]. Similar to myeloid cells, the stroma is considered pro-tumorigenic since it has been shown to promote tumor survival by providing growth factors and degrading the extracellular matrix [60]. ScRNA-seq studies of the tumor microenvironment in breast cancer so far have primarily sought to classify non-tumor cell types and label their activation states using known markers. Despite the relative simplicity of this goal, these investigations have

revealed that many 'known' tumor-associated phenotypes in immune and stromal cells have unexpected co-expression patterns and novel prognostic capabilities.

Tumor associated myeloid cells, including monocytes, neutrophils, macrophages (TAMs), and monocyte-derived suppressor cells (MDSCs), are prevalent in the breast cancer microenvironment and their pro- or anti-tumorigenic behaviors are associated with distinct gene expression patterns. A slightly oversimplified model states that myeloid cells with 'M1' signatures are pro-inflammatory and cytotoxic T cell promoting, while myeloid cells with 'M2' signatures are anti-inflammatory, pro-angiogenic, and T cell suppressive [61]. The studies that have investigated TAMs in breast cancer using scRNA-seq have all noted a pro-tumorigenic M2 phenotype in their captured macrophages based on gene scoring and marker gene investigations, though Azizi, Carr, Plitas, & Cornish et al noted an important correlate [54,62,63]. This group found that genes marking the M1 phenotype were co-expressed with M2 markers in individual myeloid cells from both healthy and tumor tissues across all breast cancer subtypes [62]. While M1 and M2 signatures have been known to have a complex balance in breast cancer, this is the first demonstration that a mixed signature is found at the single-cell level and it is unclear whether these hybrid cells will be pro- or anti-tumoral [64]. On the other hand, MDSCs, named based on their ability to suppress T cell function, are almost certainly pro-tumorigenic, though they are not always easy to identify in transcriptomic data due to their similarities to non-suppressive neutrophils and monocytes [65]. To find a more discriminating MDSC gene signature, Alshetaiwi et al used scRNA-seq to compare neutrophils and monocytes from the spleens of wild type and tumor-bearing MMTV-PyMT mice, which are enriched for MDSCs [66]. Using a straightforward clustering and marker gene analysis, they identified a conserved gene signature for MDSCs from tumor-

bearing mice, and found a novel flow sorting marker for these cells, CD84, which they validated marks neutrophils with high T cell suppressive capacities in the PyMT mouse and is also upregulated on MDSCs in the blood of human breast cancer patients [66]. These observations from scRNA-seq studies of myeloid cells in breast cancer so far have generated new activation profiles and sorting markers to facilitate improved functional studies in the future.

T cell infiltration patterns are known to be predictive of patient survival and response to therapeutics in breast cancers, so it is not surprising that their activation profiles have been a topic of interest in breast cancer related single-cell studies [67–69]. Similar to myeloid cells, T cells can have multiple states, including antitumoral cytotoxic CD8+ T cells and Th1 cells as well as immune suppressive Th2 and T regulatory (Treg) cells. Savas & Virassamy et al investigated T cell states in luminal, Her2-enriched, and triple-negative breast cancers using both scRNA-seq and bulk RNA-seq and found that the gene signatures of resident memory, CD103+CD8+ T cells in triple-negative tumors is more predictive of patient survival than the gene signatures for other T cell states, suggesting that CD103+CD8+ T cells are particularly important for tumor clearance in this subtype [70]. While T cell activation is vital to tumor clearance, prolonged activation can be followed by a state of exhaustion with increased expression of immune suppressive markers and Treg infiltration that facilitates tumor outgrowth [67,71]. Interestingly, Azizi, Carr, Plitas, & Cornish et al modeled T cell activation in breast cancer with diffusion maps and revealed a positive correlation between T cell activation/exhaustion and hypoxia, connecting the local tumor environment to T cell phenotypes [62]. Additionally, they found that T cell clonotypes (antigen-specific T cell populations), determined by VDJ sequencing, were aligned almost perfectly with scRNA-seq

driven clustering in those same T cells [62]. Therefore, there appears to be a connection between phenotype, genotype, and the metabolic environment in breast cancer-associated T cells, mirroring what was seen in the breast cancer cells themselves.

Stromal cells have more recently been studied using scRNA-seq, predominantly in a cancer context. In Bartoschek et al, they use the MMTV-PyMT mouse model to characterize three distinct populations of cancer associated fibroblasts (CAFs), associated with vasculature (vCAF), mammary residence (mCAF), and tumor cells having undergone epithelial-to-mesenchymal transition (dCAF) [72]. They find that the gene signature of vCAFs, very likely to be a population of pericytes, is associated with increased risk of developing metastatic disease [72]. Further, they connect the mCAF phenotype (PDGFRα+/CD146−) to the specification of hormone negativity in triple-negative breast cancers based on previous functional studies [73,74]. Similar fibroblast populations have been identified in humans and these studies have dug deeper into their range of cell states. Using flow cytometry, Pelon et al and Costa et al characterize four CAF subtypes, CAF-1 (FAP^{hi} $CD29^{med\text{-}hi}$ SMA^{hi}), CAF-2 (FAP^{neg} $CD29^{lo}$ SMA^{neg}), CAF-3 (FAP^{neg} $CD29^{med}$ SMA^{neg}), and CAF-4 (FAP^{neg} SMA^{hi} $CD29^{hi}$) found in metastatic lymph nodes and primary tumors from breast cancer patients[75,76]. Two of these populations, CAF-1, most similar to the mCAFs described in mice, and CAF-4, similar to the vCAFs or pericytes, are particularly enriched in triple-negative breast cancer and lymph node metastases[75,76]. Since Costa et al also found that the CAF-1 population was associated with triple-negative tumors with high Treg infiltration, and low CD8+ T cell infiltration, suggestive of a tumor supportive immune environment[76], Kieffer & Hocine et al investigated heterogeneity within the CAF-1 subpopulation in an additional breast cancer patient cohort using scRNA-seq to determine whether an even smaller subset of these

fibroblasts were driving the observed microenvironmental shifts[77]. Kieffer & Hocine et al sorted CAF-1 fibroblasts from six Luminal A and two triple-negative breast cancer patients, which identified eight subclusters of CAF-1 fibroblasts, which broadly fell into the three categories of myofibroblastic ("myCAF"), inflammatory ("iCAF"), or antigen presenting ("apCAF") fibroblasts. Interestingly, they found that two subpopulations of myCAFs, ecm-myCAFs, associated with extracellular matrix machinery, and TGFβ-myCAFs, associated with TGFβ signaling, were negatively correlated with CD8+ T cells and positively correlated with CTLA4+CD4+ T cells in a different cohort of seven breast cancer patients [77]. To investigate whether these fibroblasts were specifically associated with immunotherapy resistance, they performed GSEA using their scRNA-seq derived ecm-myCAF, TGFβ-myCAFs, and iCAF signatures on bulk RNA-seq from immunotherapy treated melanoma and lung cancer samples[77]. From this, they found that both ecm-myCAF and TGFβ-myCAF signatures were significantly associated with non-responders, while only one of the four iCAF signatures (wound-iCAFs, named due to their wound-healing associated gene expression profiles) had a significant association with poor patient responses[77]. These are still primarily correlative observations, but this work provides a large amount of data-driven support for fibroblasts as a key player in the tumor microenvironment across multiple cancer types, and suggests a convincing stratification scheme for immunotherapy responses that could facilitate improved clinical trial outcomes for the recently attempted applications of immunotherapy in triple-negative breast cancer[68,77].

1.4 Conclusions and future directions

Taken together, the studies reviewed in this manuscript have highlighted cellular plasticity, connected tumor subtypes to their metabolic and cellular microenvironment, and revealed novel cell states and markers. ScRNA-seq has shown that cells in the breast are even more dynamic than we expected, with states reflective of both their current and past environments. It has also become clear that bulk transcriptional profiles cannot fully characterize intratumoral heterogeneity and a reassessment of how we stratify patients may be warranted. In particular, scRNA-seq facilitates the development of prognostics using rare tumor cells with metastatic risk markers and profiles identified from the microenvironment, both of which require the separation of individual cells from the bulk tumor. In addition to supporting the use of scRNA-seq in stratifying patients directly, the studies covered in this review have suggested that spatial elements, such as hypoxic tumor regions, could be drivers of pro- or anti-tumoral microenvironmental behaviors and future works in breast cancer should consider using spatial transcriptomics to test these hypotheses.

ScRNA-seq has already shown great promise in this field, but we should also discuss its limitations. As an example, scRNA-seq may be ill-equipped to fully resolve progenitor cell debates, especially ones regarding the MaSC. MaSC-mimicking stromal contaminants can be sorted along with epithelial cells using standard epithelial sorting strategies in the mammary gland (e.g. $Epcam^{+}$, $Cd24^{mid/hi}$ $CD29^{hi/lo}$) and these accidentally captured stromal cells can co-express gene markers of basal cells (e.g. *Acta2*, *Vim*) and MaSC-associated markers (e.g. *Procr*). Each scRNA-seq dataset has natural variation in the capture of these cell types and paper authors have different interpretations of the captured populations; unfortunately, without functional follow up, it is impossible to know whether these “epithelial” or “stomal” labels are accurate. Further, pseudotime analysis tools are equipped to suggest

transcriptional gradients from scRNA-seq data but have no way of contrasting true lineage relationships with other sources of gene expression changes, so they are always open for reinterpretation. In fact, most computational tools in scRNA-seq give results that must be carefully contextualized and can be sensitive to noise. This was specifically mentioned in the case of breast cancer subtyping, where the tools expect "complete" gene expression profiles to stratify subtypes but scRNA-seq datasets are unable to provide this. Because of these potential issues, it is important to cater scRNA-seq analyses to specific questions and keep hypotheses grounded in functional data.

In our subsequent chapters, included in chronological order, we will see how scRNA-seq analysis can be catered to breast-related questions, and how these analyses led to better understandings of the breast in homeostasis, breast cancer metastasis, and the responses of a non-breast resident microenvironment to breast cancer cells. While we represent only a small subset of the available literature mentioned in this review, our work similarly coveys a recurrent theme of heterogeneity and plasticity, but optimistically suggests that this heterogeneity is bounded and can be generalized, even in seemingly stochastic processes like breast cancer metastasis. We hope this work builds a foundation that allows us to connect each source of breast cancer heterogeneity with the others to form a systems-level understanding of what drives breast cancer and its often lethal metastases.

CHAPTER 2: Identification of conserved gene expression changes in the human breast epithelial hierarchy

2.1 Introduction

Breast cancer arises from the breast epithelium, which forms a ductal network embedded into an adipose tissue that connects the nipple through collecting ducts to an intricate system of lobules, which are the milk producing structures during pregnancy and lactation. Throughout the duct and lobular system, the breast epithelium is composed of two known cell types, an inner layer of secretory luminal cells and an outer layer of basal/myoepithelial cells. A series of recent reports have indicated that further heterogeneity exists within these two cell layers in mice[1]. Two landmark papers published in 2006 identified a functionally distinct subpopulation of basal epithelial cells that harbors stem cell capacity and is capable of reconstituting a fully developed mammary epithelial network when transplanted into the cleared mammary fat pads of mice, referred to as a mammary stem cell (MaSC) [78,79]. Moreover, a subpopulation of luminal progenitor cells identified by high expression of KIT as well as a subpopulation of mature luminal cells have been identified using flow cytometry (FACS) isolation strategies[80,81]. It remains to be determined if other distinct cell types exist within the breast epithelium and how these different epithelial populations relate to the well-characterized heterogeneity of breast cancer.

The goal of the present study is to generate a molecular census of cell types and states within the human breast epithelium using unbiased scRNA-seq. Focusing on the breast

epithelium, our work provides a critical first impetus toward generating large-scale single cell atlases of the tissues comprising the human body as part of the international human cell atlas initiative. This molecular census can shed light on lineage relationships and differentiation trajectories in the human system and how it relates to breast cancer. To that end, we also propose a novel heuristic procedure to optimize the identification of conserved differentiation-associated genes for pseudotime inference from our scRNA-seq data. This procedure allowed us to generate a data-driven lineage trajectory of the human breast epithelium in homeostasis which will serve as a valuable resource to understand how the system changes during early tumorigenesis and tumor progression.

2.2 Results

2.2.1 scRNA-seq of cell types and states in the breast epithelium

To investigate the cell types and states present in the human breast epithelium, we utilized a droplet-mediated scRNA-seq platform (10X Genomics Chromium)[23] on reduction mammoplasty samples from four nulliparous women, who were chosen to reduce the variability associated with pregnancy-related changes of the breast. We isolated both luminal and basal cells together ($EpCAM^{+}/CD49f^{hi/lo}$) by flow cytometry and sequenced them, averaging 5000 cells per sample (**Fig 2.1A**). We sequenced a total of 24,646 cells from four individuals (Ind4-7) at an average ~60,000 reads per cell. Cells were subsequently filtered to remove any cells with low gene detection (<500 genes) and high mitochondrial gene coverage (>10%).

We determined cell types by first analyzing a single individual (Ind4) and identified three main epithelial cell types, namely Basal (*KRT14+*), Luminal 1 (L1; *KRT18+/SLPI+*) and

Luminal 2 (L2; *KRT18+/ANKRD30A+*), and what we suspect are unintentionally captured stromal and endothelial cells which we refer to as Unassigned (X; *VIM+/ESAM+*) (**Fig 2.1B**).

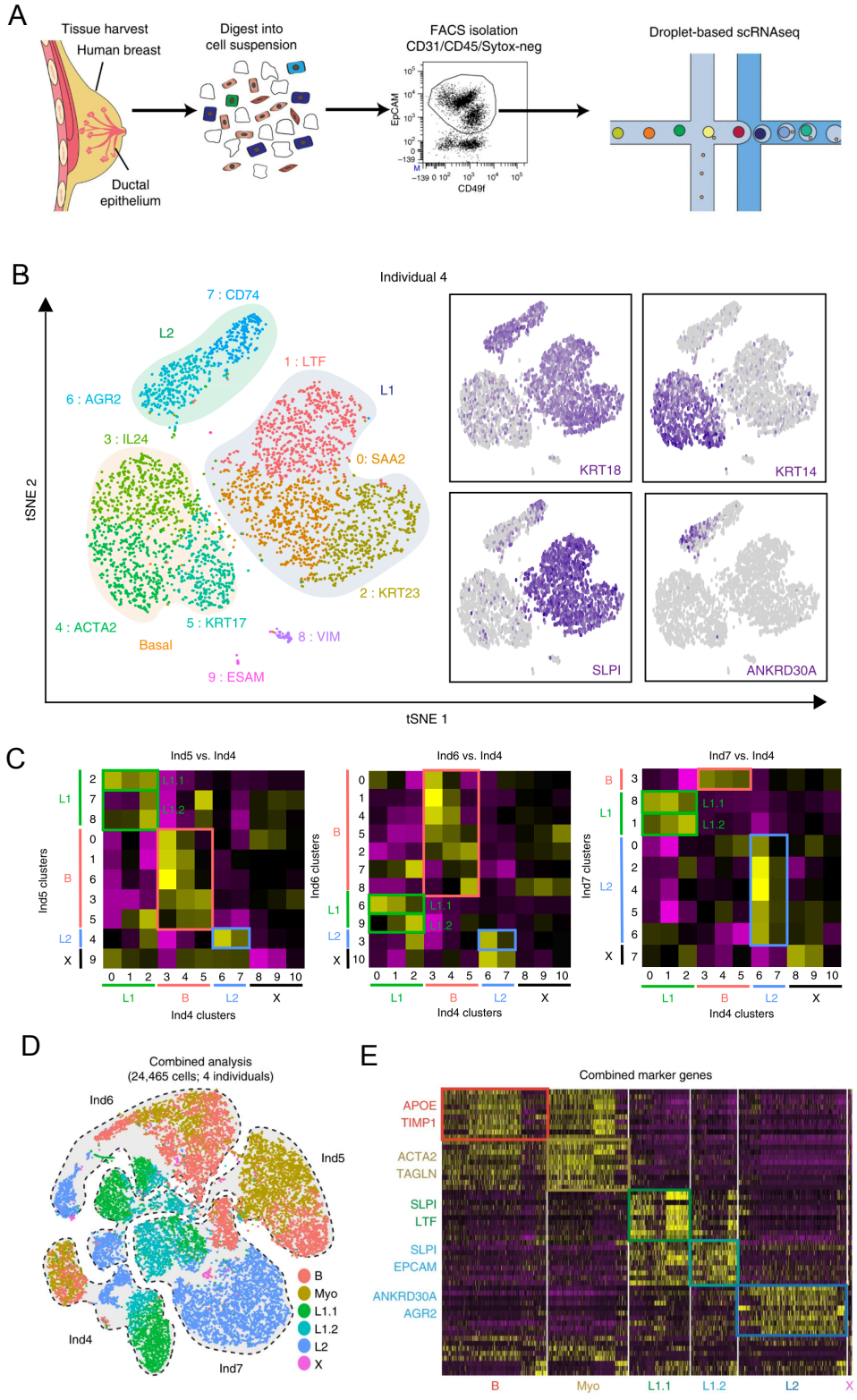

Figure 2.1 scRNA-seq reveals conserved cell types and states in the breast epithelium (A) Overview for droplet-enabled scRNA-seq approach as described above; basal and luminal epithelial cells were sorted together and subjected to combined scRNA-seq analysis using the droplet-based scRNA-seq. (B) Data from individual four was analyzed using Seurat and the distinct clusters (0–10) are displayed in tSNE projection with selected marker gene for each cluster, and main epithelial cell types (Basal, L1, L2) are outlined. Feature plots of characteristic markers for the three main cell types are shown on the right showing expression levels as gradient of purple. (C) Heatmaps showing gene scoring results using marker genes for Ind4 clusters (0–10; on bottom of heatmap) in all clusters from Ind5, Ind6, and Ind7. (D) Combined tSNE projection of all individual datasets (outlined) is shown including the cell state identity marked by different colors. (C) Heatmap showing the expression pattern of the top ten markers per cell state with selected markers indicated (yellow = high expression; purple = low expression). Full figure reprinted and adapted with permission from Nguyen & Pervolarakis et al, 2018, *Nat Commun*.

Generally, it appears that Luminal 1 represents the secretory, milk producing cells of the breast (*LTF+/SAA2+*) and Luminal 2 represents the hormone responsive population (*AGR2+*). We compared these Ind4 derived clusters to clusters identified in each of our other three individuals using gene scoring, which revealed that Luminal 1 could be further broken into two generalizable states that we refer to as Luminal 1.1 (L1.1) and Luminal 1.2 (L1.2). Further, we separated basal cells into two states we refer to as Basal (B) and Myoepithelial (Myo) based on the increased expression of genes associated with contractile, myoepithelial cell function (e.g. *ACTA2, TAGLN, KRT14*) in a subset of the basal population (**Fig 2.1C**). All four individuals were then combined into a single dimensionality reduction with cells labeled by the conserved epithelial states (**Fig 2.1D**) and common marker genes for each cell state were identified using the Wilcoxon rank sum test (**Fig 2.1E**).

We then investigated the spatial localization of these epithelial cell types using immunofluorescent (IF) analyses. KRT14 expression is a standard marker for basal cells, and our differential gene expression analysis confirmed that *KRT14* is predominantly expressed within basal cells. However, it exhibited surprising variability across basal cell populations in the scRNA-seq data with particularly high expression in the Myoepithelial cell state. IF analysis for KRT14 confirmed this, and revealed that KRT14 high cells localized to the basal cell layer within ductal regions, while lobular basal cells generally displayed lower and more

variable staining for KRT14 (**Fig 2.2A**). The scRNA-seq analyses also revealed that the luminal compartment harbors two discrete epithelial cell types (L1, L2). To determine if L1 and L2 correspond to ductal and lobular anatomical location within the tissue, we stained for specific markers for L1 (SLPI) and L2 (ANKRD30A). Interestingly, these analyses showed that both L1 and L2 are located next to each other within both ducts and lobules (**Fig 2.2B**) with no apparent anatomical skewing. Thus, the Myoepithelial cell state may have transcriptome differences to the Basal state driven by its ductal localization, but Luminal 1 and Luminal 2 transcriptome differences are likely to arise from a non-anatomical source.

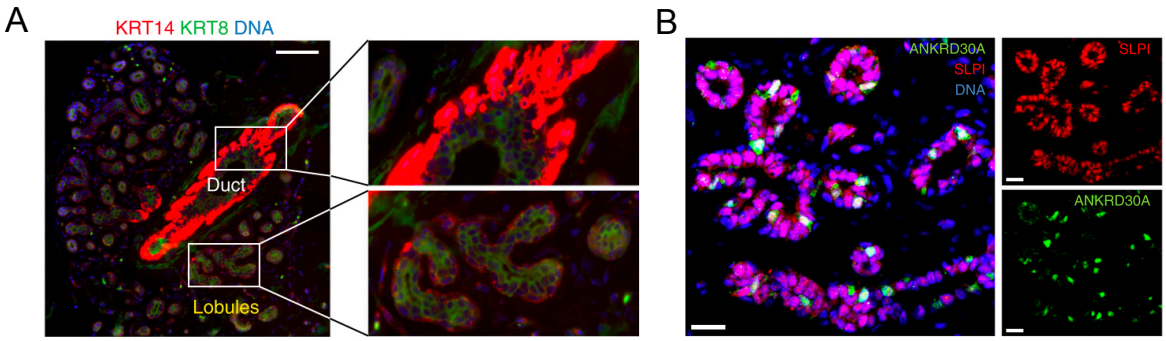

Figure 2.2: Spatial integration of cell types and states in the breast epithelium. (A) KRT14 and KRT8 double immunostaining revealed highest expression of KRT14 in ductal basal cells, while lobular basal cells show more diverse KRT14 positivity. Scale bar =75 µm. (B) Immunofluorescence analysis of NY-BR-1 protein (ANKRD30A) expression (green) in combination with basal marker SLPI (red) and DNA stain using DAPI (blue) within tissue sections from primary human reduction mammoplasty samples revealed that NY-BR-1 (ANKRD30A) and SLPI are markers for distinct luminal subpopulations. Scale bar = 25 µm. Full figure reprinted and adapted with permission from Nguyen & Pervolarakis et al, 2018, *Nat Commun*.

2.2.2 Reconstruction of the breast epithelial hierarchy in individual patients

To understand how these observed cell types and states are related to each other, we next reconstructed differentiation trajectories by pseudotemporal ordering of single cells using Monocle2 [82] in each individual. Our ordering genes for every patient were the top 20 marker genes for each of the conserved cell type and states for that individual, which

resulted in a relatively consistent trajectory, where the Basal and Myoepithelial lineage appear on one branch, and luminal cells have two branches with either Luminal 1 or Luminal 2 cells (**Fig 2.3**). However, Individual 5 shows a very different trajectory which has Myoepithelial cells branching into one Basal state and one single, tangled luminal state (**Fig 2.3**). Given that myoepithelial cells are expected to be a more differentiated basal cell state[83] and that we have seen a clear separation of luminal states 1 and 2 in other clustering and dimensionality reduction methods, this is most likely a cell type capture artifact. This patient has far more cells from the basal lineage than the luminal lineage and therefore the algorithm is picking up on minute basal cell state differences while missing the major distinctions between luminal states. To generate a combined trajectory and ensure that our hypothesized lineage relationships are indeed conserved, we developed a novel heuristic procedure to generalize a list of ordering genes

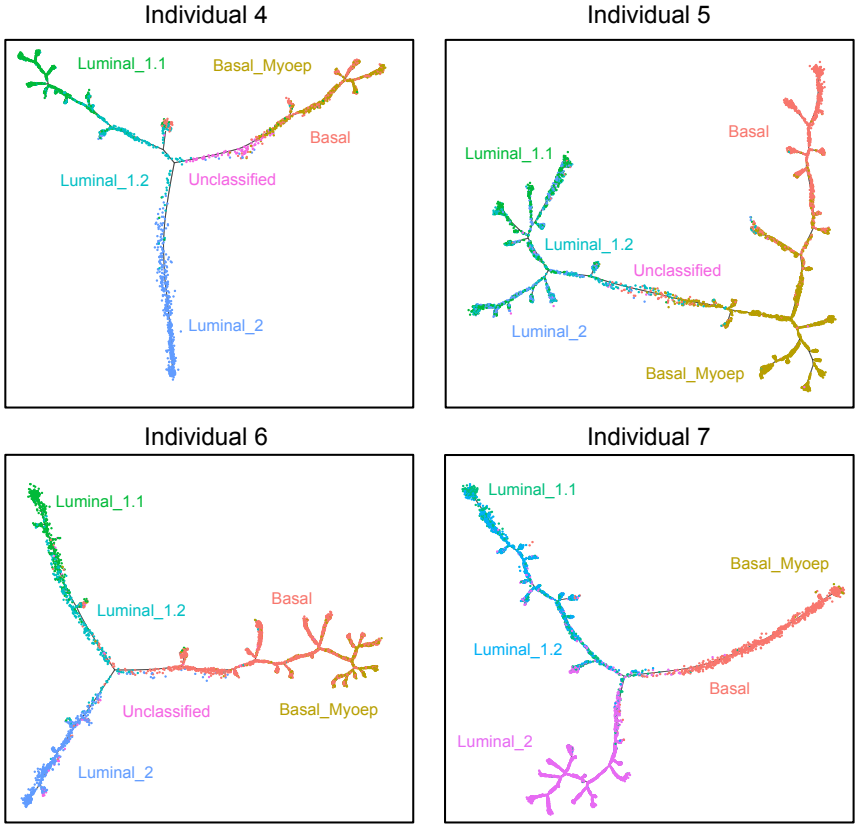

Figure 2.3: Reconstruction of the breast epithelial hierarchy in individual patients. Monocle2 generated pseudotime trajectories for each individual ordered based on the top 20 marker genes for their labeled cell sates. Data is previously unpublished.

across patients that will be less sensitive to cell type capture and individual gene expression differences (**Appendix A**).

2.2.3 Steps in the identification of a maximally conserved gene list for breast epithelial differentiation

Before discussing our heuristic, it is important to discuss our chosen pseudotemporal analysis software, Monocle2, in more detail[41] Monocle2 takes as input a count matrix and a user-defined gene list of "differentiation" associated genes to use for ordering. Its underlying algorithm, DDRTree, uses reversed graph embedding, which seeks to simultaneously map a high dimensional dataset to a low dimensional space and learn a fundamental graph structure which allows for branch points in the data to separate transcriptionally distinct cell types or states. The reversed graph embedding process is computationally intensive, so it requires that an scRNA-seq dataset be subset to a minimally noisy, maximally informative gene list, which it uses for defining the graph and dimensionality reduction. Monocle2 also offers an unbiased feature selection heuristic, dpFeature, which combines multiple dimensionality reduction and cluster marker identification methods to optimize an input gene list for trajectory inference. However, when dimensionality reduction and clustering are driven by individual differences as they are in our dataset (see **Fig 2.1D**), this method is unable to separate cell state associated gene expression changes from individual driven gene expression changes. We therefore developed the heuristic pipeline that orders each individual based on their own cell state markers, and then identifies gene modules that are internally correlated across all four individual's Monocle2 trajectories. This forces us to remove genes that have inconsistent co-expression patterns across separate individuals,

which are likely to result in some degree of individual-driven branching in a combined trajectory and are less likely to be biologically meaningful pathway changes. Below, we explain each step in our procedure, with its associated rationale and example computations where we felt it informative.

1. For each individual or biological replicate, create a Monocle2 trajectory using an individual-specific marker gene list for the conserved cell states as input.
2. Identify genes differentially expressed across Monocle2-defined States for each individual's Monocle2 trajectory. Continue analysis with genes found to be significantly differentially expressed (FDR<0.05) between States in all individuals.

   Rationale: States are determined by branch points, so the cells found together on a terminal branch or between two branch points constitute a cluster. We are only interested in genes that vary between branches, and further, only interested in genes that vary between branches in all individuals. This is because we expect that varied genes are in some way important to determining transition points and driving cell fates.
3. Estimate the expression of each of the shared differentially expressed genes from step two in the States for every individual trajectory. Note that this analysis was originally performed with Monocle 2.2.0 default parameters which 'over-branch' compared to later default parameter sets, and the following steps will work best if each individual's data is broken into a large number of small groupings (i.e. over-clustering the data is likely optimal to avoid misleading averaging or Simpson's paradox).

   Rationale: Single-cell data has a very high false drop-out rate (zeros that should not be zero) and therefore, looking at average expression of genes in cell

subpopulations can vastly reduce the noise levels of our data. Monocle2 defines States based on the transcriptome similarity of cells; because of this, we hypothesize that we can treat all the cells in a State as a cellular subtype and use pseudo-bulk data (generated through gene expression averaging across cells in a State) for our downstream analysis.

Example computation: If we have 63 branches in our trajectory and 3,000 genes kept from our previous differential expression analysis, we will generate a 3000x63 matrix where each row is a gene, each column is a State, and every position has the average expression of the listed gene in the listed State. All individuals will have the same number of rows, but a different number of columns based on how many branches were found for that individual's trajectory.

4. Generate correlation matrices for each individual across all genes.

   Rationale: We want to find sets of genes that are always expressed together both within and between individuals. We predict that identifying correlated sets of genes will provide another mechanism for reducing dropout associated noise since we can rely on multiple genes to represent the same transition or terminal state.

5. Average all individual correlation matrices from step four into a single "average correlation" matrix.

   Rationale: We want to ensure that the genes we choose will behave the same way on average across all individuals and therefore, we want to work with the average correlation value rather than the four correlation matrices separately.

   Example Computation: Each individual will have a correlation between genes A and B. Say that the correlations are the following for each individual: (0.8,-0.4,0.5,-0.5).

The corresponding position in our "average correlation" matrix will be (0.8+−0.4+0.5+−0.5)/4= 0.1. This process is intended to diminish the correlation value for gene sets that show inconsistencies across individuals because they are less likely to be meaningful modules.

6. Reduce the number of genes in our average correlation matrix to only genes highly correlated (in our data, we used a Pearson correlation >0.8) with at least one other gene.

   Rationale: This step is a logical addition to step four, which assumes that gene modules will help eliminate dropout noise. We do not want to keep genes that have no close connection to another gene since their expression changes are more likely to be driven by technical artifacts than a biologically meaningful pathway.

7. Further filter the genes in the average correlation matrix to only keep genes that were cell state markers from our individual trajectory inferences in step one. This is the final gene list which should be used as input for a combined Monocle2 trajectory.

   Rationale: Our primary goal for this ordering is the same as it was for individuals; take cell state clusters and place them into a differentiation trajectory. Therefore, we want to stick to cluster markers and remove genes that drive differences across cells not related to the major cell states defined in our original clustering, since these are often meaningful, but not interesting (e.g. proliferation status or sorting stress).

8. Cluster average correlation matrix into gene modules and assess gene module correlation consistencies across individuals both before and after a combined ordering. This should look similar to **Fig 2.4**.

Rationale: This step is not necessary for gene identification or generating a combined trajectory, but it provides a sanity check that the process gave a reasonable gene set. For example, we should find that many of the cell state marker genes made it past filtering (sanity check one), genes for different cell states predominantly appeared in different clusters (sanity check two), and that the genes in a combined trajectory have a similar correlation structure to the average of the individuals to demonstrate that cell state differences are stronger than any individual differences (sanity check three).

2.2.4 Combined pseudotemporal analysis supports smooth transitions between breast epithelial cell types

We utilized the conserved gene list identified in the previous section to generate a pseudotemporal trajectory including cells from all four individuals. This revealed three major branches (Basal, Luminal 1, and Luminal 2) where Basal and Myoepithelial cells share a branch with a few Luminal 1.2 cells, and Luminal 1.2 cells span a branch between Luminal 1.1 or Luminal 2 states (**Fig 2.5A**). We found that while individuals are slightly separated based on their contributed cell states for each epithelial cell type, the trajectory was reasonably well-covered in each patient and clearly separated cell states more than

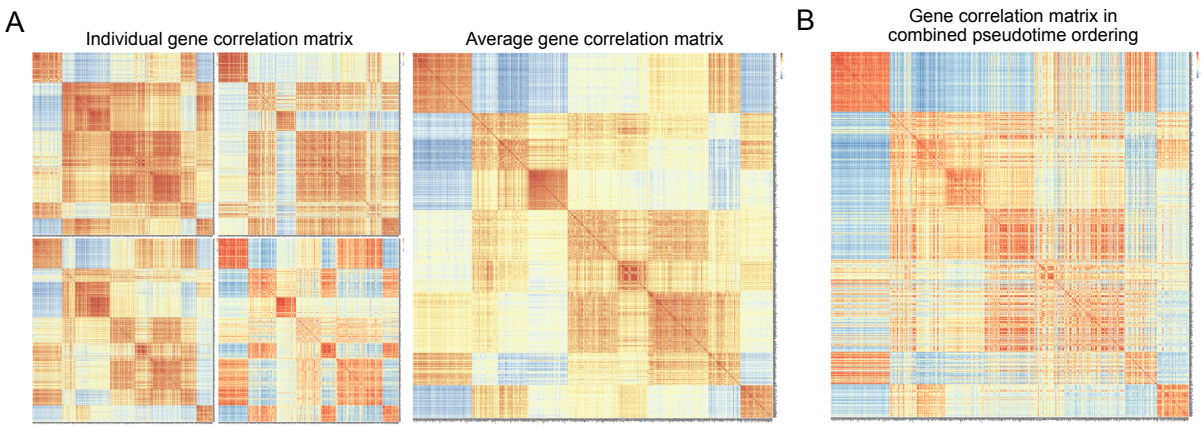

Figure 2.4: Identification of a maximally conserved gene list for breast epithelial differentiation. (A) Correlation matrices for all genes identified in our heuristic procedure in individual trajectories (left) and the average of those matrices (right), ordered based on the clustering results from the average correlation matrix (EM algorithm, MClust package). The final ordering genes for a combined pseudotime trajectory would be additionally filtered from the average gene correlation matrix to only include marker genes of our cell states of interest. (B) The gene correlation matrix from a combined pseudotime trajectory that includes all four individuals and was ordered using the filtered cell state markers as described in our heuristic. Genes shown are the same genes and cluster orders as (A) to facilitate a direct comparison. Data is previously unpublished.

individuals (**Fig 2.5B**). It is notable that in this analysis, we did not remove the small number of suspected stromal cells, and we see that this stromal population is enriched between basal and luminal cell types. It has been hypothesized that bipotent basal cells (MaSCs) resemble mesenchymal-like stromal cells[1,4], so while we still do not believe these populations are of epithelial origin, our proposed trajectory supports the idea of a mesenchymal gene expression profile acting as an intermediate between basal and luminal gene expression profiles. Since Monocle2 requires manual identification of a starting point for its pseudotime calculation, we also set this to be the small branch between the basal and luminal populations since the MaSC is hypothesized to be the most stem-like population in the breast.

We then revisited the cell type markers we expected to drive cell type and state differences and investigated how their expression changed over pseudotime (**Fig 2.5C**). This analysis agreed well with our expectation since the marker genes remained associated with

their labeled branches and increased as cell state maturity was expected to increase, though it was surprising that hypothesized markers of luminal progenitor cells (*ELF5*/*KIT*), were

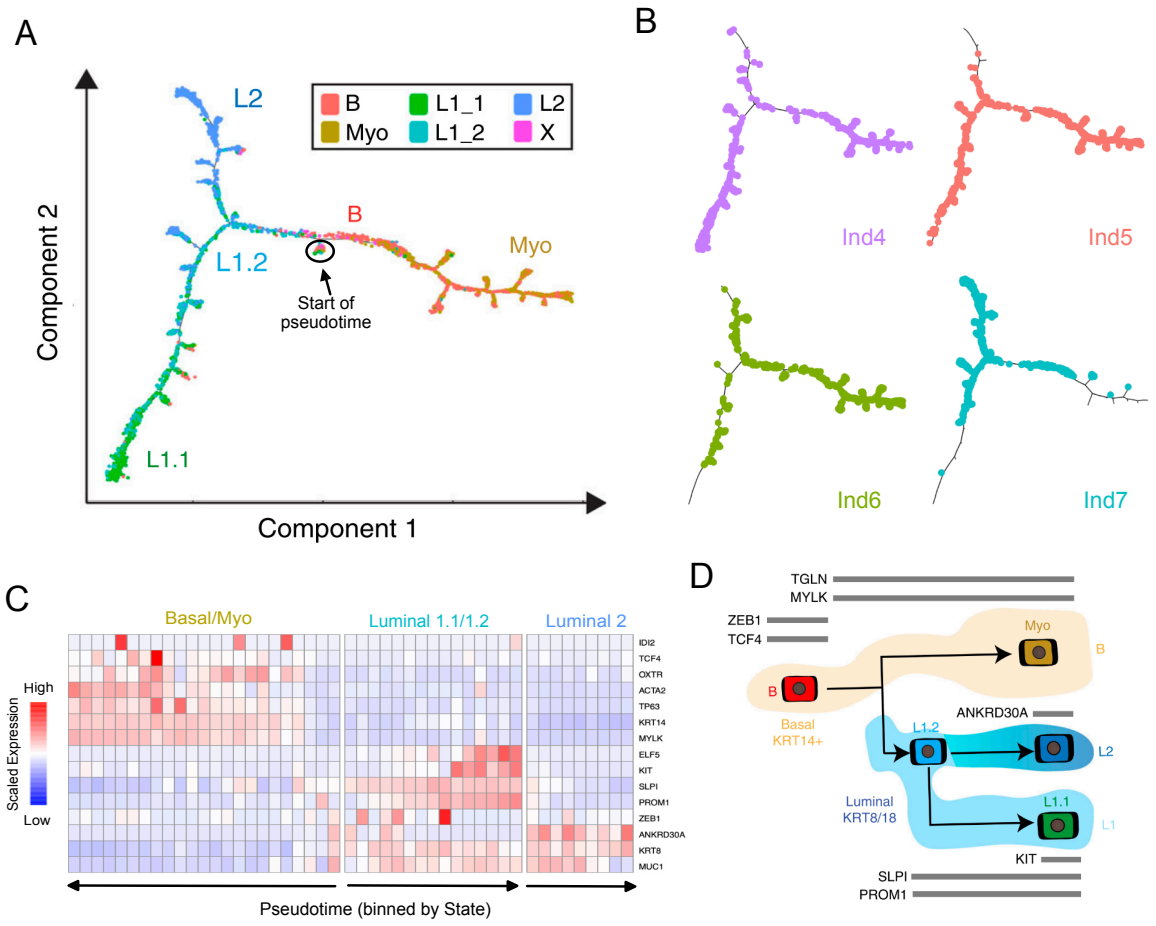

Figure 2.5: Combined pseudotemporal analysis supports smooth transitions between breast epithelial cell types. (A) Monocle-generated pseudotemporal trajectory of a randomly subsampled population of cells (n =1000) from each of our four individuals analyzed using droplet-mediated scRNA-seq is shown colored by cell state designation. (B) Trajectories show cells split and colored by individual. (C) Heatmap shows the scaled, average expression of key marker genes of basal and luminal cell types in each State, ordered by their average pseudotime value, and split across each of the major three visual branches. The start of pseudotime was set to be the small branch (State) between B and L1.2 labels. (D) Proposed model summarizing the lineage hierarchies within the breast epithelium based on one continuous differentiation trajectory from basal stem cells to three distinct differentiated cell types with overlaid marker genes of interest shown (black on gray bars). Subfigures (A) and (D) were reprinted and adapted with permission from Nguyen & Pervolarakis et al, 2018, *Nat Commun*. Subfigures (B) and (C) are previously unpublished.

found at the terminal end of the Luminal 1 branch rather than the starting point [80,84] (**Fig 2.5B,C**). There are a few possible interpretations of this result. One possibility is that both Luminal 1.2 and Luminal 1.1 cells are progenitors, and that luminal cells can take multiple

paths to reach differentiated, secretory (L1) and hormone responsive (L2) populations. Another option is that Luminal 1.1 cells have progenitor capacities under perturbed conditions but that they do not perform this function in tissue homeostasis. Without extensive follow up experiments, we cannot distinguish these cases, and therefore present our model as a unidirectional trajectory with the Luminal 1.1 cells as an endpoint (**Fig 2.5D**). Taken together, our computationally-derived trajectory of the breast epithelial hierarchy supports the existence of a shared progenitor for basal and luminal cells, as well as a bipotent, intermediate luminal progenitor (Luminal 1.2) which can generate both secretory and hormone responsive populations.

2.3 Discussion

In this work, we have described a heuristic for the identification of a conserved gene list to be used in pseudotemporal reconstruction of scRNA-seq data and demonstrated its utility in a novel scRNA-seq dataset of healthy human breast epithelial cells. We identified three major cell types (Basal, Luminal 1, Luminal 2), and five conserved cell states in the human breast epithelium (Basal, Myoepithelial, Luminal 1.1, Luminal 1.2, Luminal 2) across four reduction mammoplasty samples. We next generated a lineage trajectory for these healthy breast epithelial populations, which supported a mesenchymal-like MaSC cell type between the basal and luminal lineages in the adult breast, as well as smooth transitions between secretory (Luminal 1) and hormone responsive (Luminal 2) luminal cells.

While it is tempting to treat our scRNA-seq derived trajectory as solid evidence of *in vivo* lineage relationships unmarred by the perturbation effects one might find using mouse models, there are still many confounders in Monocle2. One major issue is that this algorithm

cannot on its own differentiate between lineage and other causes of phenotypic gradients. Spatial relationships are the most likely source of non-lineage gradients, and sometimes these trajectories are interpreted as “pseudospace” rather than “pseudotime”. In the breast epithelium, it is possible that spatial gradients are contributing to our trajectory, but it must be more complicated than simple ductal/lobular anatomy. Specifically, we saw that Luminal 1 and Luminal 2 cell states comingled in ducts and lobules in the tissue while they cleanly branched in our trajectory, suggesting that this branch point is driven by a non-anatomical source. Another potential concern is the inability of Monocle2 to break its trajectory, which could force basal and luminal populations together when they do not have any meaningful lineage connection in the adult breast. We cannot fully eliminate this possibility, but we can note Monocle2 tends to have a large amount of sparsity between truly unrelated lineages in a trajectory. In our combined trajectory, almost no gap is present. Further the cells that do reside between basal and luminal cells in our trajectory are similar to cells that have been shown to have bipotent basal-luminal progenitor capacities both *in vivo* and *in vitro*, which makes our trajectory consistent with demonstrated biology [1,4].

Our lineage reconstruction of the breast epithelium from single-cell transcriptomic profiles has provided support for a hierarchy involving multiple shared progenitor populations, but further functional experiments will be necessary to prove or disprove each transition point. To facilitate these experiments, we also identified a conserved list of genes that drive the breast epithelial cell state transitions we discussed and provide this list to the community as part of our study. These genes now serve as good candidates for perturbation in mechanistic investigations of the human epithelial hierarchy. Further, the clustering of our gene modules can assist in the identification of intrinsic or extrinsic factors that drive these

co-regulated expression changes across the trajectory. This could help to deconvolve spatial, lineage, and other signaling gradient effects. Overall, this dataset and analysis have brought us a few steps closer to understanding how the incredibly plastic breast epithelium functions in homeostasis and will allow us to better identify any lineage-related transitions that occur during breast cancer tumorigenesis and metastatic adaptation.

2.4 Materials & methods

Origin of tissue samples.

Anonymous reduction mammoplasty samples were acquired from NCI Cooperative Human Tissue Network (CHTN) and from Department of Surgery, Feinberg School of Medicine, Northwestern University. Other investigators may have received specimens from the same tissue specimens obtained through NCI CHTN. Specimens were anonymized then collected and distributed by CHTN, specimens are covered under collection/distribution of tis- sues under consent or waiver of consent. Samples were washed in PBS (Corning 21-031-CV) and mechanically dissociated using a razor blade. Dissociated samples were digested overnight in DMEM (Corning 10-013-CV) with Collagenase Type I, 2 mg/mL (Life Technologies 17100-017). Viable organoids were separated using differential centrifugation and viably frozen in 50% FBS (Omega Scientific FB-12), 40% DMEM, and 10% DMSO (Sigma-Aldrich D8418) by volume.

Single-cell RNA sequencing.

Viable organoids were thawed and washed using DMEM, and digested with 0.05% trypsin (Corning 25-052-CI) containing DNase (Sigma Aldrich D4263-5VL) to generate single cell

suspension. Cells were stained for FACS using fluorescently labeled antibodies for CD31 (eBiosciences 48-0319- 42), CD45 (eBiosciences 48-9459-42), EpCAM (eBiosciences 50-9326-42), CD49f (eBiosciences 12-0495-82), and SytoxBlue (Life Technologies S34857). We only proceeded with samples showing at least 80% viability as measured using SytoxBlue in FACS.

For droplet-enabled scRNA-seq, flow cytometry sorted cells were washed in PBS with 0.04% BSA and resuspended at a concentration of ~1000 cells/µl. Library generation for 10× Genomics v1 chemistry was performed following the Chromium Single Cell 3' Reagents Kits User Guide: CG00026 Rev B. Library generation for 10× Genomics v2 chemistry were performed following the Chromium Single Cell 3' Reagents Kits v2 User Guide: CG00052 Rev B. Quantification of cDNA libraries was performed using Qubit dsDNA HS Assay Kit (Life Technologies Q32851) and high-sensitivity DNA chips (Agilent. 5067- 4626). Quantification of library construction was performed using KAPA qPCR (Kapa Biosystems KK4824). For droplet-enabled scRNA-seq, we used the Illumina HiSeq4000 platform to achieve an average of 50,000 reads per cell.

Processing of scRNA-seq data.

Cluster identification using Seurat. For cluster identification in droplet-enabled scRNA-seq datasets, we utilized the Seurat R package version 2.0.0 [85]. Data was read into R as a counts matrix and transformed into log-space. Due to the difference in gene detection across the two platforms, differences in chemistry for the library prep, as well as sequencing depth per cell, a minimum cutoff of 500 and a maximum cut-off of 6000 genes per cell for this dataset

was used. In addition, cells with a percentage of total reads that aligned to the mitochondrial genome (referred to as percent mito) greater than 10% were removed, since increased detection of mitochondrial genes can be associated with cells undergoing stress and cell death. To account for the possibility of individual cell complexity driving cluster separation, we employed Seurat's "RegressOut" function to reduce the contribution of both the number of UMI's and the percent mito. Variable genes were then determined for subsequent PCA for each separate individual. For tSNE projection and clustering analysis, we used the first ten principal components. We used the feature plot function to highlight expression of known marker genes for basal (e.g., KRT5, KRT14) and luminal cells (e.g., KRT8, KRT18) to identify which clusters belonged to which epithelial cell type. The specific markers for each cluster identified by Seurat were determined using the "FindAllMarkers" function.

Cluster comparisons and assignment.

Cluster specific marker genes from the individual library analyses were used as input lists to the previously described gene scoring method (described in more detail below) to compare cluster signatures in a pairwise manner between individuals. To visualize pairwise gene scoring results, we generated heatmaps displaying averaged gene scoring results for each cluster. We overlaid individual-specific cluster designations onto these heatmaps to find which individual clusters best match to each other. Clusters were merged together in the case that multiple clusters scored highly. We performed a separate Seurat analysis using combined basal cells from all four individuals, and then matched clusters using the gene scoring method on a set of genes curated to represent a myoepithelial cell fate[83] to score and classify the clusters as either Basal (B) or Myoepithelial (Myo) cell state.

Gene scoring.

To compare gene signatures and pathways in epithelial sub- populations, we utilized individual gene scores as described previously12. Briefly, each score was generated by calculating total gene expression for each of the analyzed genes and separating them into 25 bins of similar expression. For every gene in each target pathway or signature, 100 "control" genes were selected from its corresponding bin and added to a "control" pathway. The resulting "control" pathway contained an equivalent expression distribution as the target pathway and its average represents an equivalent sampling of 100 pathways of equal size to the target pathway. The expression of genes in the target pathway and the "control" pathways was averaged across each cell to generate a target score (STarget) and control score (SCtrl). The cell's score for the target pathway (SPath) is the difference between the target score and control score: SPath = STarget − SCtrl. To determine statistical significance, we used the unpaired Wilcox test with a 95% confidence interval.

Immunofluorescence analysis.

Tissues were fixed in 4% formaldehyde for 24 h, dehydrated in solutions of increasing concentrations of ethanol, cleared with xylene, and embedded in paraffin. Slides of 10-μm sections were prepared using a Leica SM2010 R Sliding Microtome (Leica Biosystems, Wetzlar, Germany). Slides were heated at 65 °C for 1 h, followed by two 5-min incubations in Histo-Clear (National Diagnostics, Cat. No. HS-200, Atlanta, Georgia, USA) for paraffin removal. Tissues were rehydrated with solutions of decreasing concentrations of ethanol, washed in double-distilled H2O and PBS, and subjected to antigen retrieval using a

microwave pressure cooker with 10mM citric acid buffer (0.05% Tween 20, pH 6.0). Tissues were blocked in blocking solution (0.1% Tween 20 and 10% Goat Serum in PBS) for 20 min at room temperature, incubated with primary antibodies prepared in blocking solution at 4 °C overnight, washed in PBS, incubated with secondary antibodies diluted in PBS for 1 h at room temperature, and washed in PBS. Slides were mounted with VECTASHIELD Antifade Mounting Medium with DAPI (Vector Laboratories, Cat. No. H-1200, Burlingame, California, USA) and micrographs were taken with the BZ-X700 Keyence fluorescent microscope. For quantification of staining (e.g., ZEB1 and KRT14 staining), we manually counted positive cells as signal around nuclei (DAPI) and utilized the BZH Hybrid Cell Count software (Keyence) in at least three different fields of view using a 40× objective in at least two different samples. Primary Antibodies: Estrogen Receptor (ER) rat mAb diluted 1:50 (Cat. No.916201); KRT14 rabbit pAb diluted 1:500 (Cat. No. PRB-155P) (Biolegend, San Diego, CA, USA); SLPI goat pAb diluted 1:200 (R&D Systems, Cat No. AF1274-SP, Minneapolis, MN, USA); KRT18 rabbit pAb diluted 1:500 (Cat. No. GTX112978) (GeneTex, Inc., Irvine, California, USA); NY-BR-1 mouse mAb diluted 1:500 (Cat. No. MS-1932-P0); KRT14 mouse mAb diluted 1:100 (Cat. No. MA511599); and KRT18 mouse mAb diluted 1:100 (Cat. No. MA512104) (Thermo Fisher Scientific Inc., Carlsbad, California, USA). Secondary Antibodies: Donkey anti-mouse Cy5.5-conjugated IgG (Novus Biologicals, Cat. No. NBP1-73774, Littleton, CO, USA); Goat anti-rabbit IgG conjugated with Alexa Fluor 568 and 488 (Cat. No. A21069 & A11034); Goat anti- mouse IgG conjugated with Alexa Fluor 568 and 488 (Cat. No. A11004 & A11001); Goat anti-rat IgG conjugated with Alexa Fluor 488 (Cat. No. A11006); Donkey anti-rabbit FITC-conjugated IgG (Cat. No. A16030); and Donkey anti-goat IgG conjugated to FITC

and Alexa Fluor 568 (Cat. No. A16006 & A11057) (Thermo Fisher Scientific Inc., Carlsbad, California, USA).

Code availability.

Custom scripts are available at:

https://github.com/kessenbrocklab/Nguyen_Pervolarakis_Nat_Comm_2018.

Data availability.

The authors declare that all data supporting the findings of this study are available within the article and its supplementary information files or from the corresponding author upon reasonable request. All scRNA-seq data quantified data matrices along with their associated meta data have been deposited in the GEO database under accession code GSE113197.

Portions of the Introduction, Results (2.2.1, 2.2.4), and Methods in this section were reprinted and adapted with permission from:

Nguyen, Q.H., Pervolarakis, N., Blake, K. et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. Nat Commun 9, 2028 (2018). https://doi.org/10.1038/s41467-018-04334-1

CHAPTER 3: Biomarkers of micrometastasis in triple-negative breast cancer

3.1 Introduction

Despite major advances in the detection and treatment of early stage disease, metastasis remains the cause of nearly all mortality associated with breast cancer[86,87]. Previous work suggests that metastasis can be seeded by rare primary tumor cells with unique biological properties that enable them to surpass each step in the metastatic cascade, though the exact mechanism of metastasis is not yet known in breast cancer and it is still possible that metastasis is a fully random process [88–90]. Although the properties promoting cell motility and migration have been well studied, the mechanisms governing micrometastasis seeding and dormancy in distal tissues are poorly understood. This is in part because metastatic seeding cannot be studied in humans and because it is technically challenging to detect and analyze rare cells at this transient stage in animal models. Further insights into the mechanisms driving the seeding and maintenance of clinically undetectable micrometastases are critical to inspire new strategies for the prevention of metastatic spread and reduction in mortality of patients with breast cancer.

We have developed a robust approach for the capture and analysis of individual cancer cells during the seeding of micrometastasis in human patient-derived xenograft (PDX) models using single-cell RNA sequencing (scRNA-seq) technology. Using this data, we were able to identify novel markers of micrometastatic tumors, and developed a forward selection, logistic regression model to prioritize potential biomarkers for follow up. From this procedure, we found *PHLDA2* and *BHLHE40* were conserved markers of

micrometastasis and primary tumors respectively, with prognostic capabilities for relapse-free survival in breast cancer patients. Further, *in situ* validation of *PHLDA2* identified rare micrometastatic-like cells in the primary tumor, supporting the possibility of non-random metastatic progression in breast cancer.

3.2 Results

3.2.1 Single-cell RNA sequencing of matched primary tumors and micrometastases.

To identify fundamental properties of micrometastasis in breast cancer, we investigated transcriptome programs uniquely expressed by cancer cells during the seeding and establishment of micrometastatic lesions. We utilized three previously established patient-derived xenograft (PDX) models of triple-negative breast cancer (TNBC), HCI001,HCI002, and HCI010, which have been shown to both possess and maintain intra-tumoral heterogeneity in mice.[59,91,92] As is seen in many patients with breast cancer, metastatic progression is slow and sporadic in these models, where most animals display dispersed micrometastases in the lung and lymph nodes and very low metastatic burden at endpoint.[59,91] This enabled us to investigate the transcriptional changes associated with early events in the seeding and establishment of micrometastasis. We utilized a previously developed protocol for the isolation of metastatic cells from PDX models using flow cytometry with human (CD298) and mouse (MHC-I) species-specific antibodies[59], sorted individual cancer cells from the lungs, lymph nodes, and primary tumors from PDX mice into 96-well PCR plates, and performed full-length scRNA-seq using an optimized version of the Smart-seq2 protocol [22] (**Fig 3.1A,B**).

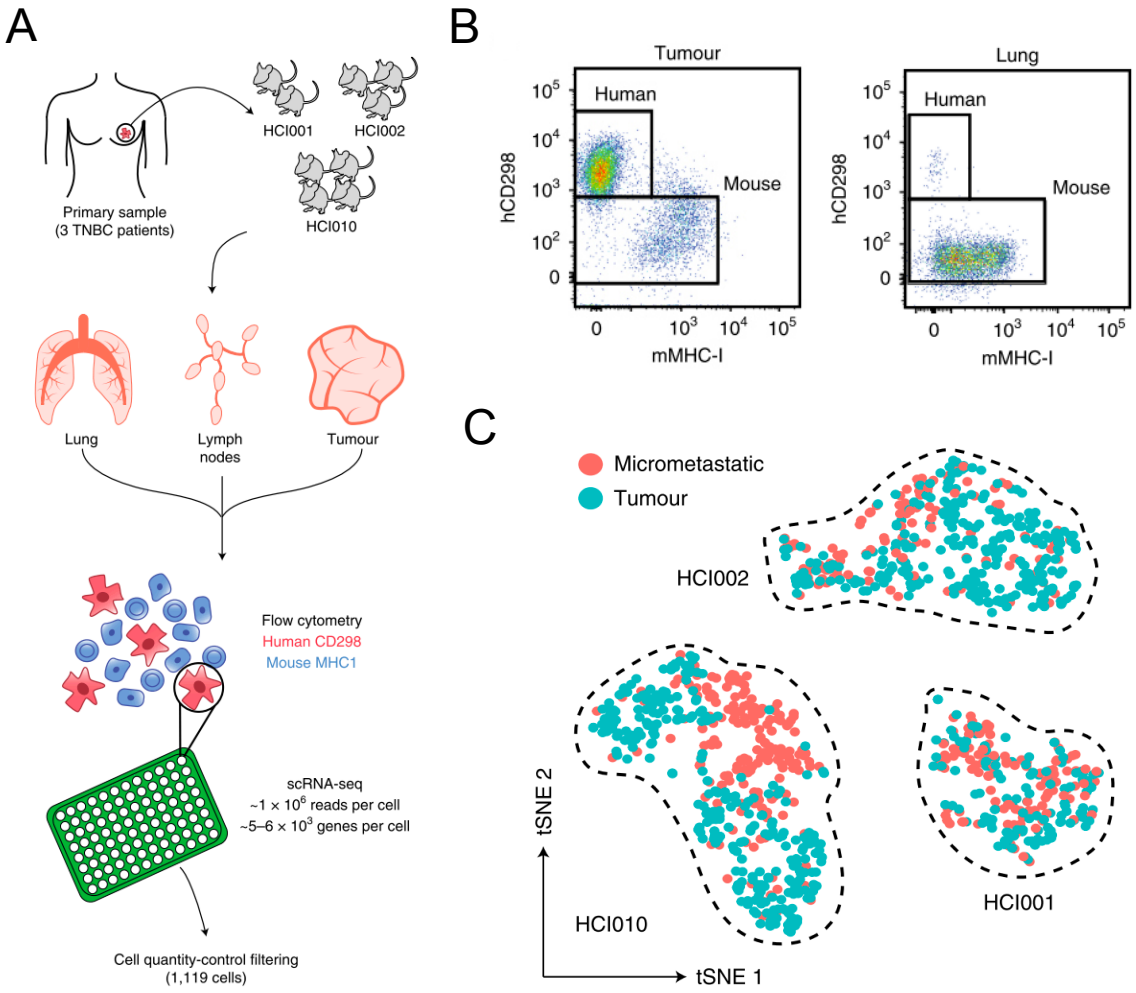

Figure 3.1: Single-cell RNA sequencing of matched primary tumor and micrometastatic cells. (A) Overview of the experimental workflow. The primary tumor, lungs and lymph nodes of each PDX animal were digested to make single-cell suspensions. Single CD298^{+}MHC-I^{-} human tumor cells were isolated by flow cytometry, deposited into individual wells of 96-well plates and single-cell cDNA libraries were prepared using Smart-seq2 chemistry. Matched primary tumor and micrometastatic cells from nine mice and three PDX models (HCI001, HCI002 and HCI010) were analyzed, and 1,119 cells passed quality- control filtering. (B) Left, flow cytometry-based strategy for the isolation of human CD298^{+}MHC-I^{-} cells from micrometastatic (bottom) and primary tumor (top) cells. (C) T-distributed stochastic neighbor embedding (tSNE) plot showing all metastatic and primary tumor cells from the HCI001, HCI002 and HCI010 models. Reprinted and adapted with permission from Davis et al, 2020, *Nat Cell Biol*.

After removing cells with low gene detection (<2500 genes) or a high mitochondrial percentage (>50%), we analyzed 1,119 matched tumor and micrometastatic cells from nine PDX mice across our three PDX models. Dimensionality reduction using t-distributed stochastic neighbor embedding (tSNE)[34] revealed a strong transcriptional separation of

patients, but relatively little separation of tumor and micrometastatic cells within patients (**Fig 1C**). This large amount of inter-tumoral heterogeneity is consistent with findings from prior single-cell studies, and may be due to distinct copy number variations driving global gene expression shifts across patients.[38,55] Thus, these data suggest that investigating intra-tumoral heterogeneity and the differences between tumor and micrometastatic cells requires individual patient analyses.

3.2.2 Transcriptional diversity in micrometastatic and primary tumor cells.

Performing dimensionality reduction and unbiased clustering for each patient separately revealed distinct populations of tumor and micrometastatic cells with biologically meaningful behaviors, including metabolic specializations (e.g. OXPHOS, Glycolysis, and Fatty-acid metabolism), epithelial-to-mesenchymal transition (EMT), and extracellular matrix modulation (ECM) (**Fig 3.2.1 A, B**). This intra-tumoral heterogeneity and the specific biological specializations were well-conserved across patients, xenograft passages, and tissue of origin, though not of equal abundance in each condition (**Fig 3.2.2A-C**). While no clusters were purely of tumor or micrometastatic origin, we could identify skewing towards OXPHOS in micrometastatic cells (A1, B1,C2, C3) as well as an enrichment of EMT or ECM modulation in primary tumor cells for patients HCI002 (B3, B5) and HCI010 (C4) (**Fig 3.2.2A,B**). A requirement to undergo EMT for migration and MET (mesenchymal-to-epithelial transition) for metastatic outgrowth has been characterized[93], though a role of

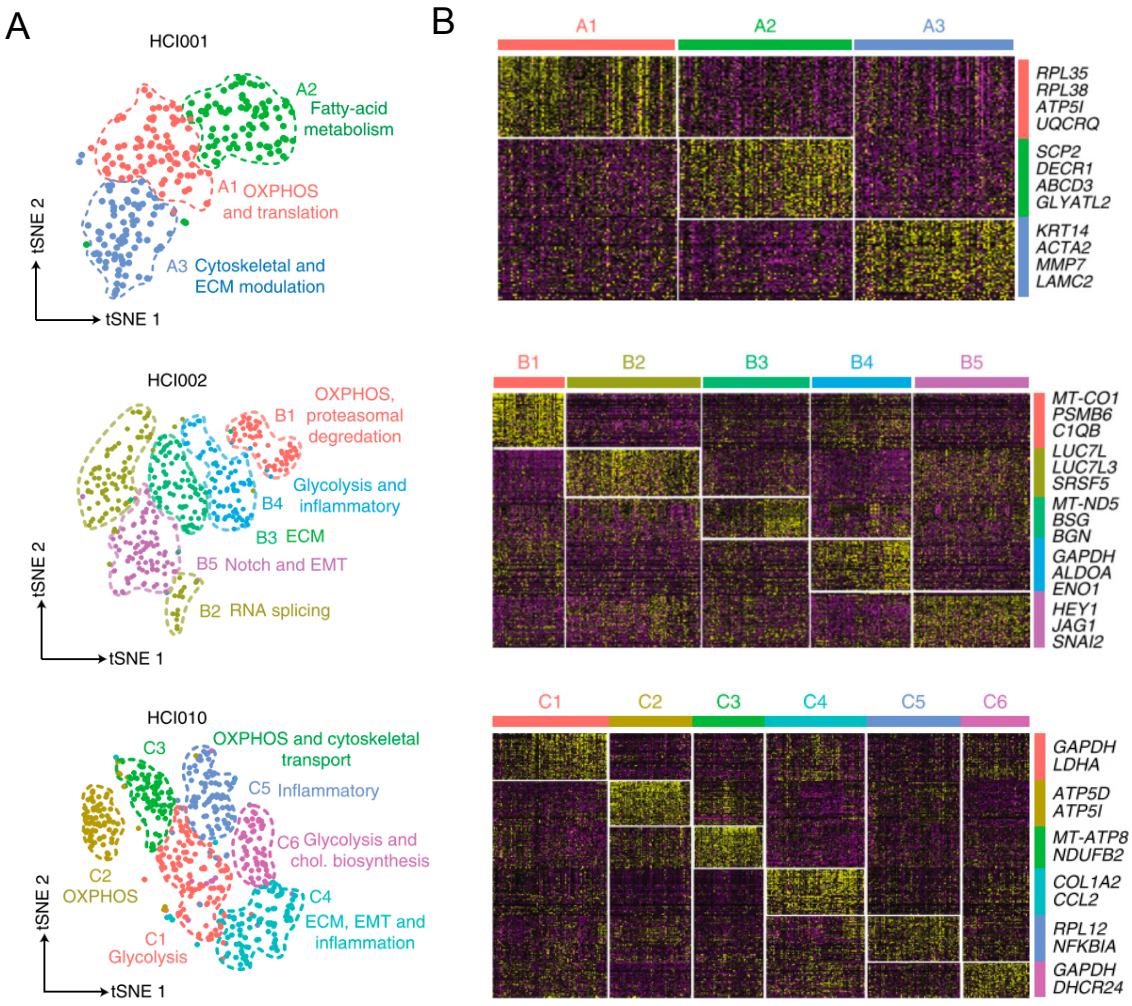

Figure 3.2.1: Transcriptional diversity in micrometastatic and primary tumor cells. (A) Clustering of cells from the HCI001 (top; n = 247 cells), HCI002 (middle; n = 401 cells) and HCI010 (bottom; n = 471 cells) PDX models shown in tSNE plots. The cells are colored according to their cluster identity. The biological features defining each population identified by GO term analysis of marker genes are indicated. Chol. = cholesterol. (B) Heatmaps show the top marker genes in each cluster in (A) based on average log fold enrichment and the Wilcoxon rank sum test ($P < 0.05$). Select marker genes are listed for each cluster on the left. Reprinted and adapted with permission from Davis et al, 2020, *Nat Cell Biol*.

OXPHOS in micrometastasis is novel and it is not yet apparent how these extremely small micrometastatic tumors develop or maintain the level of heterogeneity seen here. Taken as a whole, these data suggest that triple-negative breast cancers are transcriptionally heterogeneous both between and within tumors. However, there also appears to be

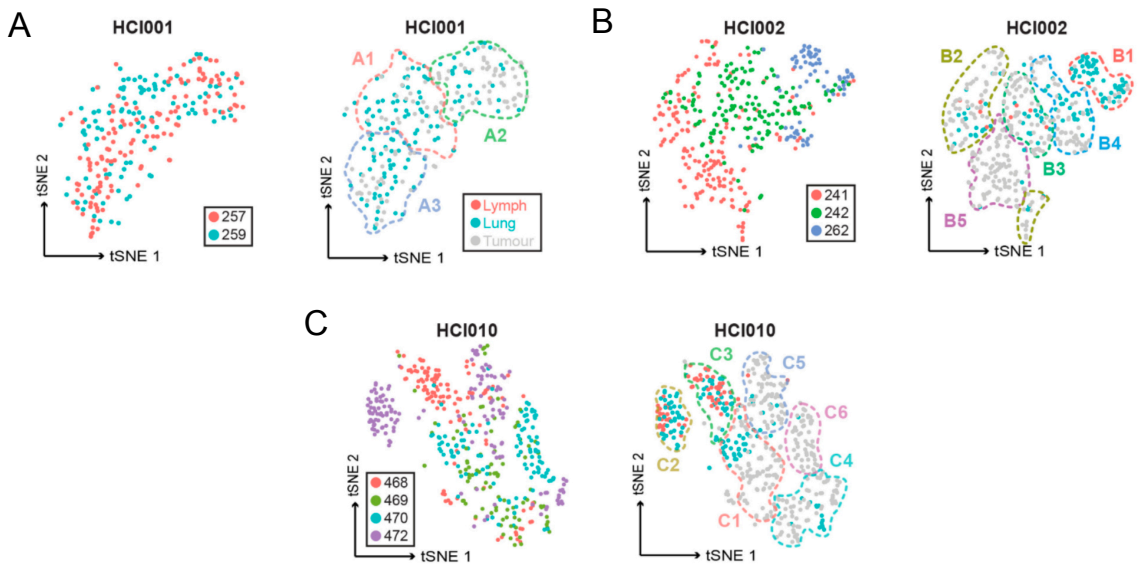

Figure 3.2.2: Cluster associations for individual mice and tumor cell origins. tSNE plots display individual patient analyses, colored by mouse of origin (left) and tissue of origin (right) from PDX models (A) HCI001 (n = 247 cells), (B) HCI002 (n = 401 cells) and (C) HCI010 (n = 471 cells). Additionally, cluster designations are overlayed with colors and names corresponding to **Fig 3.2.1A**. Reprinted and adapted with permission from Davis et al, 2020, *Nat Cell Biol*.

conserved biological specializations driving intra-tumoral heterogeneity across different patient models.

3.2.3 Micrometastatic cells display a distinct transcriptome program.

To determine a set of conserved markers for micrometastatic cells, we performed a supervised differential expression analysis between micrometastatic cells and tumor cells using the tobit test[94] with patient as a latent variable to reduce the effects of inter-tumoral heterogeneity (**Fig 3.3 A**). This identified 330 significantly differentially expressed genes ($P < 0.05$, average log fold change > 0.25) including 116 genes specifically upregulated in micrometastatic cells across all three PDX models (**Fig 3.3 B,C**). The genes upregulated in micrometastasis included heat shock proteins (*HSPB1*, *HSPE1*, *HSPA8*), known to have a role in apoptotic and stress response[95], as well as multiple cytokeratins (*KRT7*, *KRT14, KRT16,*

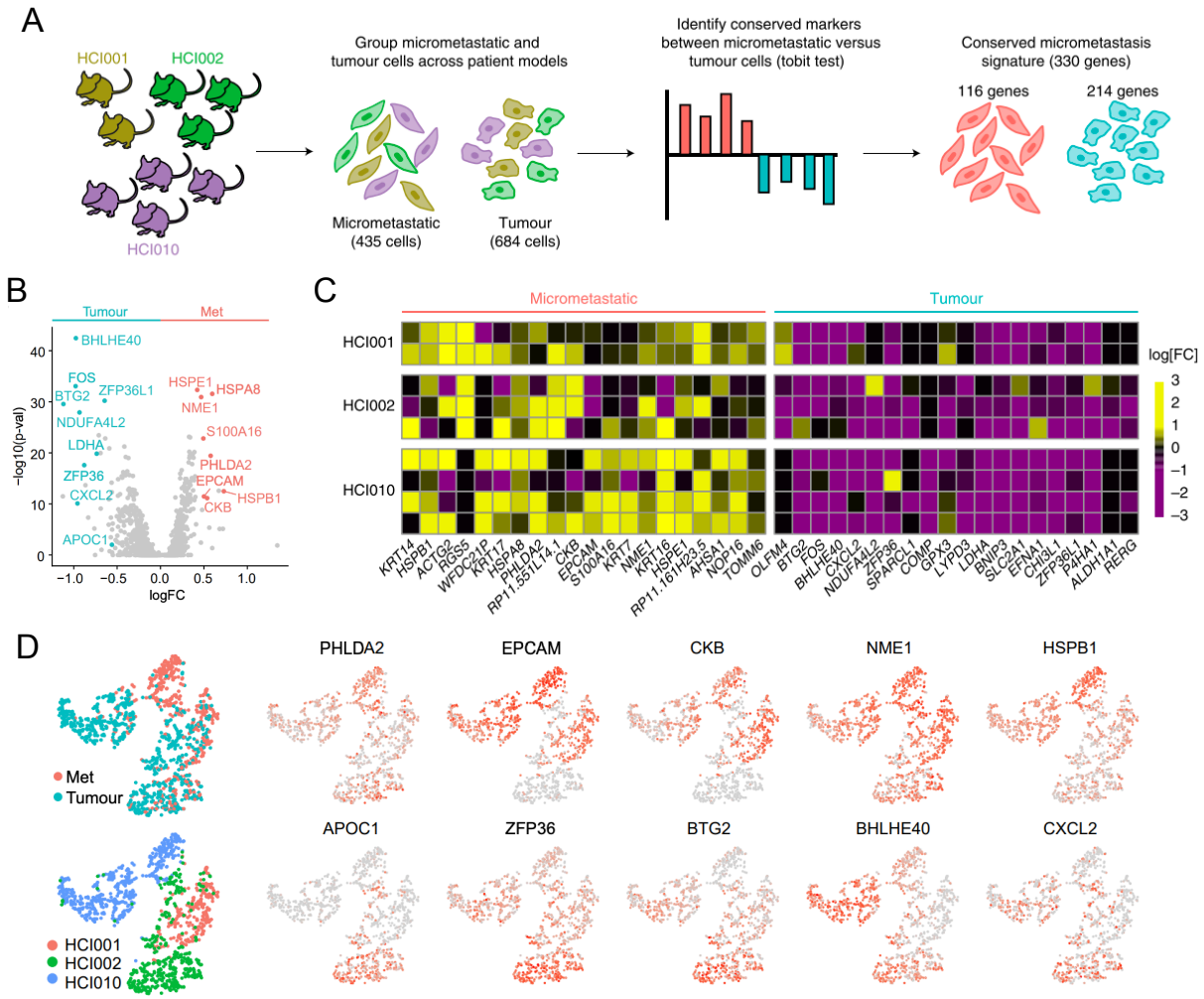

Figure 3.3: Micrometastatic cells display a distinct transcriptome program. (A) Schematic of the analysis method to identify genes that were differentially expressed between micrometastatic and primary tumor cells, conserved in all three PDX models. All of the cells in the dataset were classified as micrometastatic or tumor cells and the differential gene expression between the groups was calculated using the tobit test in Seurat with the patient ID as a latent variable; 330 differentially expressed genes were identified ($P < 0.05$, min.pct = 0.1, average log fold change > 0.25). (B) Volcano plot showing all genes that were differentially expressed between micrometastases (Met) and tumors. (C) Heatmap of the top 20 marker genes for micrometastatic and primary tumor cells. The average log fold change of each gene (x axis) in micrometastatic relative to primary tumor cells was plotted for each mouse (y axis). Yellow indicates higher expression in micrometastatic cells and purple indicates higher expression in tumor cells. (D) tSNE plots show expression of select marker genes in all cancer cells. Clustering was performed based on expression of the top 30 genes differentially expressed between tumor and metastatic cells from (A). Subfigures (A)-(C) reprinted and adapted with permission from Davis et al, 2020, *Nat Cell Biol*. Subfigure (D) is previously unpublished.

KRT17) which are intermediate filaments which have roles in epithelial tissue integrity and mechanical stress responses[96]. We also identified genes marking micrometastases that have not been previously identified in a metastasis context (e.g. *PHLDA2)*, or were shown to have

the opposite effect in other cancers, such as Metastasis Inhibition Factor Nm23 (*NME1*) (**Fig 3.3B,C**).

To further investigate how well these genes separated tumor and micrometastatic cells across patients, we performed a dimensionality reduction using only the top 30 gene markers of tumor and micrometastatic cells as input (**Fig 3.3D**). From this, we found that patients were still clearly distinguishable in the shared space; however, the micrometastatic cells for each patient were pulled together (**Fig 3.3D**). This likely indicates that some of our tumor markers have patient-specificity while the micrometastatic markers are more conserved. Visualizing individual marker genes support this hypothesis, where we see that a few tumor markers were specific to patients (e.g *APOC1* and HCI002) while others convincingly span all three patient models (e.g. *BTG2*) (**Fig 3.3D**). This data indicates that we have effectively narrowed down a list of biomarkers upregulated in micrometastasis, but further filtering for genes downregulated in micrometastasis (i.e. upregulated in tumor cells) across patients may provide improved prognostic capabilities and biological insights.

3.2.4 Improved feature selection using a predictive classifier for micrometastases

We built a logistic regression model using forward selection to optimize biomarker identification from our 330 differentially expression genes using balanced data from all nine mice as input (i.e. equal numbers of tumor and micrometastatic cells from each mouse) (**Fig 3.4A**). We repeated this procedure using ten uniform cell subsamplings, and at each subsampling, five top genes were identified since this best minimized the Akaike information criterion (AIC) (**Fig 3.4A**). Three genes were repeatedly selected in >70% of sampling events, indicating they were robustly predictive across cell subsamplings (**Fig 3.4B**). *PHLDA2* was

identified as the top gene predictive of micrometastatic status, and *BHLHE40* and *LDHA* were identified as top genes predictive of tumor status (**Fig 3.4B**). We subsequently used our model to predict the tumor or micrometastatic identity of individual cells based on their expression of only these three genes (**Fig 3.4C**). We assessed the model on all 1,119 cells in our dataset and found that it demonstrated an overall accuracy of 74.6% (split at 0.5), where it correctly classified metastatic cells 81.5% of the time and primary tumor cells 67.7% of the time (**Fig 3.4C**). Interestingly, when we dug further into the cell type classifications of high confidence (>.75 for micrometastatic label or <0.25 for tumor label), we found again that far more tumors were misclassified as micrometastases than vice versa (**Fig 3.4D**).

To interrogate the prognostic capabilities of these label-predictive genes, we utilized KM plotter, a large online database of breast cancer patient samples with matched gene expression and clinical information[97]. We investigated the effects of our micrometastatic (*PHLDA2*) and tumor predictive (*LDHA*, *BHLHE40*) genes on RFS and found that high levels of *PHLDA2* resulted in a worse prognosis (HR = 1.42) and high levels of *BHLHE40* resulted in an improved prognosis (HR=0.7) (**Fig 3.4E**). High levels of the transcription factor *BHLHE40* has previously been shown to inhibit tumor invasion[98], while *PHLDA2* is known to regulate placental growth, improve engraftment in xenograft models, and improve invasive capacities of tumor cells *in vitro*[99]. In contrast, *LDHA* did not follow the expected pattern of prognostic improvement, which we hypothesize is due to its known role in glycolysis, which may cause it to mark more proliferative, aggressive tumor types with an indistinguishable relationship to metastasis associated phenotypes[100]. We next looked at the RFS predictive capabilities of *PHLDA2* and *BHLHE40* together, and found a small, but notable improvement (HR=1.55) (**Fig 3.4E**).

Because *PHLDA2* was a top gene identified for both single-cell classification and predicting RFS in patient cohorts, we further evaluated its expression in primary tumor and micrometastases using fluorescence *in situ* hybridization technology (RNAscope). We found

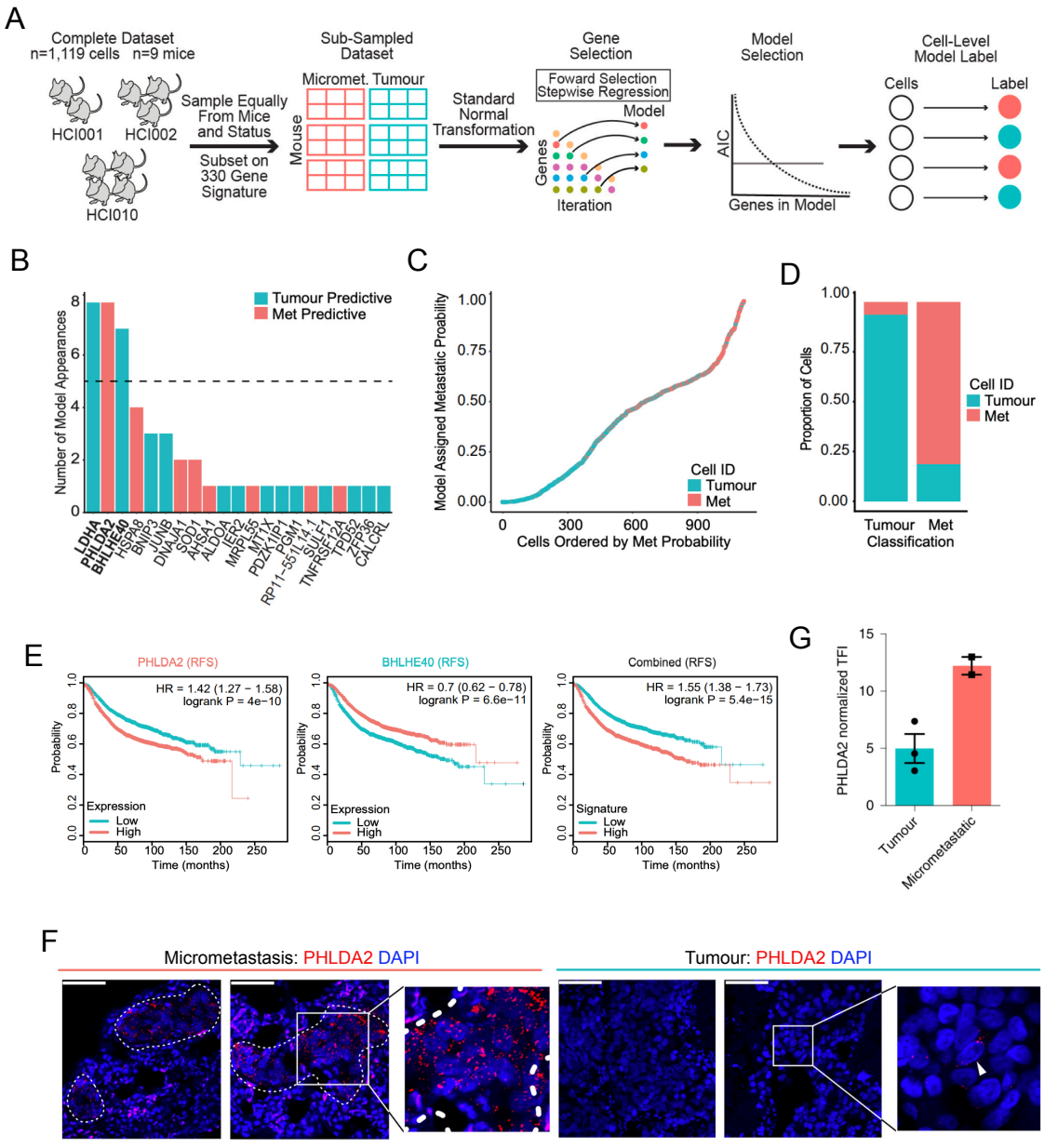

Figure 3.4: Improved feature selection using a predictive classifier for micrometastases (A) Schematic for the construction of a stepwise logistic regression model to identify top biomarker candidates descriptive of primary tumor or micrometastatic cells. The data was subsampled to analyze equal numbers of micrometastatic and tumor cells from each mouse. The model was run on 10 subsamplings of the data, with the number of genes in each model determined by AIC (n=5). (B) Bar plot showing the number of model appearances for each gene out of 10 data subsamplings. (C) Result of logistic regression using PHLDA2, BHLHE40, and LDHA. Each cell was classified as tumor or micrometastatic based on its expression of the three genes. Cells assigned a probability of >0.5 were classified as 'micrometastatic' (Met), while cells assigned a probability of <0.5 were classified as tumor. (D) Quantification of the total model accuracy, subset to cells classified with <0.25 or >0.75 model probability. (E) Kaplan-Meier curves of relapse-free survival (RFS) based on primary tumor expression of PHLDA2, BHLHE40, or both in breast cancer patient cohorts. Data for all breast cancer subtypes is shown (n=1,402). HR= hazard ratio. (F) Representative fluorescent in situ hybridization for PHLDA2 (RNAscope) on primary tumor (n = 3 biologically independent samples; right) and lung micrometastases (n = 2 biologically independent samples; left) from the PDX model HCI001. Insets, higher magnification of individual puncta. The white arrow indicates a tumor cell with high expression of PHLDA2. Scale bars, 50 μm. (G) Normalized total fluorescent intensity (TFI) of PHLDA2 in primary tumor and micrometastatic cells from the PDX model HCI001 (n = 2 lungs, >15 lesions; n = 3 tumors, 22 fields). Data are shown as the mean ± s.e.m. Subfigures (A),(B),(F), and (G) reprinted and adapted with permission from Davis et al, 2020, *Nat Cell Biol*. Subfigures (C)-(E) are previously unpublished.

that the levels of *PHLDA2* transcripts were at least twofold higher in micrometastases relative to primary tumors and only rare tumor cells were found to express the gene in any capacity (**Fig 3.4F,G**). It is tempting to speculate that these rare tumor cells may represent the cells misclassified in our logistic regression model, and perhaps do possess higher metastatic capacities than unmarked cells. Overall, these data highlight our dataset and methodology as a resource for the identification of drivers of metastatic seeding and biomarkers to predict metastatic progression in patients with breast cancer.

3.3 Discussion

In this work, we collected a scRNA-seq dataset of matched tumors and micrometastatic cells across three PDX models and nine xenograft passages, and from it identified novel biomarkers of micrometastasis using multilayered regression models. We performed a simple differential expression between tumor or micrometastatic cells using the tobit test, which identified 330 significantly differentially expressed genes, and many

conserved markers of micrometastases related to stress response. We then added in a forward selection, logistic regression to narrow our list of 330 markers to one candidate marker for micrometastasis (*PHLDA2*) and two potential markers for non-invasive tumor cells (*BHLHE40, LDHA*), the former of which was unconfounded by other fundamental tumor characteristics. *PHLDA2* and *BHLHE40* were both independently and together found to be of prognostic value for RFS in a separate patient cohort. *PHLDA2* has not been extensively studied in metastasis, and our data and *in situ* experiment validated that it is enriched in micrometastatic cells compared to primary tumor cells, and further suggested that rare primary tumor cells show expression comparable to that of micrometastases.

From a methodology standpoint, it is important to note that while we tested the predictive capabilities of our forward selection, logistic regression model on our data, it was not trained to act as a predictor for new data and its only proposed function is in feature selection. It did successfully identify genes robust to noise within our own dataset, and from a practical standpoint, this ended up being genes with minimal to no dropout across both individual cells and our three patient models. The forward selection element also helps to present only the "best" marker of each highly correlated gene set or pathway for further investigation. We propose that this addition to a standard differential expression test is generally useful for narrowing down biomarker candidates in scRNA-seq data, and in our case, resulted in us following up on a gene we otherwise would have looked past since it was neither the top marker in our differential expression analysis nor was it a well-characterized protein.

Interestingly, our logistic regression model misclassified a far larger number of tumor cells as micrometastasis than it misclassified micrometastatic cells as tumors. This may be

biologically meaningful, as it is possible that some cells in the primary tumor are primed for metastasis, and our gene signature may help identify them. The additional observations that rare tumor cell are high for *PHLDA2* and this gene had prognostic utility for RFS data based on primary tumor samples suggests that *PHLDA2* may represent an early marker for metastatic capacity in tumor sections. The additional of a separate predictor of low/non-metastatic cells (e.g. *BHLHE40*) could also improve *in situ* investigations since 'high' and 'low' expression levels of any single gene can vary widely across patients, while a two-gene ratio may have more direct interpretability. Generally, the utility of these two genes as predictors of patients with a high likelihood of having metastasis below clinical detection warrants more investigation, as there are currently very few strategies for this and early therapeutic intervention could save lives.

3.4 Materials & methods

PDX models

The samples from patients were provided by A. L. Welm at the Department of Oncological Sciences at the Huntsman Cancer Institute (HCI). All of the tissue samples were collected with informed consent from individuals being treated at the Huntsman Cancer Hospital and the University of Utah under a protocol approved by the Institutional Review Board of the University of Utah[91]. HCI001 was acquired from a primary tumor biopsy of a female patient diagnosed with Stage IV ER−PR−Her2− basal-like invasive ductal carcinoma with no previous systemic treatment. HCI002 was acquired from a primary tumor biopsy of a female patient diagnosed with Stage IIIA ER−PR−Her2− basal-like medullary-type invasive ductal carcinoma with no previous systemic treatment. HCI010 was acquired from a pleural

effusion of a female patient diagnosed with Stage IIIC ER–PR–Her2– basal-like (PAM50) invasive ductal carcinoma treated with several rounds of chemotherapies12. Additional clinical details of each patient tumor can be found in Supplementary Table 1 of ref [91]. The samples were collected and deidentified by the Huntsman Cancer Institute Tissue Resource and Application Core facility before being obtained for implantation. The study is compliant with all of the relevant ethical regulations regarding research involving human participants.

Animal experiments

The Institutional Animal Care and Use Committee of the University of California, Irvine reviewed and approved all of the animal experiments. Orthotopic transplants of serially passaged human tumor samples were performed on immunocompromised three- to four-week-old NOD/SCID mice after clearing the mammary fat pads following established protocols[101]. Tumor growth was monitored by weekly caliper measurements and volumes were calculated as: length × width2× 0.51. The animals were euthanized and tissues were harvested when the tumors reached a length or width of 2.0–2.5 cm. The study is compliant with all of the relevant ethical regulations regarding animal research.

Tissue harvest and dissociation.

Animals at the endpoint were euthanized by asphyxiation with CO_2 followed by cervical dislocation and perfusion with 10 mM EDTA in D-PBS. Evan's Blue (Sigma-Aldrich, cat. no. E2129-10G) was injected into the footpads and ears of the anaesthetized mice before perfusion to aid visualization of the lymph nodes. The solid tissues from the mice—which included the primary tumor, lungs and lymph nodes—were processed for flow cytometry by

mechanical chopping with blades, followed by collagenase IV (Sigma-Aldrich cat. no. C5138-1G) digestion in medium (DMEM–F12 medium with 5% FBS, 5 μg ml^{-1} insulin and 1% penicillin/streptomycin solution) for 45 min at 37 °C. The cell suspensions were washed with 2 μg ml^{-1} DNAse I (Worthington Biochemical, cat. no. LS002139) for 5 min and further dissociated with 0.05% trypsin for 10 min. Following a wash with Hanks balanced salt solution with 2% FBS, the cells were passed through a 70-μm filter. Lung and primary tumour cells were treated with 1×RBC lysis buffer, followed by resuspension in DMEM–F12 with 10% FBS for FACS.

Flow cytometry

We used the human-specific antibody CD298 (diluted 1:100; PE; BioLegend, cat. no. 341704) and the mouse-specific antibody MHC-I (diluted 1:150; APC; Thermo Fisher Scientific, cat. no. 17-5957-80). Flow cytometry was performed using a BD FACSAria Fusion cell sorter. Cell viability was determined by negative staining with SYTOX blue (diluted 1:1,000; Thermo Fisher Scientific, cat. no. S34857). The forward-scatter area by forward-scatter width (FSC-H × FSC-A) and side-scatter area by side-scatter width (SSC-H × SSC-A) was used to discriminate single cells from doublet and multiplet cells. Mouse cells were excluded by gating out CD298−MHC-I+ cells. Human primary tumor cells and metastatic cells were selected by gating on Sytox−CD298+MHC-I− cells.

Generation of scRNA-seq data

Single cells were sorted directly into each well of a skirted 96-well PCR plate (Fisher Scientific, Eppendorf, cat. no. E951020443) containing lysis buffer (0.2% Triton X-100

(Sigma-Aldrich, cat. no. T9284), 2 U μl^{-1} RNAseOUT (Thermo Fisher Scientific, cat. no. 10777019), 10 μM oligo-dT30VN and 10 μM dNTPs (Thermo Fisher Scientific, cat. no. 18427088)) as described previously[22]. The plates were snap frozen on dry ice and stored at −80 °C until further processing. Total RNA was converted into complementary DNA using the Smart-seq2 protocol and prepared for Illumina sequencing using the Nextera XT DNA library preparation kit (Illumina, cat. no. FC-131-1096). The cells were sequenced at a depth of 1 × 106 reads per cell on a HiSeq 2500 system.

Processing of scRNA-seq data

Files from the HiSeq 2500 were demultiplexed and converted to FASTQ files. Paired-end 100 bp reads were aligned to the Gencode 21 human transcriptome using Bowtie 2 and quantified using RSEM with the following parameters: rsem-calculate-expression -p $CORES —bowtie2 —paired- end READ1 READ2 gencodehg21. The expression values were log-transformed into log[transcripts per kilobase million + 1] matrices and loaded into the Seurat analysis package with the following parameters: p10<- CreateSeuratObject(raw. data = p10.mat, min.cells = 8, min.genes = 1,000, project = 'HCI010'). We removed any cells identified as visual outliers by library complexity (<2,500 genes per cell) or overrepresentation of mitochondrial gene expression (>50%) as a further quality control. In addition, we removed any genes that were not represented in a robust population of cells (<8 cells per gene) from the downstream analysis. This resulted in a final analysis of 1,119 single-cell profiles. Using the RegressOut feature in Seurat, we calculated the z-score residuals using nGene and percent.mito as co-variates, which was used to perform principal-

component analysis and tSNE. A G1/S and G2/M score was calculated using the gene score method described below and regressed out as well for HCI001 and HCI010.

Dimensionality reduction, cell cluster identification and differential gene expression analysis

Dimensionality reduction and differential gene expression was performed using the Seurat analysis package version 2.1.0. For the main combined and individual patient analysis, highly variable genes in our dataset were identified using the MeanVarPlot function with the following parameters: FindVariableGenes(object = comb, mean.function = ExpMean, dispersion. function = LogVMR, x.low.cutoff = 0.0125, x.high.cutoff = 3, y.cutoff = 0.5). For the combined analysis for only micrometastatic and tumor differentially expressed genes, the top 30 markers of tumor or micrometastatic status were used as direct input into the principal-component analysis instead. These variable genes were then used for principal-component analysis. The principal components generated were then used to perform tSNE of the data. For the individual patient analysis, using the FindClusters function in Seurat and a granularity parameter of 1.0, we identified distinct subpopulations and defined marker genes for each of them with the FindAllMarkers function in Seurat with the default settings for the FindAllMarkers function and the 'bimod' statistical test. For the generation of the 330-micrometastatic-gene signature, metastatic cells from all PDX models were grouped together separate from tumor cells and we calculated a differential expression test in Seurat using the 'tobit' test with the following parameters: comb<- FindAllMarkers(object = comb, only.pos = TRUE, min.pct = 0.1, logfc.threshold = 0.25, test.use = 'tobit', latent.vars = 'orig.ident'). The 'orig.ident' command in the 'latent.vars' variable represents the patient ID (that is, HCI001,

and so on). By including this variable, the tobit model identifies conserved marker genes. The 'min.pct' variable in Seurat's differential expression (DE) tobit test is defined as the minimum percent of cells per group that must express a gene (log(TPM + 1) >0) to be considered in the output of the test. Gene Ontology analysis was performed using the Enrichr web resource[102,103], where the input gene set for each population was the markers identified by FindAllMarkers.

Development of logistic regression model for identifying candidate biomarkers

The classification model was calculated beginning from 1,119 single cells and 330 differentially expressed genes as calculated with a generalized additive model (tobit) in Seurat's FindAllMarkers() function. Patient ID was used as a latent variable when calculating differential expression. The gene expressions were normalized across all cells such that each gene had a mean expression of zero and a standard deviation of one. For model fitting the data was sampled equally ten times from each mouse and origin (tumor vs micrometastasis) category to avoid systematic bias. For each sampling, a stepwise regression with forward selection was performed where at each step, the model that minimized the AIC was chosen to be used as a base model for the next step. Proceeding in this fashion the p-value of each marginal gene included in the model was recorded as well as the area under the receiver operating curve (AUC). It was noted that the Bonferroni corrected p-value of the coefficient for each marginal gene fell outside of a 0.01 cutoff by the fifth gene that was included in the model. Thus, the genes included in the ten five-gene models were tabulated. From this tabulation it is apparent that the most frequently occurring genes are *LDHA*, *PHLDA2*, and *BHLHE40* occurring in eight, eight, and seven of the ten models respectively and a final model

was constructed by using those three genes in all cells to calculate a logistic regression classifier for tissue source (tumor vs micrometastasis).

Relapse-free survival analysis

For the relapse-free survival analysis, we generated Kaplan–Meier survival curves on primary tumor microarray data of patients with all subtypes of breast cancer from the KM Plotter database[97]. To generate the combined survival analysis, we calculated a weighted average of PHLDA2 and -1*BHLHE40 using the 'Use Multiple Genes' function in KM Plotter. All Kaplan–Meier plots are displayed using the 'auto select best cutoff' parameter.

Histology

Tumor and lung tissues from the PDX mouse models were fixed overnight in 4% paraformaldehyde and then dehydrated and processed for paraffin embedding in a Leica tissue processor using standard protocols. The paraffin blocks were cut into 5-µm-thick sections using a Leica microtome, rehydrated and then stained with haematoxylin and eosin. Bright-field imaging was performed using a BZ-X700 Keyence microscope.

Fluorescent in situ hybridization

Fluorescent in situ hybridization was performed on formalin-fixed paraffin-embedded sections using the RNAscope multiplex fluorescent reagent kit v2 (ACD, cat. no. 323110) according to the manufacturer's instructions. Briefly, the formalin-fixed paraffin-embedded sections were rehydrated in Histoclear and 100% ethanol before antigen retrieval using the RNAscope antigen retrieval solution and mild boiling at 100 °C for 15 min. PHLDA2 probe

(ACD, cat. no. 551441) amplification was performed according to the manufacturer's instructions with the TSA plus cyanine 3 (diluted 1:1,000; PerkinElmer, cat. no. PN NEL744001KT) fluorophore, stained with DAPI and mounted with Prolong Gold. The slides were visualized using a Zeiss LS700 confocal microscope. Image analysis was performed in ImageJ. The normalized total fluorescence intensity of the PHLDA2 probe was calculated on regions of interest across at least five different fields of view on two mouse lungs and three tumors from HCI001 according to the following equation: Normalized TFI = (Fluorescence integrated density)/(Area of region of interest). The surrounding mouse stroma was excluded from the analysis for the quantification of lung micrometastatic regions of interest and the necrotic regions or mouse stroma were excluded from the analysis for the tumor regions of interest.

Data availability

All RNA-seq data files along with their associated metadata have been deposited in the GEO database under the accession code GSE123837.

Code availability

Custom scripts are available at https://github.com/lawsonlab/Single_Cell_ Metastasis

Portions of the Introduction, Results, and Methods in this section were reprinted and adapted with permission from:

Davis, R.T., Blake, K., Ma, D. et al. Transcriptional diversity and bioenergetic shift in human breast cancer metastasis revealed by single-cell RNA sequencing. Nat Cell Biol 22, 310–320 (2020). https://doi.org/10.1038/s41556-020-0477-0

CHAPTER 4: Microglia heterogeneity in breast cancer breast metastasis

4.1 Introduction

Breast cancer brain metastasis (BCBM) is increasing in incidence and rapidly emerging as a critical clinical problem in breast cancer. 15-30% of metastatic breast cancer patients develop brain metastasis, and autopsy and imaging studies indicate an additional 30% of patients are likely to develop brain metastasis as treatments for peripheral disease improve and patients live longer [104,105]. This is alarming since there are no effective treatments for brain metastasis and median survival is only a few months [106–109]. There is growing interest in immunotherapeutic strategies to treat central nervous system (CNS) cancers, given that immune cells enter the brain during disease while most conventional therapies are precluded by the blood brain barrier (BBB) [110,111]. However, greater understanding of the immune response to BCBM will be needed to develop immunotherapy strategies effective in the unique immune microenvironment of the CNS.

The brain immune microenvironment is principally composed of specialized tissue resident macrophages called microglia that tile the brain and play diverse functions in CNS homeostasis and disease [112–114]. Microglia represent a prime immunotherapeutic target because they are the first line of defense to disease in the CNS and have the power to direct the initial immune response. BCBMs are heavily infiltrated with tumor associated macrophages (TAMs), which may be comprised of microglia, border-associated macrophages (BAMs), as well as bone marrow derived monocytes and macrophages

(BMDMs) [115–119]. Functional studies using genetic and pharmacologic approaches to deplete TAMs overwhelmingly support a tumor promoting role for these cells. Depletion of TAMs with CSF1R inhibitors results in tumor reduction and decreased metastasis in glioblastoma and melanoma models [120–124]. TAM depletion using a CX3CR1-targeted genetic ablation model similarly results in decreased BCBM [125]. However, it is unclear whether microglia or other types of TAMs produce the tumor promoting effects observed in these studies. CSF1R inhibitors have been shown to preferentially deplete microglia, but also attenuate other myeloid cells, and microglia ultimately repopulate the brain when treatment ceases. Likewise, CX3CR1 is expressed by diverse myeloid cell populations and upregulated by BMDMs upon entry into the brain [125,126]. Therefore, the impact of brain resident microglia on tumor initiation and their potential as an immunotherapy target remain unclear.

We combined single cell RNA-sequencing (scRNA-seq) with newly developed genetic and humanized mouse models to show for the first time that microglia exert a potent tumor suppressive effect on BCBM initiation. ScRNA-seq of >75,000 cells from three different models revealed that microglia mount a robust pro-inflammatory response to BCBM. Subclustering of pro-inflammatory microglia showed further specialization of their response, where distinct populations of microglia upregulate programs for antigen presentation, IFN response, phagocytosis, cytokine production, and glycolysis. ScRNA-seq showed that these discrete microglia substates were conserved in a humanized mouse model of BCBM, suggesting that human microglia have the capacity to respond similarly to disease initiation in BCBM patients. Finally, we investigated the function of microglia in BCBM

initiation using an innovative new genetic model that specifically lacks microglia while retaining other myeloid cells [127]. We find that the absence of microglia results in decreased survival and increased BCBM progression, showing that microglia play an important role in tumor suppression. This contrasts with the pro-tumorigenic function reported for other types of TAMs in CNS cancer, and highlights the potential of harnessing the natural tumor suppressive function of microglia to treat brain metastasis.

4.2 Results

4.2.1 BCBM are extensively infiltrated with activated TAMs

During homeostasis, the brain is home to microglia that tile the parenchyma as well as BAMs that reside in the meninges, choroid plexus, and perivascular surface of blood vessels **(Fig 4.1.1A)**. During inflammation, there can be substantial infiltration of BMDMs that express similar markers, making it difficult to determine the origin and function of TAMs in BCBM (**Fig 4.1.1A**). We first investigated TAM activation and localization in human patient BCBM by immunofluorescence (IF) staining for the canonical activation marker ionized calcium-binding adaptor molecule 1 (IBA1), which is expressed lowly by homeostatic microglia and highly by activated microglia and macrophages (**Fig 4.1.1A**) [128]. As expected, we find that IBA1$^+$ cells are evenly spaced throughout normal brain and display small cell bodies and ramified morphology typical of homeostatic microglia (**Fig 4.1.1B**). In contrast, we find that BCBM are heavily infiltrated with IBA1$^+$ cells that display ameboid morphology typical of activated microglia and macrophages. We subsequently turned to a well-established mouse model of BCBM, MDA-MB-231-BR2 (231BR) for further investigation of

TAM origin and function [129,130]. We performed intracardiac (i.c.) injections of 213BR cells stably expressing firefly luciferase and AcGFP reporters into Foxn1$^{nu/nu}$ mice (**Fig 4.1.1C**). Consistent with prior reports, 231BR cells arrest in blood vessels and cross into the brain two to seven days after injection, then grow along blood vessels and form micrometastases by day 14 and parenchymal metastasis by day 28 (**Fig 4.1.1D**)[131,132]. Interestingly, IF analysis shows that IBA1^{+} cells surround and directly interface with cancer cells by day seven, showing they interact with cancer cells at the initial stages of micrometastasis initiation (**Fig 4.1.1D**). We further find that day 28 parenchymal metastases are densely infiltrated with IBA1^{+} cells, in contrast to regions of normal tissue distal to metastases (**Fig 4.1.1D**). Quantification of IBA1 fluorescence intensity shows 4-fold higher signal in parenchymal metastases compared to control brains ($p<0.0001$) (**Fig 4.1.1E**). These data show that TAMs immediately interact with metastatic cells and become progressively activated in mouse and human BCBM.

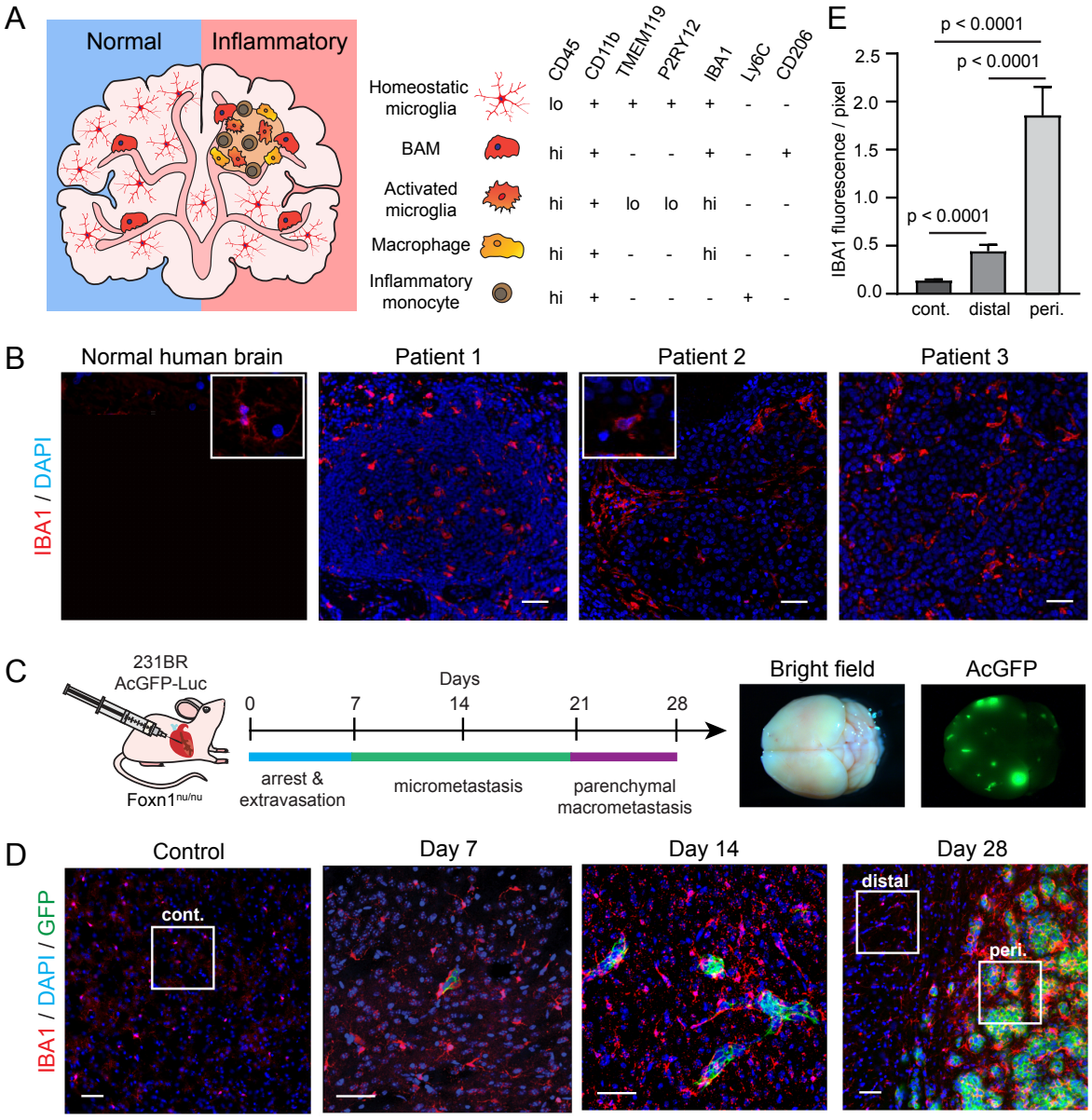

Figure 4.1.1: BCBM are extensively infiltrated with activated TAMs. (A) Schematic (left) of brain resident and bone marrow-derived macrophage cell types present in the normal and inflammatory brain microenvironment and their relative expression of canonical markers (right). BAM = border-associated macrophage. (B) IF staining shows IBA1$^+$ cells (red) in normal human brain and three resected patient BCBM tumors. Insets show cell morphology. Scale bar = 50μm. (C) Schematic showing disease progression in mouse 231BR-Foxn1$^{nu/nu}$ BCBM experimental metastasis model. 500,000 AcGFP-Luc labeled 231BR cells were injected into the left cardiac ventricle of Foxn1$^{nu/nu}$ mice and harvested 28 days later. Whole mount brightfield and fluorescence microscopy images show a representative brain with AcGFP$^+$ metastatic foci (green). (D) IF staining shows IBA1$^+$ cells (red) in control and metastatic brains at 7, 14, and 28 days post 231BR cell injection. Metastatic cells are AcGFP$^+$ (green). Boxes indicate representative peritumoral (peri.), distal and control (cont.) regions for IBA1 quantified in (E). Scale bar = 50μm. (E) Quantification of IBA1 expression in control (n=4) and metastatic (n=4) brains 28 days post 231BR cell injection. IBA1 fluorescence intensity per pixel was quantified in control (cont., n=115 fields), peritumoral (peri., n=127 fields) and distal (n=96 fields) regions as shown in (D). *P* values were generated using a two sided, unpaired Welch's *t*-test and error bars show standard deviation.

4.2.2 Microglia display a robust pro-inflammatory response to BCBM initiation

We used scRNA-seq to investigate the specific function of microglia in BCBM and discriminate them from other TAM populations. Cells were dissociated from control (n=3) and metastatic brains (n=3) by automated heated mechanical and enzymatic digestion followed by density centrifugation to remove myelin (**Fig 4.2.1A, Fig 4.1.2A**). Live metastatic ($CD45^-GFP^+$) and myeloid cells ($CD45^+CD11b^+$) were subsequently isolated by flow cytometry (**Fig 4.1.2B**). Astrocytes ($CD45^-ASCA2^+$) were also sorted since they have been previously implicated in BCBM [132–134](**Fig 4.1.2B**). Isolated cells were captured and prepared for sequencing using droplet-based technology (Chromium) (**Fig 4.2.1A**).

Mouse cells were identified by aligning to a merged human (GRCh38) and mouse (mm10) genome, where cells were identified as mouse if >87.5% of reads aligned to mm10 (**Fig 4.1.2C**). We also removed poor quality cells and doublets by excluding cells with <500 genes, >2000 genes, or a mitochondrial gene percentage >10% (**Fig 4.1.2D**). The remaining cells were integrated across sequencing batch using the mutual k-nearest neighbors (kNN) algorithm adaptation in the Seurat pipeline [135,136]. Analysis of the 42,891 cells that passed further filtering revealed seven distinct cell types identified by lineage-specific markers and visualized by t-distributed stochastic neighbor embedding (tSNE) (**Fig 4.2.1B**, **Fig 4.1.2E,F**). This included the targeted cell types, astrocytes (*Aldoc, Atp1a2*), microglia (*Tmem119, P2ry12*) and non-microglia myeloid cells including dendritic cells, monocytes and macrophages (*Lyz2, Plac8*) (**Fig 4.2.1B**, **Fig 4.1.2E,F**). We also recovered small numbers of ependymal cells (*Ccdc153, Rarres2*), oligodendrocytes (*Mbp, Ptgds*), vascular cells (*Cldn5,*

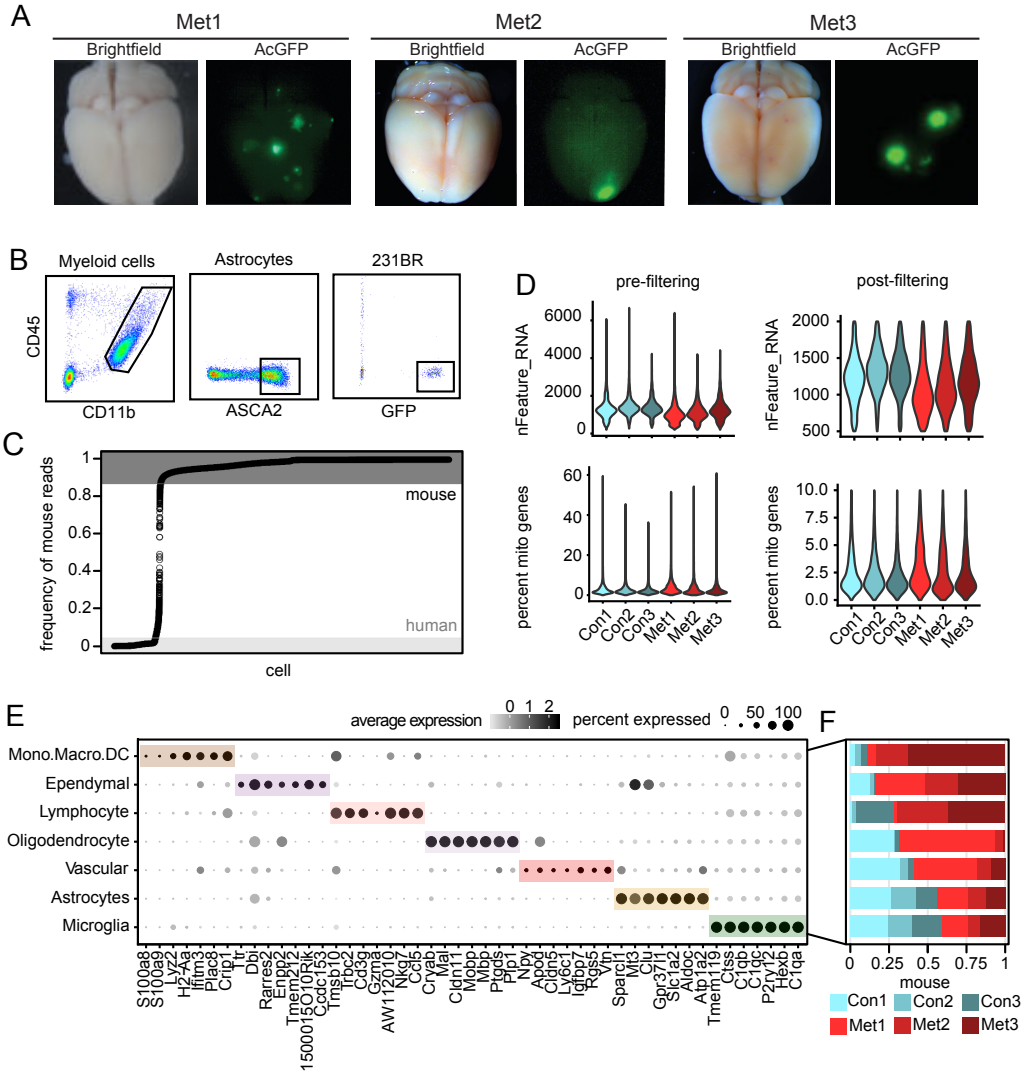

Figure 4.1.2 : Quality control and exclusion criteria for Foxn1$^{nu/nu}$ scRNA-seq cell libraries. (A) Whole mount brightfield and fluorescent microscopy images of metastatic brains (Met1-3) used to generate the scRNA-seq dataset described in **Fig 2**. Metastatic lesions are AcGFP^{+} (green). (B) Representative FACS plots show gating for single, live (Sytox negative) myeloid cells (CD45^{+}CD11b^{+}), astrocytes (CD45^{-}ASCA2^{+}) and 231BR cells (CD45^{-}GFP^{+}) isolated for scRNA-seq. (C) Identification of mouse and human cells by the frequency of reads that align to the mm10 mouse genome. Cutoffs used to identify mouse cells (>0.875 aligned, n=51,418 cells), human cells (<0.05 aligned, n=7336 cells) and doublets (0.05-0.875 aligned, n=913 cells) are shown. (D) Violin plots show cell distributions for key quality control metrics pre- and post-filtering and removal of poor quality cells. Cells were removed that displayed <500 or >2000 genes (nFeature_RNA), or >10% of genes mapped to the mitochondrial genome (percent mito genes). (E) Dot plot shows top marker genes for each cell type ranked by the average natural logFC and determined by the Wilcoxon rank sum test. Dot size represents the percentage of cells that express the gene, and dot greyscale represents the average expression level. (F) Bar chart showing the frequency of cells contributed by each mouse that localize to each cell type in (E).

Vtn), and lymphocytes (*Cd3g, Gzma*) (**Fig 4.2.1B, Fig 4.1.2E-F**). Peripheral immune cells,

namely non-microglia myeloid populations and lymphocytes, were found preferentially in the metastatic condition (**Fig 4.1.2E,F**).

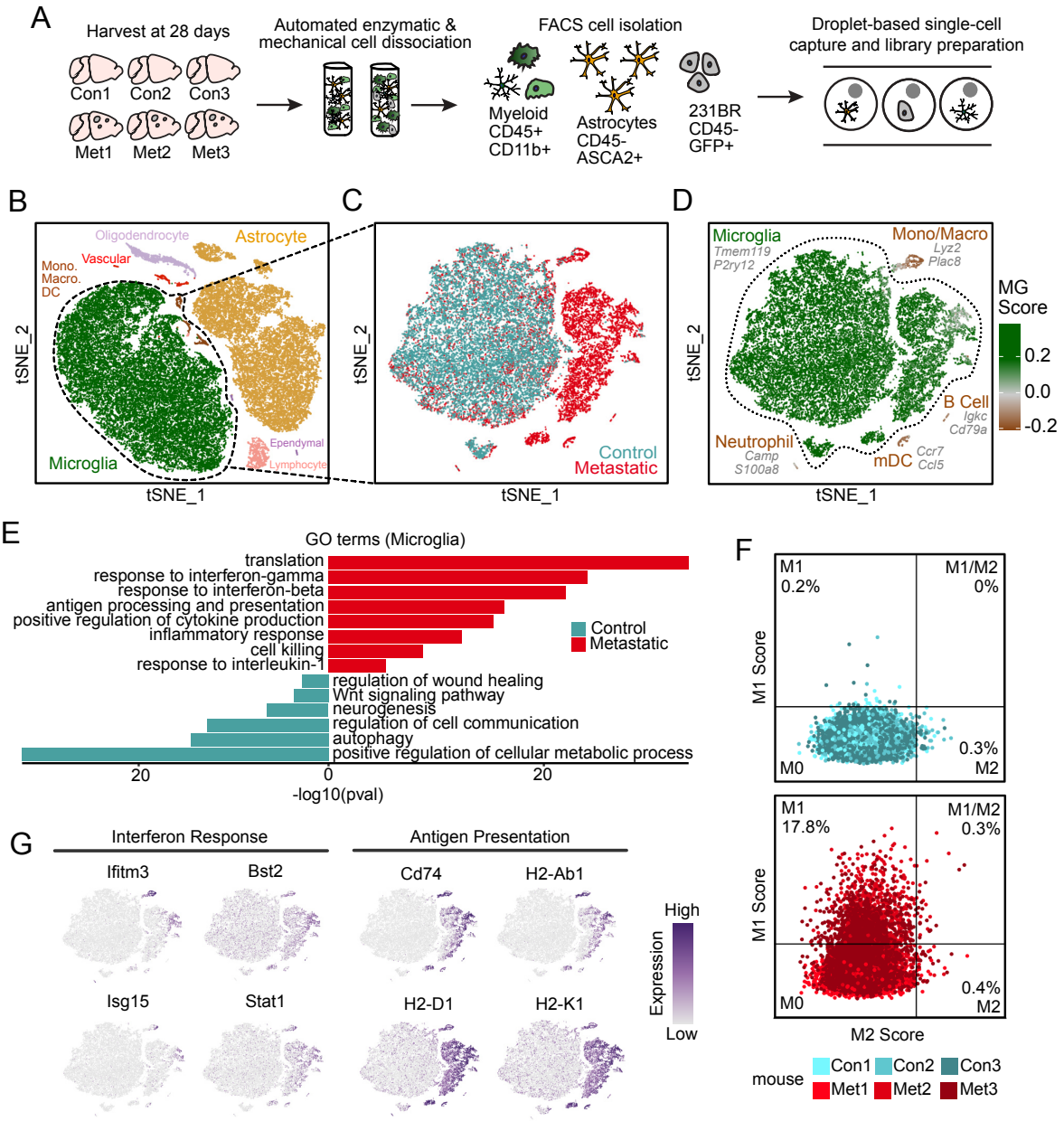

Figure 4.2.1: Microglia display a robust pro-inflammatory response to BCBM (A) Schematic showing experimental design for generation of scRNA-seq dataset. Foxn1$^{nu/nu}$ mice were injected with 500,000 AcGFP-Luc labeled 231BR cells and brains were harvested 28 days later. Three metastatic (Met1-3) and three control (Con1-3) and brains were digested, and myeloid cells, astrocytes and 231BR cells were isolated by flow cytometry for droplet-based scRNA-seq. (B) tSNE plot shows mouse cells that passed filtering (n=42,891), colored and labeled by cell type. (C) tSNE plot shows clustering of myeloid cells (n=24,348), colored by condition. (D) tSNE plot shows each myeloid cell colored by its MG-score, the core microglia gene signature from Bowman et al (2016). Scores were calculated using the AddModuleScore function in Seurat. Top marker genes (gray) for each myeloid cell type were identified using the Wilcoxon rank sum test in Seurat v3. mDC = mature dendritic cell; Mono/Macro = monocytes and macrophages. (E) Bar plot shows selected top GO terms identified for microglia from control (n=3,083 genes, adj. p<0.05) and metastatic (n=609 genes, adj. p<0.05) brains. Differentially expressed genes were determined using the Wilcoxon rank sum test. GO terms were determined using MouseMine and select terms with Holm-Bonferroni adjusted *P* values <0.05 were retained. (F) Scatter plots showing gene scores for M1 or M2 macrophage gene signatures in microglia from control and metastatic brains based on the lists from Azizi et al (2018). Control mice were used to draw boundaries for positive or negative M1 and M2 scores. (G) Feature plots show relative expression in each cell for key marker genes associated with top GO terms.

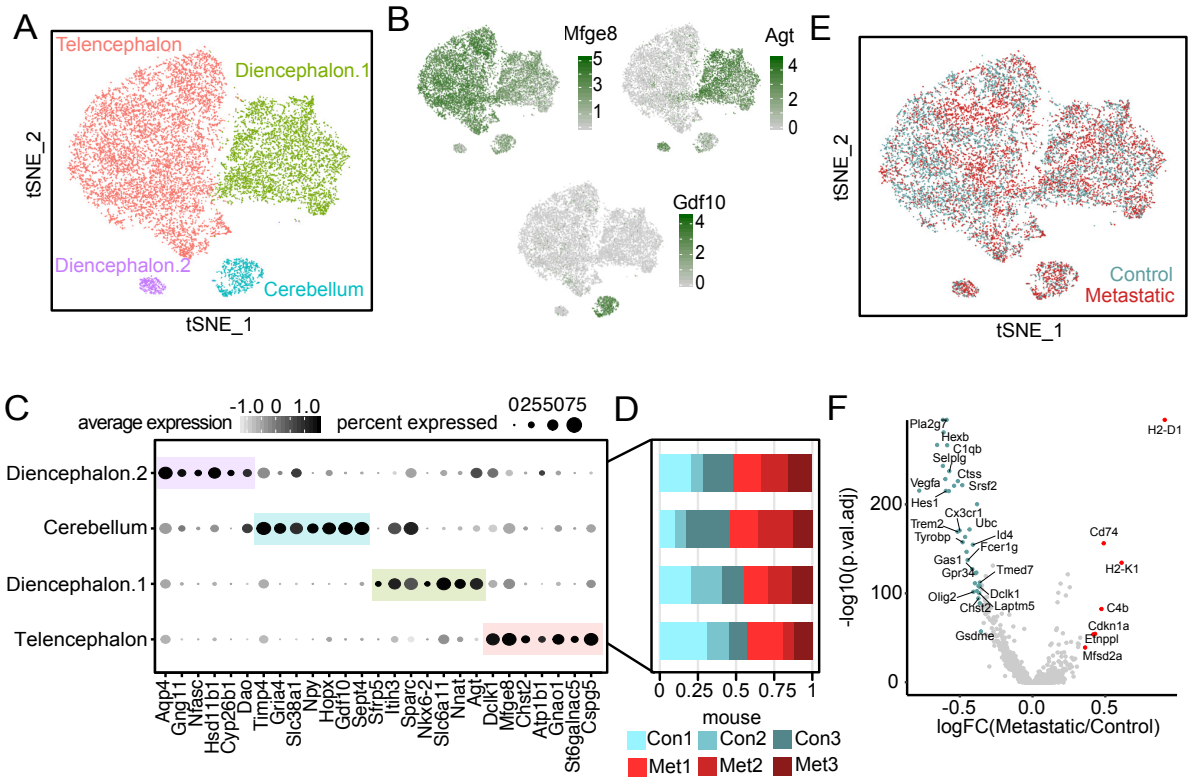

Figure 4.2.2: Astrocytes display regional heterogeneity but limited response to BCBM. (A) tSNE plot shows clustering of astrocytes (n=15,288) colored and labeled by brain region. (B) Feature plots show relative expression in each cell for key marker genes of astrocytes from the telencephalon (Mfge8), diencephalon (Agt) and cerebellum (Gdf10). (C) Dot plot shows top marker genes for each cluster ranked by average natural logFC and determined by the Wilcoxon rank sum test. Dot size represents the percentage of cells that express the gene, and dot greyscale represents the average expression level. (D) Bar chart shows the frequency of cells contributed by each mouse that localize to each cluster in (C). (E) tSNE plot of astrocytes colored by condition. (F) Volcano plot shows genes differentially expressed (n=6,542) between astrocytes from control and metastatic brains determined by Wilcoxon rank sum test, ($p < 0.01$). Select genes with an absolute value average natural logFC >0.35 are colored and labeled. The y-axis represents the -log10 of Bonferroni corrected *P* values, and the x-axis represents average natural logFC between conditions.

Further analysis of the 231BR cells showed limited heterogeneity beyond cell cycle differences. We found substantial heterogeneity amongst astrocytes, but it was principally associated with regional localization (**Fig 4.2.2A,B**). Consistent with prior work, astrocytes formed discrete subpopulations associated with the telencephalon (*Mfge8)*, diencephalon strong transcriptomic shifts associated with BCBM using unbiased clustering or supervised differential expression analysis (**Fig 4.2.2E,F**).

In contrast to astrocytes, tSNE visualization of myeloid cells showed strong separation of control and metastatic conditions (**Fig 4.2.1C**, **Fig 4.2.3E**). Microglia were distinguished from other myeloid populations by scoring cells for the core microglia signature (MG-score) developed in Bowman et al (2016), which compared microglia to BMDMs using bulk RNA-seq from lineage labeled mice (**Fig 4.2.1D**) [126]. Marker gene analysis confirmed the presence of two major microglia populations *(Tmem119, P2ry12)* (**Fig 3.1.1A**, **Fig 4.2.1D**, **Fig 4.2.3A-B**)**,** where one contained microglia from both control and BCBM and the other was almost fully from BCBM (**Fig 4.2.1C**). We also identified two small populations of microglia that display an increased stress response (**Fig 4.2.3D**), which is common post tissue manipulation [30], as well as populations of neutrophils (*Camp*, *S100a9*), monocytes/macrophages (*Ly6c2, Lyz2*), mature dendritic cells (*Ccr7, Flt3*), and B cells (*Igkc,*

Cd79a) (**Fig 4.2.1D**, **Fig 4.2.3A-D**). The latter were predominantly recovered from metastatic animals (**Fig 4.2.1C**, **Fig 4.2.3C**), suggesting they are recruited to the CNS in response to metastatic outgrowth.

To identify gene expression changes in microglia associated with BCBM, from here on called BCBM-response (BCBM-R) microglia, we performed differential gene expression and pathway analyses. Supervised analysis revealed 3,715 genes differentially expressed between microglia from control mice and mice with BCBM (adjusted $p<0.05$). Gene Ontology (GO) analysis of this BCBM-R signature identified 'cytokine production,' 'antigen processing and presentation,' 'cellular response to IL-1,' 'response to IFN-gamma,' and 'response to IFN-beta' as top GO terms, suggesting that microglia undergo a primarily pro-inflammatory response to brain metastasis (**Fig 4.2.1E**) [137]. This is further supported by scoring each cell for a list of genes associated with pro-inflammatory (also known as M1) versus alternatively activated, anti-inflammatory (M2) macrophage responses[62]. Microglia from two of the three mice with BCBM (Met2 and Met3) showed a strong M1 upregulation with minimal M2 upregulation in all mice (**Fig 4.2.1F**). Most M2 markers that were expressed in the brain during BCBM (e.g. *Cd163*, *Ccl17*, *Mrc1*) were enriched in non-microglia TAM populations (**Fig 4.2.3F**).

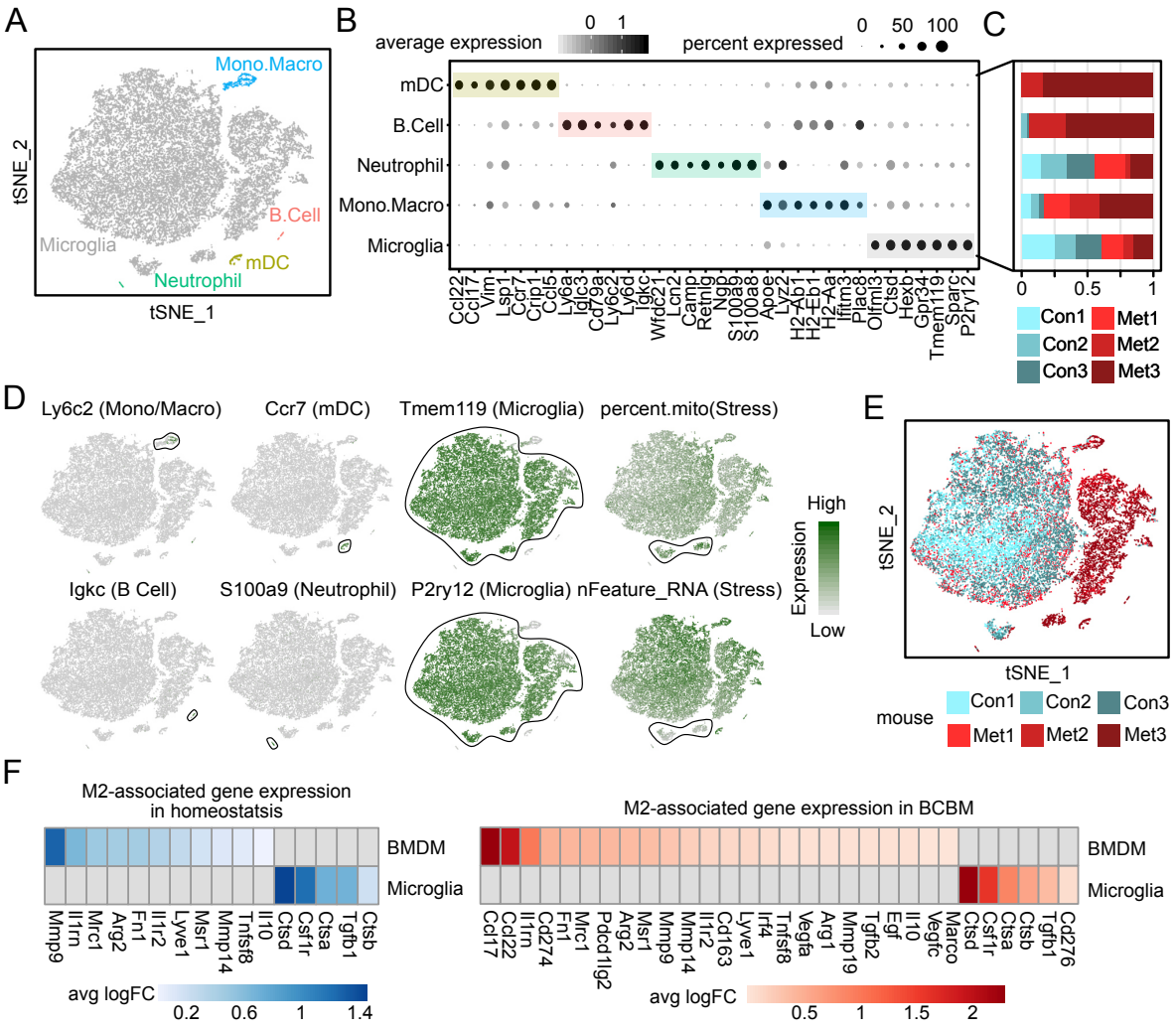

Figure 4.2.3: Identification of myeloid cell types in BCBM. (A) tSNE plot shows myeloid cells (n=15,288) colored and labeled by cell type. mDC = mature dendritic cell. (B) Dot plot showing top marker genes for each cell type ranked by average natural logFC. Dot size represents the percentage of cells that express the gene, and dot greyscale represents the average expression level. (C) Bar chart showing the frequency of cells contributed by each mouse that localize to each cell type in (B). (D) Feature plots show myeloid cells colored by top lineage-specific marker genes or features. Stressed cells were identified by increased expression of mitochondrial genome (percent.mito) genes, and decreased number of genes detected (nFeature_RNA). (E) tSNE plot of myeloid cells, colored by mouse. (F) Heatmaps show M2-associated genes differentially expressed between microglia and BMDM in homeostasis and BCBM. Differentially expressed genes (unadjusted $P < 0.01$) were determined using the Wilcoxon rank sum test and are displayed as average natural logFC. BMDMs include neutrophils, mono/macro, and mDC.

Examination of the genes associated with each GO term showed that BCBM-R microglia upregulate a series of IFN-beta (Type I) response genes typical of an inflammatory response, including *Bst2*, *Ifitm3, Isg15,* and *Stat1* (**Fig 4.2.1G**). BCBM-R microglia also

upregulate extensive genes associated with antigen presentation, including the MHC-I genes *H2-D1* and *H2-K1,* the MHC-II genes *H2-Ab1* and *Cd74*, as well as the proteasome activator subunits *Psme1* and *Psme2* (**Fig 4.2.1G**). Additionally, they upregulate key pro-inflammatory cytokines *Il1b, Tnf*, *Mif*, and *Spp1*, as well as many chemokines that promote immune cell recruitment [138], including *Ccl2* and *Ccl12* for inflammatory monocyte trafficking, *Ccl3*, *Ccl4* and *Ccl5* for macrophage and NK cell migration, and *Cxcl9* and *Cxcl10* for CD8 T cell recruitment for a Th1 response[139]. These data show that microglia mount a robust pro-inflammatory response to BCBM, characterized by increased IFN response genes, cytokine production, and antigen presentation machinery.

4.2.3 The microglia pro-inflammatory response is conserved in diverse BCBM models

We investigated the microglia pro-inflammatory response in three BCBM models, the human 231BR (Foxn1$^{nu/nu}$) and two mouse immune competent models, 4T1 (BALB/c) and EO771 (C57BL/6) [125,130,140–142]. We evaluated protein expression of three representative markers by flow cytometry; the IFN-beta response gene bone marrow stromal antigen 2 (BST2), and major histocompatibility complex II (MHC-II) and CD74 which are critical for antigen presentation [143–146]. In the 231BR-Foxn1$^{nu/nu}$ model, tissues were harvested 28 days post injection and microglia were identified by gating on CD45^{lo}CD11b^{+}Ly6C^{-} cells (**Fig 4.3.1A**) [147,148]. Remarkably, we found a 10-fold increase in the frequency of CD74 (p=0.001) and BST2 (p=0.0001), as well as a 20-fold increase in MHC-II (p=0.04) positive microglia in metastatic (n=14) versus control (n=7) brains (**Fig 4.3.1A**), validating our findings from scRNA-seq. In situ IF analysis further showed that the response is specific to microglia proximal to metastatic lesions. Co-staining of CD74 with the microglia-specific marker,

transmembrane protein 119 (TMEM119) (**Fig 4.1.1A**), showed that CD74 is specifically upregulated by microglia that directly interface with micrometastatic lesions, while distal microglia remain CD74 negative (**Fig 4.3.2A**) [118,149].

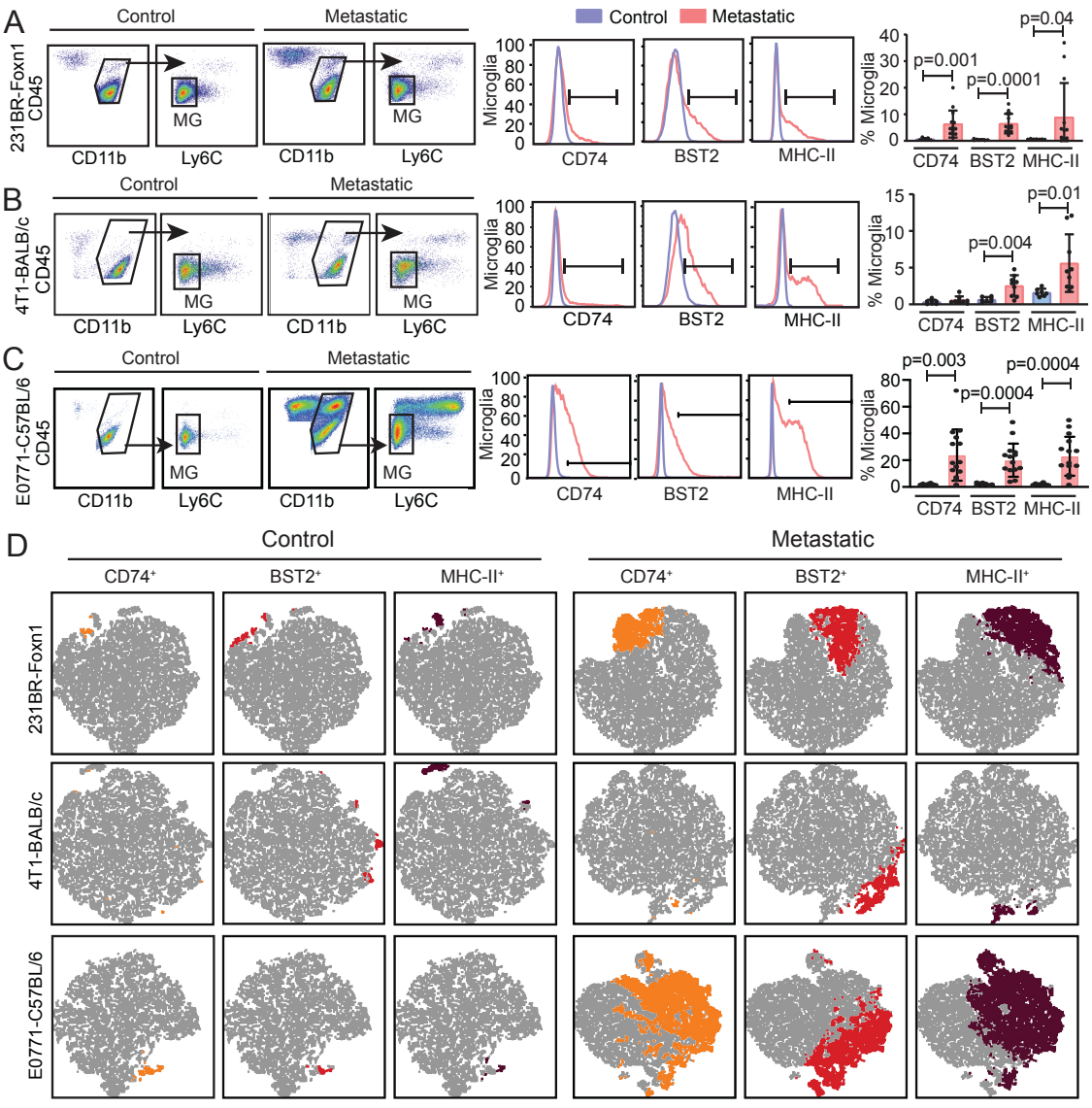

Figure 4.3.1: The microglia pro-inflammatory response is conserved in diverse BCBM models. (A) Flow cytometry analysis of CD74, BST2 and MHC-II in microglia harvested 28 days post intracardiac injection of 231BR (500,000) cells into Foxn1$^{nu/nu}$ animals. Representative plots show gating for single, live (Zombie^{neg}) CD45^{lo}CD11b^{+}Ly6C^{-} microglia (left panel) followed by analysis for CD74, BST2, and MHC-II (middle panel). Bar graph (right panel) shows the percent of microglia that express each marker in control (n=7) and metastatic (n=14) brains. *P* values were generated by an unpaired two-sided Student's *t*-test, and error bars indicate standard deviation. (B) Flow cytometry analysis of CD74, BST2 and MHC-II in microglia harvested 14 days post intracardiac injection of 4T1-GFP (100,000) cells into BALB/c animals. Representative plots are gated as in 3A. Bar graph (right panel) shows the percent of microglia that express each marker in control (n=7) and metastatic (n=7) brains. *P* values were generated by an unpaired two-sided Student's *t*-test, and error bars indicate standard deviation. (C) Flow cytometry analysis of CD74, BST2 and MHC-II in microglia harvested 14 days post intracranial injection of EO771-GFP (100,000) cells into C57BL/6 animals. Representative plots are gated as in 3A. Bar graph (right panel) shows the percent of microglia that express each marker in control (n=8) and metastatic (n=14) brains. *P* values were generated by an unpaired two-sided Student's *t*-test, and error bars indicate standard deviation. (D) Representative tSNE plots of microglia gated from (A-C). Colored cells indicate those gated as positive for CD74 (orange), BST2 (red) and MHC-II (brown) in the 231BR-Foxn1$^{nu/nu}$, 4T1-BALB/c and EO771-C57BL/6 models.

The immune competent models also displayed marked expansion of pro-inflammatory microglia. In the 4T1-BALB/c model, GFP-labeled 4T1 cells were injected i.c. and tissues were analyzed two weeks later (**Fig 4.3.2B,C**). IF analysis showed infiltration of metastatic lesions with IBA1^{+} cells similar to the 231BR model (**Fig 4.3.2D**). Analysis of CD45^{lo}CD11b^{+}Ly6C^{-} microglia by flow cytometry showed a 3.8-fold increase in the frequency of BST2 (p=0.004) and a 3.5-fold increase in MHC-II (p=0.01) in metastatic (n=9) versus control (n=7) brains, but no increase in the frequency of CD74^{+} cells (**Fig 4.3.1B**). In the EO771-C57BL/6 model, GFP-labeled EO771 cells were injected intracranially according to previously established protocols and analyzed two weeks later (**Fig 4.3.2E,F**)[121,123]. IF analysis showed similar infiltration of tumor lesions with IBA1^{+} cells (**Fig 4.3.2G**). However, flow cytometry analysis showed a remarkably robust response in this model, with >20 fold increase in CD74 (p=0.003), BST2 (p=0.0004), and MHC-II (p=0.0004) positive microglia from metastatic (n=14) versus PBS-injected control (n=8) brains (**Fig 4.3.1C**). These data show that the microglia pro-inflammatory response is conserved in three distinct BCBM

models. The less robust response observed in the 4T1-BALB/c model is consistent with prior reports of proclivity towards Th2 over Th1 immunity in the BALB/c background strain [150].

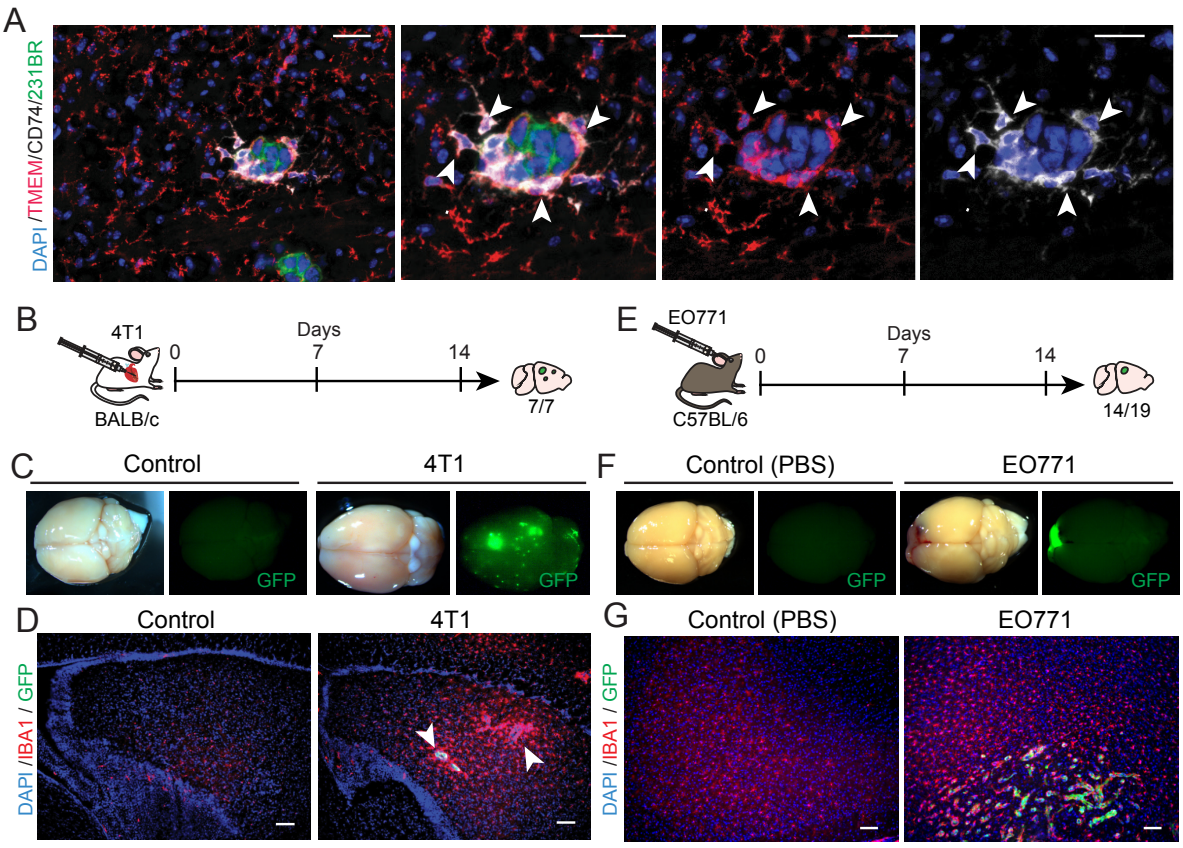

Figure 4.3.2: Disease progression and microglia activation in the 4T1-BALB/c and EO771-C57BL/6 models. (A) Representative images show IF analysis of CD74 in microglia from the 231BR-Foxn1$^{nu/nu}$ model (n=3). Arrowheads indicate CD74^{+} (white) and TMEM119^{+} (TMEM, red) microglia surrounding metastatic lesions (green). Scale bar= 50μm. (B) Schematic shows 4T1-BABL/c i.c. experimental metastasis model. 100% of animals (7/7) develop brain metastasis two weeks after injection of 100,000 GFP labeled 4T1 cells. (C) Whole mount brightfield and fluorescence microscopy images show metastatic lesions (green) in brains from representative control and metastatic animals. (D) Representative images show IF analysis for IBA1 in control and 4T1-BALB/c metastatic brains. White arrowheads indicate metastatic lesions (green) surrounded by IBA1^{+} (red) microglia. Scale bar = 100μm. (E) Schematic shows EO771-C57BL/6 intracranial injection model. 73% of animals (14/19) develop tumors 2 weeks after injection of 100,000 GFP labeled EO771 cells. (F) Whole mount brightfield and fluorescence microscopy images show tumors (green) in brains from representative control (PBS injected) and tumor-bearing (EO771) animals. (G) Representative images show IF analysis for IBA1 (red) in control and EO771-C57/BL/6 (green) injected brains. Scale bar = 100μm.

Finally, we investigated whether the microglia response to metastasis is homogeneous or heterogeneous by determining whether the protein markers are expressed by the same or different cells. We plotted microglia using tSNE to visualize the expression of BST2, MHC-II, and CD74 in each individual cell. Interestingly, this shows extensive overlap of the markers but also reveals subpopulations of microglia that express only one or two of the individual markers (**Fig 4.3.1D**). For example, in the EO771-C57BL/6 model, the IFN-response protein BST2 is only expressed by a subpopulation of $CD74^{+}MHC\text{-}II^{+}$ microglia. This shows that the microglia response to metastasis is heterogeneous, and raises the question of whether there are discrete substates of microglia that carry out distinct functions in BCBM.

4.2.4 BCBM-R microglia are heterogeneous and display specialized responses to metastasis

We further investigated heterogeneity within BCBM-R microglia at the whole transcriptome level using an iterative analysis of our single cell dataset (**Fig 4.4A**). To find conserved substates, we first performed sequencing batch correction using Seurat's integration protocol on myeloid cells from all three animals with BCBM and then unbiasedly clustered the integrated cells (**Fig 4.4A**) [135,136]. We next used the MG-score to discriminate microglia from other myeloid cells, and subsequently identified BCBM-R microglia by scoring for genes significantly upregulated in our metastatic condition compared to control (**Fig 4.4A**). This identified two clusters of BCBM-R microglia, which we extracted and further subclustered to investigate heterogeneity (**Fig 4.4A**).

Our iterative analysis revealed six distinct subpopulations of BCBM-R microglia, which we named Cycling, IFN responsive, APC, Secretory, Glycolytic, and Homeostatic (**Fig**

4.4B). The Homeostatic cluster was named as such because it displayed high levels of canonical microglia markers (**Fig 4.1.1A**) (*Tmem119, P2ry12*) and appeared similar to

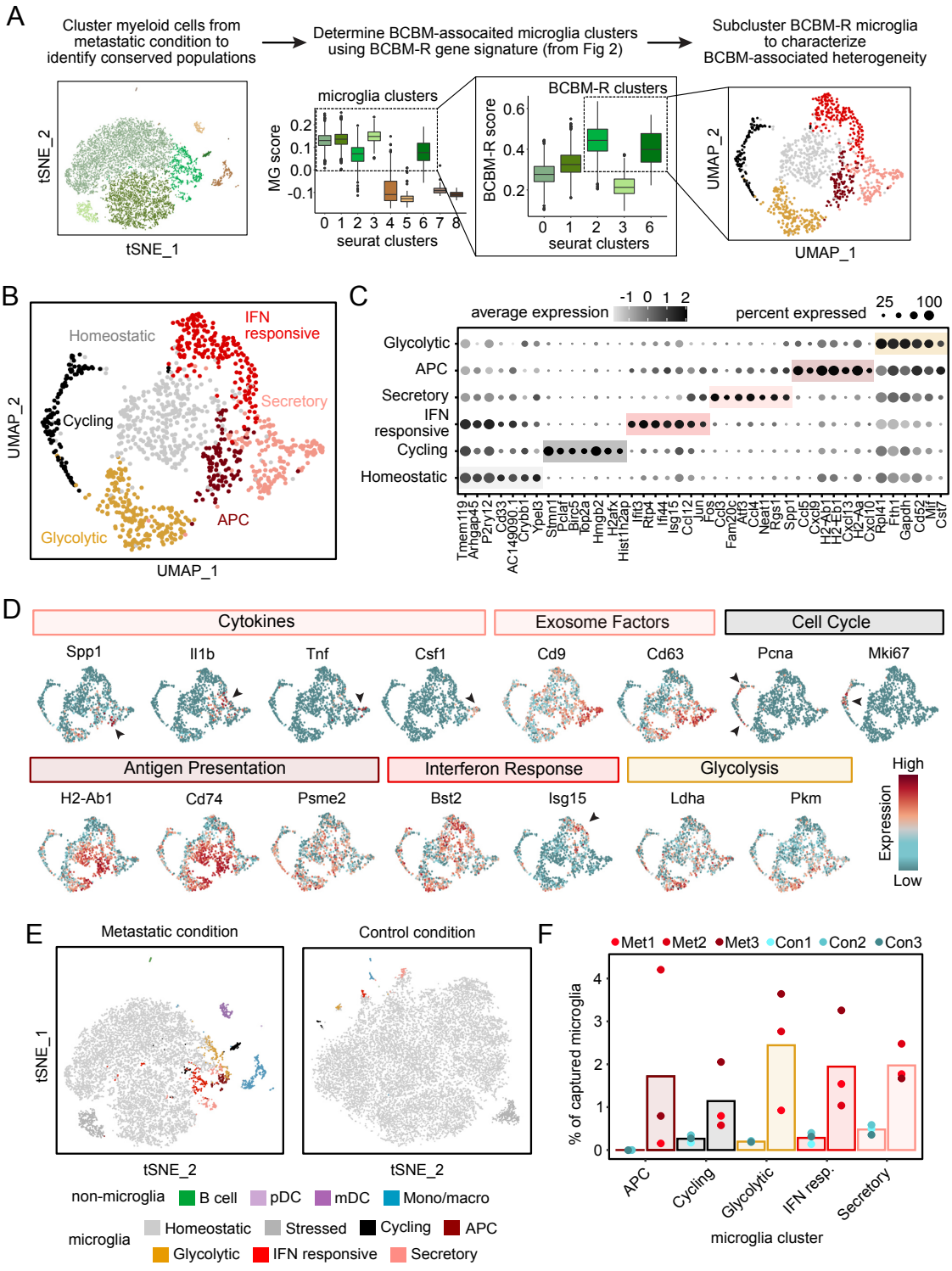

Figure 4.4: BCBM-R microglia are heterogeneous and display specialized responses to metastasis. (A) Schematic overview of iterative approach for BCBM-R microglia selection for subclustering analysis. Briefly, myeloid cells from mice with BCBM were reclustered and microglia were identified using the MG-score. Next, microglia were scored for the BCBM-R microglia signature to identify most robust responders. Finally, these microglia were separated and subclustered to investigate heterogeneity within the BCBM-R microglia. Note that Fig 2 refers to Fig 3.2.1 in this manuscript. (B) UMAP of BCBM-R microglia subpopulations, colored by cluster label. (C) Dot plot shows top marker genes for each cluster ranked by average natural logFC and determined by the Wilcoxon rank sum test. Dot size represents the percentage of cells that express the gene, and dot greyscale represents the average expression level. (D) Feature plots show relative expression in each cell for key marker genes associated with each BCBM-R microglia cluster. Arrows indicate regions of high expression for indicated genes. (E) tSNE plots show myeloid cells from metastatic (left) and control (right) animals, integrated by sequencing batch and colored by cell types and states. Control cell labels were determined using label transfer from the metastatic condition in Seurat v3. (F) Barplots show the average percentage of microglia in each subcluster that came from control and metastatic animals. Points represent individual mice

control microglia other than upregulation of MHC-II genes (**Fig 4.4C,D**). The Cycling cluster was marked by proliferation genes, such as *Top2a*, *Mki67, and Pcna* (**Fig 4.4C,D**). The IFN responsive (*Bst2*, *Ifitm3*, *Isg15*) and APC (*H2-Ab1*, *Cd74*, *Psme2*) clusters showed upregulation of the classic M1 pro-inflammatory genes identified in our BCBM-R signature (**Fig 4.4C,D**). The Secretory and Glycolytic clusters displayed unique expression programs not strongly captured by the BCBM-R signature. The Secretory cluster was marked uniquely by cytokines (*Spp1*, *Tnf*, *Il1b, Csf1*) and exosome factors (*Cd9*, *Cd63*) and shared markers of lipid metabolism (*Lpl*, *Apoe*) and phagocytosis (*Trem2*, *Tlr2*) with the APC subpopulation (**Fig 4.4C,D**). This suggests that the Secretory cluster may represent more classic microglia functions, which include supporting the local inflammatory environment with cytokines and phagocytosing dead or dying cells, leading to eventual antigen presentation. The Glycolytic cluster showed a shift towards increased glycolysis (*Pkm*, *Ldha, Gapdh*), which is a key feature of inflammatory macrophages and has been shown to increase metabolic output for microglia proliferation and cytokine production during neuroinflammation (**Fig 4.4C,D**) [151,152]. Interestingly, clustering of control microglia showed very limited heterogeneity, and label transfer in Seurat to the control condition showed that the BCBM-R subclusters were

specifically enriched in mice with BCBM with very few cells observed in control mice (**Fig 4.4E,F**). These data show that the microglia response to BCBM is heterogeneous, where microglia demonstrate distinct, specialized responses to metastasis that are unlikely to derive from pre-existing subtypes of microglia.

4.2.5 The pro-inflammatory response to BCBM is conserved in human microglia

Since it is difficult to study how microglia respond to metastasis initiation in BCBM patients, we developed a humanized mouse model of BCBM which allowed us to control the timing of tumorigenesis and investigate the full range of human microglia responses using scRNA-seq. We utilized the MITRG mouse model in which human *CSF1*, *IL3* and *TPO* are knocked into a *Rag2*−/−*Il2rg*-/- background to support the engraftment of human monocytes and macrophages [153]. In prior work, transplantation of human induced pluripotency-derived hematopoietic progenitor cells (iHPSCs) into the postnatal brain of MITRG mice was shown to result in differentiation into microglia and CNS macrophages [154,155]. We injected MITRG mouse pups with GFP-labeled iHPSCs, allowed engraftment for 10 weeks, and injected them i.c. with mCherry-labeled 231BR cells (**Fig 4.5.1A**). Control (n=3) and metastatic (n=3) mice were harvested three weeks later and analyzed by whole mount fluorescence microscopy, which confirmed the engraftment of GFP^+ human microglia and $mCherry^+$ 231BR metastases (**Fig 4.5.2A**). Dissociated cells from each sample were indexed using the MULTI-seq method and mouse cells were subsequently removed using anti-mouse MHC-I magnetic beads (**Fig 4.5.1A**) [156]. The remaining human cells, consisting of both myeloid and cancer cells, were then captured for sequencing using droplet-based technology (**Fig 4.5.1A**).

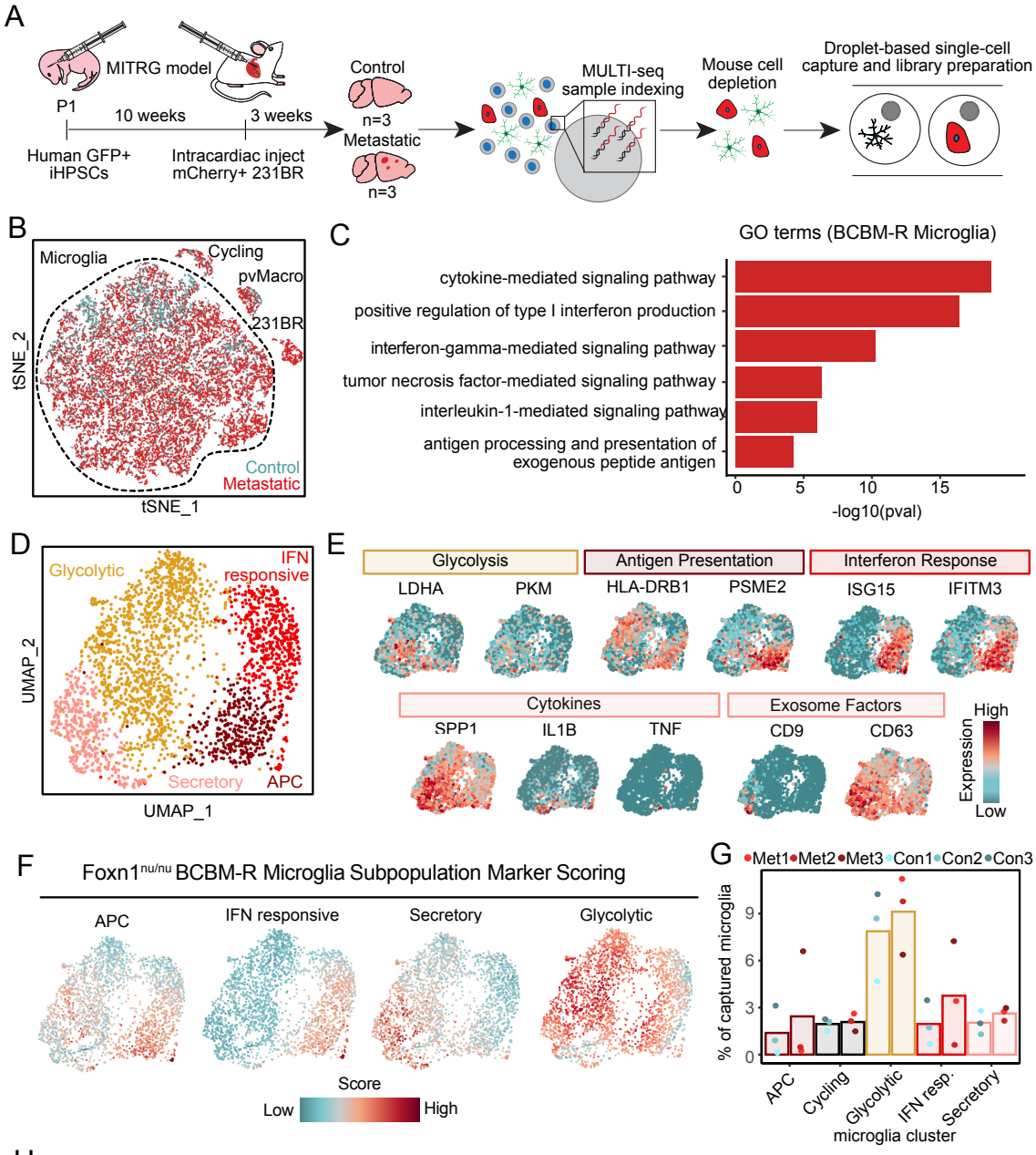
A
MITRG model
P1
10 weeks
3 weeks
Human GFP+ iHPSCs
Intracardiac inject mCherry+ 231BR
Control
n=3
Metastatic
n=3
MULTI-seq sample indexing
Mouse cell depletion
Droplet-based single-cell capture and library preparation
B
Microglia
Cycling
pvMacro
231BR
Control
Metastatic
tSNE_2
tSNE_1
C
GO terms (BCBM-R Microglia)
cytokine-mediated signaling pathway
positive regulation of type I interferon production
interferon-gamma-mediated signaling pathway
tumor necrosis factor-mediated signaling pathway
interleukin-1-mediated signaling pathway
antigen processing and presentation of exogenous peptide antigen
0
5
10
15
-log10(pval)
D
Glycolytic
IFN responsive
Secretory
APC
UMAP_2
UMAP_1
E
Glycolysis
Antigen Presentation
Interferon Response
LDHA
PKM
HLA-DRB1
PSME2
ISG15
IFITM3
Cytokines
Exosome Factors
SPP1
IL1B
TNF
CD9
CD63
High
Expression
Low
F
Foxn1nu/nu BCBM-R Microglia Subpopulation Marker Scoring
APC
IFN responsive
Secretory
Glycolytic
Score
Low
High
G
Met1 Met2 Met3 Con1 Con2 Con3
% of captured microglia
0
3
6
9
APC
Cycling
Glycolytic
IFN resp.
Secretory
microglia cluster
H
DAPI
GFP-hMG
mCherry-231BR
HLA-DR
Merge

Figure 4.5.1: The pro-inflammatory response to BCBM is conserved in human microglia. (A) Schematic shows experimental design for scRNA-seq of human microglia from humanized MITRG mice transplanted with 231BR cells. MITRG mouse pups were injected with GFP-labeled iHPSCs, aged to 10 weeks and injected i.c. with mCherry-labeled 231BR cells. Brains from control (n=3) and metastatic (n=3) mice were digested to make single cell suspensions three weeks later. Dissociated cells from each sample were indexed using the MULTI-seq method. Mouse cells were removed using anti-mouse MHC-I magnetic beads, and recovered cells were collected and pooled into two samples for scRNA-seq, metastatic and control. (B) tSNE plot shows cells (n=21,353) colored by condition and labeled by cell type. pvMacro=perivascular macrophages. (C) Bar plot shows selected top GO terms identified for microglia from metastatic (n=4,146 genes, adj. $p<0.05$) brains. GO terms were determined using MouseMine and select terms with Holm-Bonferonni adjusted *P* values <0.05 were retained. (D) UMAP of BCBM-R microglia, colored by cluster label. BCBM-R microglia were identified for subclustering analysis using the iterative approach described in **Fig 4.4A**. (E) Feature plots show relative expression in each cell for key marker genes associated with each BCBM-R microglia cluster. (F) UMAP plots show similarity of human and mouse microglia substates by gene scoring analysis. Each human cell from (D) was scored for gene signatures for the mouse microglia substates identified in **Fig 4.4**. Scores were calculated using the AddModuleScore function in Seurat. Gene signatures were translated from mouse to human using the biomaRt package in R. See **Appendix B**. (G) Barplots show the average percentage of microglia in each labeled cluster that came from control and metastatic animals. Points represent individual mice (H) Representative images showing IF analysis of HLA-DR (white) in human microglia (green) near 231BR metastatic cells (mCherry) in transplanted MITRG mice from (A). Scale = 1000μm. hMG=human microglia.

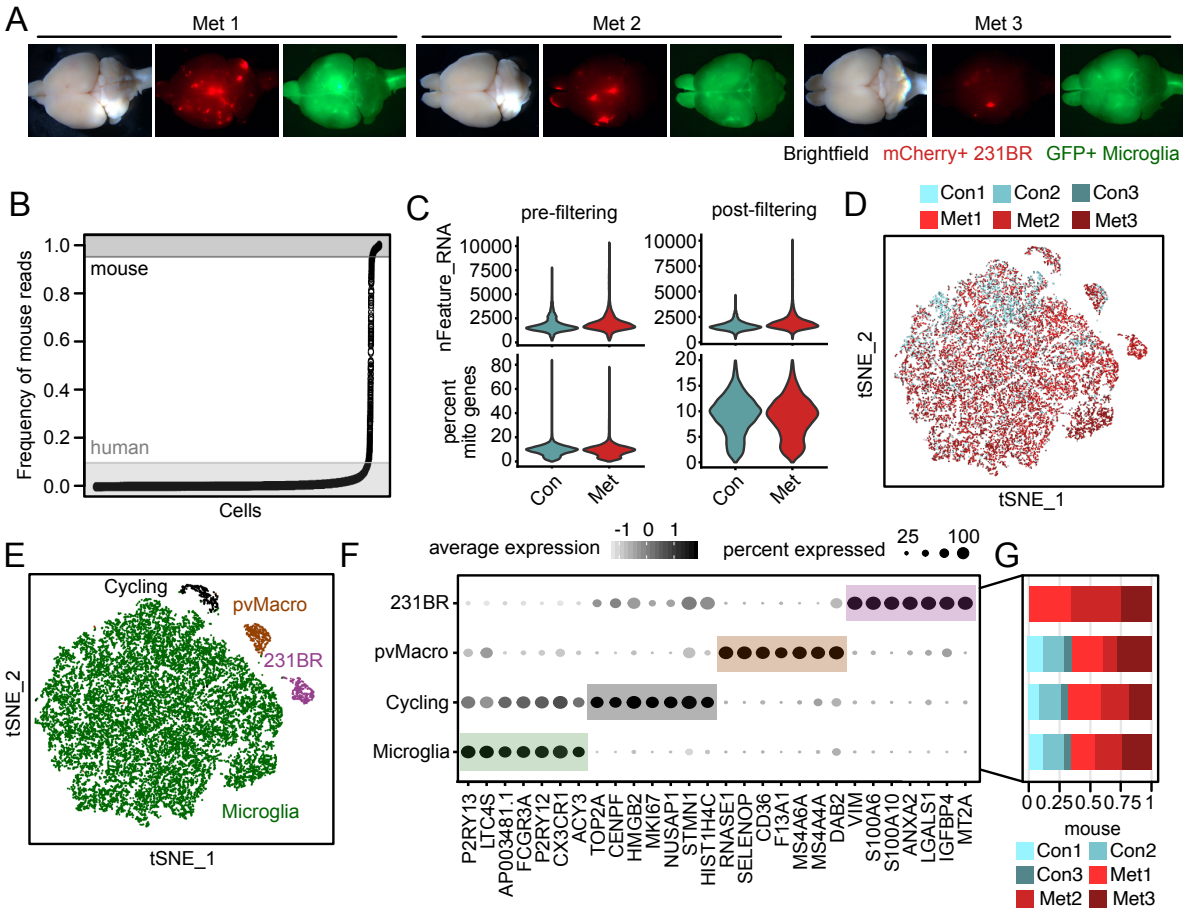

Figure 4.5.2: Experimental design, quality control and cell type identification for scRNA-seq cell libraries from transplanted MITRG mice. (A) Whole mount brightfield and fluorescence microscopy images show brains from MITRG mice transplanted with GFP-labeled iHPSC cells and mCherry-labeled 231BR cells. See **Fig 4.5.1A**. (B) Identification of mouse and human cells by the frequency of reads that align to the mm10 mouse genome. Cutoffs used to identify mouse cells (>0.95 aligned, n=641 cells), human cells (<0.1 aligned, n=25,287 cells) and doublets (0.1-0.95 aligned, n=387 cells) are shown. (C) Violin plots show cell distributions for key quality control metrics pre- and post- filtering and removal of poor quality cells. Cells were removed that displayed >20% of genes mapped to the mitochondrial genome (percent mito genes). (D) tSNE plot shows clustering of human cells (n=21,353) from MITRG brains colored by mouse ID. Mouse ID was assigned to each cell based on MULTI-seq barcode analysis. (E) tSNE plot shows human cells, colored by cluster and labeled by cell type. pvMacro=perivascular macrophages, Cycling = cycling myeloid cells. (F) Dot plot shows top marker genes for each cell type determined by the Wilcoxon rank sum test and ranked by average natural logFC. Dot size represents the percentage of cells that express the gene, and dot greyscale represents the average expression level. pvMacro=perivascular macrophages. (G) Bar chart shows the frequency of cells contributed by each mouse to the cell types shown in (F).

Human cells were further distinguished bioinformatically by aligning to a merged human (GRCh38) and mouse (mm10) genome, which identified 25,287 human cells (**Fig 4.5.2B**). Cells that were identified as doublets, contained no MULTI-seq index, or displayed a mitochondrial gene percentage >20% were removed from downstream analysis (**Fig 4.5.2C**). Clustering and marker gene analysis of the 21,353 human cells that passed filtering showed limited batch effects (**Fig 4.5.2D**), and revealed a distinct population of 231BR cells *(VIM)* and several populations of myeloid cells (**Fig 4.5.1B**, **Fig 4.5.2E-G**). These included clusters of human perivascular macrophages (*CD163*), microglia (*TMEM119*), and a population of proliferating myeloid cells (*MKI67*) (**Fig 4.5.2E-G**).

Supervised analysis of genes differentially expressed between human microglia from the control and metastatic conditions revealed GO terms similar to the mouse BCBM-R signature, including cytokine response, interferon response, and antigen presentation (**Fig 4.5.1C**). Strikingly, subclustering of the BCBM-R human microglia also revealed similar substates as observed in mouse microglia (**Fig 4.5.1D**). Using the same iterative analysis as described in **Fig 4.4A**, we identified four distinct subclusters marked by the same top genes that delineated APC (*HLA-DRB1, PSME2*), IFN responsive (ISG15, IFITM3), Secretory (*SPP1, IL1B, CD9, CD63*), and Glycolytic (*LDHA, PKM*) microglia in the mouse (**Fig 4.5.1D,E**). Gene scoring for subpopulation markers of each mouse microglia substate further supported this finding and showed that signatures derived from mouse BCBM-R microglia can be directly applied to human microglia to determine their phenotypic state (**Fig 4.5.1F**, **Appendix B**). However, the BCBM-R substates showed less relative expansion in the human than the mouse models (**Fig 4.5.1G**). IF staining for the APC gene HLA-DR confirmed upregulation at the protein level, and showed the response is strongest in microglia proximal to BCBM

lesions (**Fig 4.5.1H**). Overall, these data show that human and mouse microglia demonstrate similar pro-inflammatory responses to BCBM and suggest that human microglia may have the same capacity to respond to metastasis initiation in BCBM patients.

4.2.6 Microglia demonstrate a potent tumor suppressive effect on BCBM initiation

Prior work using pharmacologic and genetic depletion strategies has shown a clear pro-tumorigenic role for TAMs in BCBM and CNS cancers [122–125]. These studies primarily utilized CSF1R inhibitors and CX3CR1-targeted genetic ablation strategies that can target microglia as well as other TAM populations, leaving the specific role of microglia unclear [123,125,157,158]. A new genetic model was recently developed that specifically and completely lacks microglia due to deletion of a newly discovered, highly conserved super-enhancer in the CSF1R locus called the fms-intronic regulatory element (FIRE) (**Fig 4.6.1A**) [127]. The *Csf1r*$^{\Delta FIRE/\Delta FIRE}$ (FIRE-KO) model lacks microglia while retaining BAMs and BMDMs, making it an important new tool to specifically explore microglia function in disease [127].

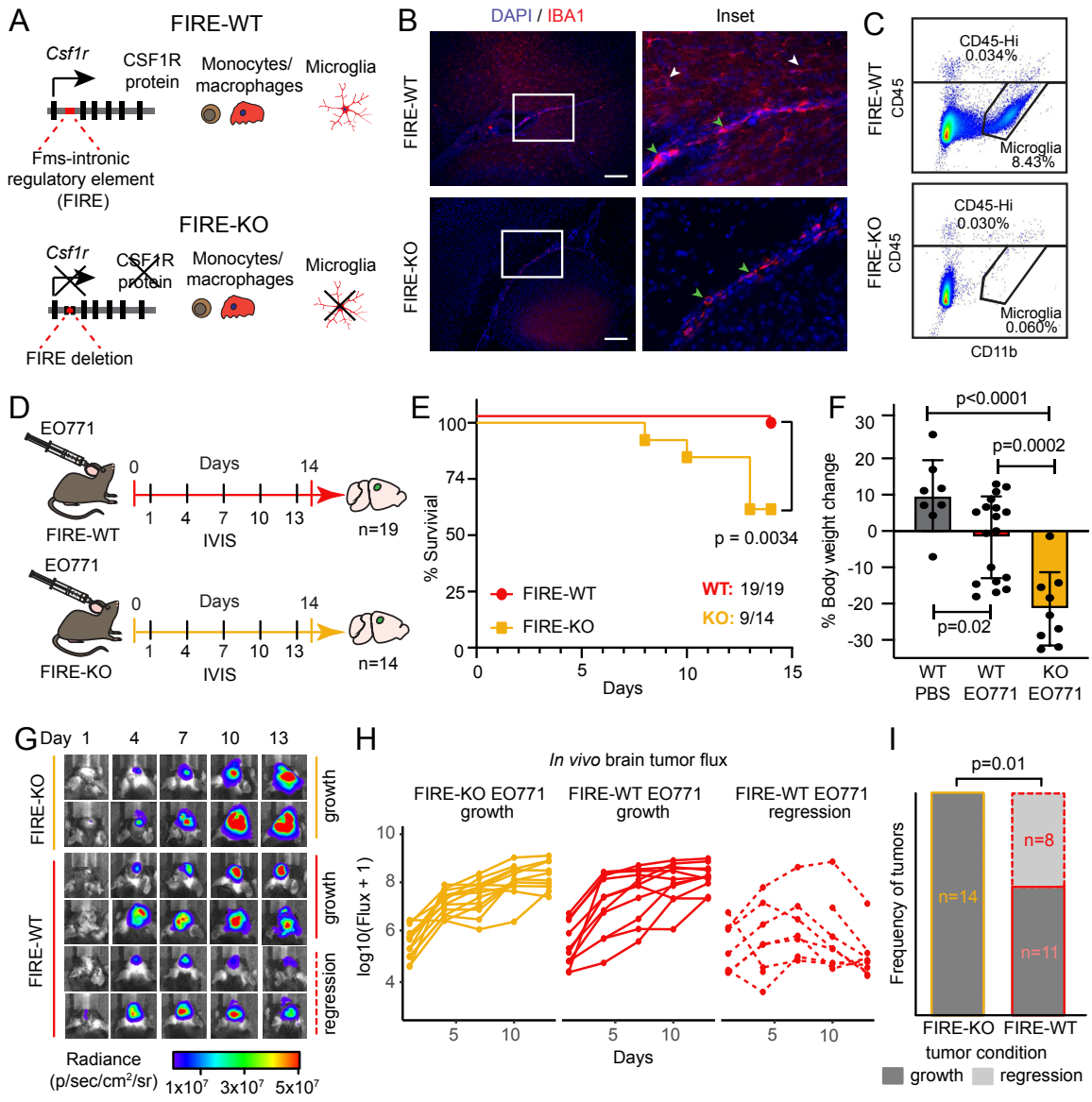
A
FIRE-WT
Csf1r
CSF1R protein
Monocytes/ macrophages
Microglia
Fms-intronic regulatory element (FIRE)
FIRE-KO
FIRE deletion
B
DAPI / IBA1
Inset
FIRE-WT
FIRE-KO
C
CD45-Hi 0.034%
Microglia 8.43%
CD45-Hi 0.030%
Microglia 0.060%
CD45
CD11b
D
EO771
FIRE-WT
FIRE-KO
Days
IVIS
n=19
n=14
E
% Survivial
FIRE-WT
FIRE-KO
p = 0.0034
WT: 19/19
KO: 9/14
Days
F
% Body weight change
p<0.0001
p=0.0002
p=0.02
WT PBS
WT EO771
KO EO771
G
Day 1 4 7 10 13
growth
regression
Radiance (p/sec/cm²/sr)
1x10⁷ 3x10⁷ 5x10⁷
H
In vivo brain tumor flux
FIRE-KO EO771 growth
FIRE-WT EO771 growth
FIRE-WT EO771 regression
log10(Flux + 1)
Days
I
p=0.01
n=14
n=8
n=11
Frequency of tumors
FIRE-KO FIRE-WT
tumor condition
growth regression

Figure 4.6.1: Microglia demonstrate a potent tumor suppressive effect on BCBM initiation. (A) Schematic depiction of *Csf1r*$^{\Delta FIRE/\Delta FIRE}$ mouse model. Deletion of FIRE super-enhancer in FIRE-KO mice leads to loss of CSF1R protein expression and lack microglia development. (B) IF staining shows IBA1^{+} cells in FIRE-WT and FIRE-KO mouse brains. Green arrows show choroid plexus macrophages in FIRE-KO and FIRE-WT, and white arrows show microglia only in FIRE-WT. Scale bar = 50μm. (C) Representative flow cytometry plots show the percentage of CD45^{lo}CD11b^{+} microglia and CD45^{hi} immune cells gated from live (sytox negative), single cells in FIRE-WT (n=2) and FIRE-KO (n=2) mouse brains. (D) Schematic of experimental design to compare disease progression in FIRE-WT and FIRE-KO mice. FIRE-WT (red, n=19) and FIRE-KO (yellow, n=14) mice were injected intracranially with 100,000 GFP and luciferase labeled EO771 cells. Control FIRE-WT mice (n=8) were also injected with PBS. Animals were imaged for luminescence (IVIS) every three days before dissection at endpoint on day 14. (E) Kaplan-Meier plot shows survival in FIRE-WT (19/19, 100%) and FIRE-KO (9/14, 64%) mice from (D). *P* value determined by log-rank (Mantel-Cox) test. (F) Bar graph shows percentage body weight change for each PBS injected (n= 8), FIRE-WT (n=19), and FIRE-KO (n=9) animal from (D) at day 14 relative to day 0. *P* values determined by unpaired two-sided Student's *t*-test and error bars represent standard deviation. (G) IVIS images show luminescence signal change over time in FIRE-WT and FIRE-KO animals from (D). Representative animals that displayed continuous signal increase (tumor growth, solid line) vs. signal decrease (tumor regression, dashed line) are shown. Pseudocoloring of luminescence shows quantification of radiance (p/sec/cm^2/sr). (H) Line graphs show quantification of luminescence signal change over time in all FIRE-WT and FIRE-KO animals from (D). Solid lines indicate animals that demonstrated tumor growth and dashed lines indicate those that showed tumor regression. Growth was defined by signal increase over time, and regression was defined as either baseline signal (<10^6) or >5-fold decrease in signal relative to maximum. (I) Bar graph summarizes the frequency of animals that displayed tumor growth and tumor regression in FIRE-WT and FIRE-KO backgrounds. *P* value was determined by Fisher's exact test.

We investigated the role of microglia in BCBM using FIRE mice and the E0771 model. We first compared the immune composition of FIRE-KO and FIRE-WT mice. IF and flow cytometry analysis confirmed a complete absence of microglia and retention of BAMs in non tumor-bearing FIRE-KO animals (**Fig 4.6.1B,C**). We observed the same phenomenon in tumor-bearing animals by scRNA-seq. CD45^{+} cells from E0771-injected FIRE-WT (n=4) and FIRE-KO (n=4) brains were isolated by flow cytometry, pooled and captured for sequencing using droplet-based technology. Clustering and marker gene analysis of the 10,827 cells that passed quality control filtering identified 11 immune cell types, including one cluster of microglia (**Fig 4.6.2A**). As expected, microglia were only observed in FIRE-WT mice (**Fig 4.6.2A,C**). The proportions of other immune cell types were also not skewed between FIRE-WT and FIRE-KO mice, excluding this as a confounding variable in future experiments (**Fig 4.6.2D**). Importantly, this contrasts with the Cx3cr1$^{CreERT/+}$:ROSA26i$^{DTR/+}$ model used by

Guldner et al (2020) which found a pro-tumorigenic function for TAMs in BCBM. Reanalysis of scRNA-seq data from these mice showed retention and potentially enrichment of microglia in depleted vs. control animals (**Fig 4.6.2B,E**). Depleted animals also demonstrated a decrease in the proportion of macrophages and Ly6c^{hi} monocytes (**Fig 4.6.2E**). These data confirm the specificity of microglia depletion in FIRE-KO mice and emphasize the value of the model for studies of microglia function.

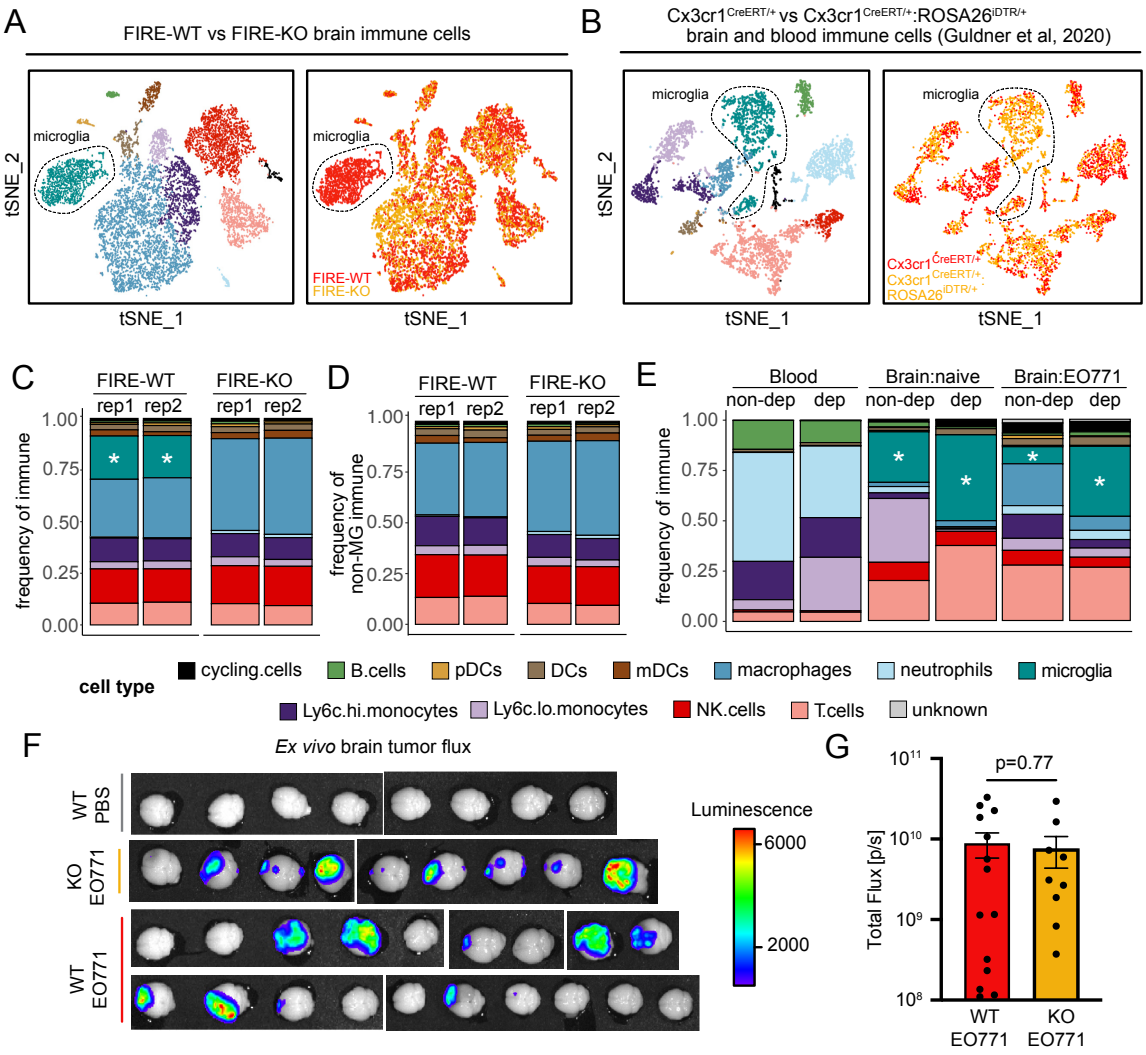

Figure 4.6.2: Analysis of immune cell composition and tumor burden in FIRE-WT and FIRE-KO animals. (A) ScRNA-seq analysis of immune cell composition in FIRE-WT and FIRE-KO mice. FIRE-WT (n=4) and FIRE-KO (n=4) mice were injected intracranially with 100,000 EO771 cells and harvested 14 days later. Immune cells ($CD45^+$) were isolated by flow cytometry and pooled for droplet-based capture and sequencing. The genotype of each cell was determined using Souporcell. tSNE plots (n=10,827 cells) show cells labeled by cell type (left) and condition (right). (B) tSNE plots show reanalysis of single cell CITE-seq data collected from $Cx3cr1^{CreERT/+}$:$ROSA26i^{DTR/+}$ (CNS-myeloid depleted) and $Cx3cr1^{CreERT/+}$ (control non-depleted) mice as described in Guldner et al, (2020) (GSE139971). Left plot shows cells colored by cell type, which was assigned by label transfer from the FIRE dataset in (A). Right plot shows cells colored by genotype using Souporcell, which were cross-referenced with HTO-barcodes from the CITE-seq dataset. (C) Barplot shows the frequency of each immune cell type captured from FIRE-WT and FIRE-KO mice in the scRNA-seq experiment from (A). Stars indicate microglia. Replicates 1 and 2 indicate two separate pooled samples generated for droplet capture (rep1, n=5,466; rep2, n=5,361). (D) Barplot shows the frequency of each immune cell type as in (C) but excluding microglia (MG). (E) Barplots show the frequency of each immune cell type out of total immune cells captured from the blood and brain of $Cx3cr1^{CreERT/+}$:$ROSA26^{iDTR/+}$ (CNS-myeloid depleted, dep) and $Cx3cr1^{CreERT/+}$ (non-depleted, non-dep) mice from Guldner et al. (2020) (GSE139971) as shown in (B). Cells were assigned to mouse stains and tissue type based on HTO-barcodes from CITE-seq dataset. (F) Ex vivo whole brain luminescence of FIRE-KO (n=9) and FIRE-WT (n=19) mice that survived to endpoint (day 14) relative to PBS injected controls. Pseudocoloring of luminescence is shown in counts(p/s). (G) Barplots show the quantification of total flux for brains shown in (F). *P* value was determined by unpaired two-sided t-test.

We next compared BCBM progression in FIRE-KO and FIRE-WT mice. FIRE-WT (n=19) and FIRE-KO (n=14) mice were injected with GFP and luciferase-labeled EO771 cells and monitored by IVIS every three days until endpoint at day 14 (**Fig 4.6.1D**). Surprisingly, many FIRE-KO mice quickly developed overt clinical symptoms of advanced disease (**Fig 4.6.1E,F**). Five of 14 FIRE-KO mice died before endpoint (36% mortality), while all 19 FIRE-WT survived (0% mortality) (p=0.0034) (**Fig 4.6.1E**). Surviving FIRE-KO mice displayed >20% decrease in body mass compared to FIRE-WT (p=0.0002), also indicating increased morbidity in mice lacking microglia (**Fig 4.6.1F**). Analysis of tumor growth over time by IVIS revealed interesting differences in the kinetics of tumor progression between FIRE-WT and FIRE-KO animals (**Fig 4.6.1G-I**). After initial engraftment, we observed a decrease in luciferase signal in eight of 19 FIRE-WT mice over time (42% mice decrease), while signal continued to increase in all 14 FIRE-KO animals (0% mice decrease) (p=0.01036) (**Fig 4.6.1G-I**). This indicates that tumors regress in FIRE-WT but not FIRE-KO animals,

suggesting microglia suppress BCBM specifically through tumor rejection. Consistent with this hypothesis, ex vivo analysis of tumors at endpoint confirmed the presence of tumors in 9/9 of the surviving FIRE-KO animals (eight parenchymal and one meningeal) and 14/19 FIRE-WT mice (**Fig 4.6.2F**). But no difference in tumor size was observed between the groups (p=0.77), showing that the absence of microglia attenuates the animal's capacity for tumor rejection rather than for slowing tumor growth (**Fig 4.6.2G**). Taken together with previous work, these findings suggest that microglia are innately pro-inflammatory and tumor suppressive, while other TAM populations are anti-inflammatory and tumor promoting. This highlights the importance of developing therapeutic approaches that target specific TAM populations in order to effectively treat BCBM and other CNS cancers.

4.3 Discussion

We utilized a diverse array of approaches to investigate the role of microglia in the development of BCBM. ScRNA-seq analysis of whole transcriptome profiles allowed us to discriminate brain resident microglia from other types of TAMs, a longstanding challenge in CNS diseases. This enabled us to discover that microglia upregulate a pro-inflammatory (M1) response to BCBM, in contrast to prior dogma that they and other TAMs favor an alternatively activated (M2) response. We further describe heterogeneity within pro-inflammatory microglia, where we find that distinct subpopulations of microglia upregulate programs for proliferation, IFN response, cytokine and exosome secretion, glycolysis, and antigen presentation. We validated the pro-inflammatory response at the protein level in three distinct models of BCBM, as well as in human microglia, showing the response is highly conserved and highlighting its relevance in human BCBM. Most importantly, we utilized the

newly developed, genetically engineered FIRE-KO mouse model that completely and specifically lacks microglia to investigate their impact on metastasis. We find that animals lacking microglia develop more metastasis and less tumor regression than controls. Together with scRNA-seq data, these results show that microglia are pro-inflammatory and anti-tumorigenic, and suggest that microglia suppress metastasis through facilitating tumor rejection. This contrasts with the anti-inflammatory, pro-tumorigenic role previously ascribed to microglia and other TAMs, and raises the prospect of augmenting their tumor suppressive function to treat BCBM.

An interesting phenomenon emerging from our study is a clear conservation of microglia response phenotypes to different diseases of the CNS. Several of the microglia substates we observed in BCBM have been recently reported in single cell studies of microglia in other CNS diseases. In a study comparing microglia diversity during development, aging, and demyelinating injury, Hammond et al. (2019) identified subpopulations of cells that resemble our Glycolytic (*Mif*, *Pkm*), Secretory (*Lpl*, *Ccl4*), IFN responsive (*Ifitm3*, *Isg15*), and Cycling (*Pcna*, *Mki67*) microglia [159]. Interestingly, specific subpopulations of microglia were preferentially found in different contexts; Glycolytic and Secretory were predominantly found in the developing brain, while IFN response and Cycling microglia were enriched in aging and injury. Similar response phenotypes have also been observed in neurodegenerative diseases. Keren-Shaul et al (2017) found that disease-associated microglia (DAMs) in Alzheimer's Disease (AD) and amyotrophic lateral sclerosis (ALS) upregulate the phagocytosis and lipid metabolism associated genes *Apoe*, *Cst7*, *Cd9*, and *Lpl,* which are markers of Secretory microglia in our dataset [160]. They show that DAMs phagocytose plaques and are protective against AD development, suggesting Secretory

microglia may perform similar functions in BCBM. Another study of AD found that microglia display distinct phenotypes at progressive stages of disease development, where early response (ER) microglia upregulate markers of proliferation while late response (LR) microglia upregulate IFN response and antigen presentation genes [161]. ER and LR microglia share many markers in common with Cycling, IFN responsive, and APC microglia in our study. The overlap of markers between ER and LR microglia led the authors to suggest that they represent progressive stages of microglia activation in response to disease development, raising the question of whether the microglia substates identified in our study also represent different stages of temporal activation. An alternative hypothesis could be that the substates represent microglia responding to different local stimuli that elicit distinct responses. An important question for future studies will be to determine whether the microglia subpopulations carry out distinct functions, and what their independent effects on metastasis are.

Another fascinating phenomenon revealed by our study is the opposing outcomes achieved using different TAM depletion strategies. While we find increased tumor progression in FIRE-KO mice lacking microglia, previous work clearly showed decreased tumor progression following TAM depletion using CSF1R inhibitors and CX3CR1-targeted genetic ablation models [123,125] .There are several possible explanations for these discrepant findings. It is clear that microglia depletion in the FIRE-KO model is more complete and restricted to microglia than other approaches [123,125,162]. Furthermore, microglia cannot rebound and repopulate the brain in FIRE-KO mice as has been observed in other depletion models. The massive cell death produced in the $Cx3cr1^{CreERT/+}$:$ROSA26i^{DTR/+}$ depletion model has also been shown to induce cytokine storm and astrogliosis, which may have confounding

effects on tumor growth and the immune response [163,164]. Another important distinction in our study is that the FIRE-KO mice lack microglia from birth, while most prior studies targeted TAMs postnatally and after tumor initiation. It is therefore plausible that the timing of depletion impacts the outcome, as microglia and TAMs may become tumor promoting as disease progresses. Regardless, further investigation is critical given that several CSF1R inhibitors targeting TAMs are currently in clinical trials (e.g., NCT02829723, NCT01596751, NCT02401815, NCT02584647) for the treatment of CNS and peripheral cancers. An unintended side effect may be the depletion of protective microglia in the CNS and the creation of a permissive microenvironment for the development of CNS metastasis.

Finally, it will be important in future work to investigate the mechanism by which microglia suppress BCBM outgrowth. Since the absence of microglia in FIRE-KO mice specifically compromises their capacity for tumor rejection, it is compelling to consider that their effect is mediated through promotion of an anti-tumor T cell response. BCBM-R microglia secrete several chemokines that may promote T cell trafficking, as well as pro-inflammatory cytokines that may sustain T cell function once in the brain. BCBM-R microglia also upregulate MHC-I and MHC-II and other machinery for antigen presentation, which could enable them to present tumor neoantigens to CD4 or CD8 T cells in the brain. Previous work has shown that local APCs are critical to skew T cell differentiation and sustain their activation after arrival to the inflamed tissue, so it is reasonable that microglia function as the predominant purveyors of this function in the CNS [165]. Microglia function could therefore be critical for the efficacy of checkpoint inhibitors in CNS cancers, and boosting their function with macrophage targeting agents such as CD40 agonists could provide further therapeutic benefit.

4.4 Materials & methods

Normal human brain and human BCBM samples

FFPE sections of deidentified normal human brain and resected breast cancer brain metastasis were acquired from University of California Irvine department of Pathology and Laboratory Medicine, experimental tissue shared resource facility and the University of California Davis Pathology Biorepository.

Cell lines

MDA231-BrM (referred to as 231BR) [129] cells stably transduced with membrane targeted AcGFP (rLV.EF1.AcGFP1-Mem-9, ClonTech/Takara Bio, USA), mCherry (rLV.EF1.mCherry-9, ClonTech/Takara), and luciferase lentivirus were a generous gift from Ian Smith (Parker, 2017, Bos, 2009). 4T1 cells were purchased from ATCC (CRL-2593), stably infected with GFP lentivirus (Santa Cruz Biotechnology, copGFP Control Lentiviral Particles) at a MOI or 10, and sorted for GFP expression after two weeks growth in culture. EO771 cells were purchased from CH3 Biosystems and stably infected with pCDH-EF1a-eFFly-eGFP) lentivirus particles. pCDH-EF1a-eFFly-eGFP was a gift from Irmela Jeremias (Addgene plasmid # 104834 ; http://n2t.net/addgene:104834 ; RRID:Addgene_104834). To produce lentiviral particles, HEK293T cells were transfected with pCDH-EF1a-eFFly-eGFP together with pMD2G and psPAX2 packaging plasmids using Lipofectamine 2000 (Invitrogen). Supernatants containing lentiviral particles were used to infect EO771 cells overnight in the presence of 8 μg/ml polybrene (Sigma-Aldrich). Transduced EO771 cells were sorted on the basis of GFP expression on a BD FACSAria Fusion cell sorter. MDA213-BRm and 4T1 cell lines

were cultured in DMEM, 5% FBS, 10U/ml penicillin, 0.1mg/mL streptomycin, at 37 °C, 5% CO_2, 95% relative humidity. EO771 cells were cultured in RPMI 1640, 5% FBS, 10U/ml penicillin, 0.1mg/mL streptomycin, 10mmol/L HEPES at 37 °C, 5% CO_2, 95% relative humidity. Cells were passaged for one-two weeks prior to intracardiac or intracranial injections.

Mouse strains

Female Foxn1$^{nu/nu}$ mice (stock #007850), C57BL/6J (stock #000664), and BALB/cJ (stock #000651) were purchased from The Jackson Laboratories. Female MITRG mice (Jackson laboratories stock #017711) which are C:129S2- Rag2$^{tm1.1Flv}$ Csf1$^{tm1(CSF1)Flv}$ CSF2/IL3$^{tm1.1(CSF2,IL3)Flv}$ Thpo$^{tm1.1(TPO)Flv}$ Il2rg$^{tm1.1Flv}$/J were bred, housed and maintained by the laboratory of Mathew Blurton-Jones (IACUC protocol #AUP-17-162). *Csf1r*$^{\Delta FIRE/\Delta FIRE}$ (FIRE-KO) and *Csf1r*$^{FIRE/FIRE}$ (FIRE-WT) mice were a gift from Claire Pridans and Mathew Blurton-Jones laboratories and were housed and maintained by the Lawson laboratory. All animal studies were performed in accordance with an IACUC approved protocol #AUP-19-051 at the University of California Irvine.

Immunofluorescence analysis of human BCBM samples

4-µm sections were heated at 65 °C for 30 min, then deparaffinized by two sequential five-min incubations in Histo-Clear (National Diagnostics, #HS-200, Atlanta, Georgia, USA). Tissues were rehydrated with graded solutions of ethanol (100%-50%) and washed in double-distilled H_2O and 1XPBS. Antigen retrieval was performed using a microwave pressure cooker with 10 mM citric acid buffer (0.05% Tween 20, pH 6.0). Tissues were

blocked in blocking solution (0.1% Tween 20 and 10% Goat Serum in PBS) for 30 min at room temperature, incubated with primary antibodies diluted in blocking solution at 4 °C overnight, washed in PBS, incubated with secondary antibodies diluted in blocking solution for one hour at room temperature, and washed in PBS. Slides were mounted with VECTASHIELD Antifade Mounting Medium with DAPI (Vector Laboratories, #H-1200, Burlingame, California, USA) and micrographs were taken with the BZ-X700 Keyence fluorescence microscope.

Generation of BCBM in mice

For intracardiac injection to establish brain metastasis, as previously described by [166], cells were injected into the left cardiac ventricle of anesthetized mice (300mg/kg Avertin). For 231BR brain metastasis 500,000 cells in 100μL of DPBS were injected into nine week old Foxn1$^{nu/nu}$ or 10 week old MITRG mice. For 4T1 brain metastasis, 100,000 cells were injected into 9 week old BALB/cJ mice in 100μL of DPBS. For the intracranial injection of C57BL/6J, 100,000 EO771 cells in a volume of 10μL PBS were injected to a depth of 3mm into the right coronal suture of five week old mice [121,123]. Control mice were injected with 10μL PBS.

Dissection and visualization of mouse BCBM by whole mount fluorescence microscopy

At endpoint, mice were euthanized and perfused with 50mL of sterile ice cold 1X PBS, 1mg/mL EDTA. The brain was dissected from the cranium and meninges, and then washed in ice cold sterile 1X PBS. To visualize metastasis prior to RNAseq, flow cytometry analysis,

or fixation, the whole brain was placed on the dissection microscope (Leica Biosystems, DMC 2900) and imaged for GFP fluorescence and brightfield.

Mouse brain fixation and sectioning

Dissected brains were drop fixed into 4% PFA, 1X PBS, pH 7.4 overnight at 4°C. Fixed brains were transferred into 30% Sucrose 1X PBS for 24 hours prior to cryosectioning on sliding microtome (Leica Biosystems, SM2010R). Brains were frozen onto the stage for sagittal or coronal sectioning at 40µm thickness using dry ice powder. Serial slices were collected into 1X PBS, 0.05% sodium azide and stored at 4°C for floating section immunostaining.

Immunofluorescence staining of floating sections

Brain slices were transferred into a well of a 24 well plate containing 300µL of blocking solution (1X PBS, 5% serum, 0.3% tritonX-100) and placed on an orbital shaker for one hour. Blocking solution was removed and replaced with 500µL of primary antibody diluted in blocking solution and incubated overnight on an orbital shaker at 4°C. The next day, primary antibody was removed, and brain slices were washed with three sequential 500µL washes of blocking solution and incubated with secondary antibody for one hour at room temperature. Brain slices were transferred to a glass slide and mounted with VECTASHIELD Antifade Mounting Medium with DAPI (Vector Laboratories, # H-1200, Burlingame, California, USA). Micrographs were taken with the BZ-X700 Keyence fluorescence microscope and acquisition software.

Primary antibodies: Rabbit polyclonal anti-IBA1 diluted 1:500 (Wako #019-19741); rat anti-CD74 clone ln1/Cd74 diluted 1:100 (BioLegend #151002); rabbit monoclonal anti-

TMEM119 clone 28-3 diluted 1:500 (Abcam #ab208064); anti -Human HLA-DRB clone LN3 diluted 1:200 (Invitrogen,, REF#14-9956-82). Secondary Antibodies, diluted 1:400: Goat anti-rabbit IgG conjugated with Alexa Fluor 568 and 488 (#A21069 and #A11034); Goat anti-rat IgG conjugated with Alexa Fluor 568 and 647 (#A11006 and #A21247); Goat anti-hamster conjugated with Alexa Flour 647 (#A21451) (Thermo Fisher Scientific Inc., Carlsbad, California, USA).

Quantification of IBA1 immunofluorescence in Foxn1$^{nu/nu}$ brains

Four brain tissue sections from control (n=4) and 28-day metastatic (n=4) Foxn1$^{nu/nu}$ mouse brains were stained for IBA1. Micrographs were acquired on the BZ-X700 Keyence fluorescence microscope. Baseline exposure level for IBA1 was established using control brain tissues under 20X magnification. For controls, 8 x 16 μM Z-stack fields of brain parenchyma per mouse were taken. For 231-BR metastatic brains, AcGFP^{+} lesions were located at low magnification (2X), then images at 20X using the same exposure setting as control. Z-stack micrographs were compressed into maximum intensity projection and opened in ImagJ (https://imagej.nih.gov/ij/). Regions of interest were quantified for IBA1 fluorescence intensity as the mean fluorescence intensity per pixel for control (n=115), peritumoral (n=127) and distal (n=96). Data was tabulated and analyzed in GraphPad Prism 8 (https://www.graphpad.com/scientific-software/prism/).

Isolation of cells for scRNA-seq

Single cell suspensions from mouse brains were prepared using the Adult Brain Dissociation Kit, Mouse and Rat (Miltenyi Biotec) with some modifications. Whole dissected brains were

chopped into 8 pieces of equal size and placed into C tube (Miltenyi Biotec) containing enzyme P and A. Brain tissue was digested using gentleMACS Octo Dissociator with heaters operating the Adult brain dissociation protocol (Miltenyi Biotec). After digestion, the cell suspension was strained over a sterile 70μm strainer (Fisher Scientific) and washed with 5mL FACS buffer containing ice cold DMEM/F12, 50mM HEPES, and 2% BSA. After removal of myelin by density centrifugation, the cell pellet was washed and remaining red blood cells were lysed with red blood cell lysis buffer. Cells were then re-suspended in FACS buffer and blocked with anti-CD16/32 for 15 minutes on ice. Next, cells were stained with fluorescent antibodies on ice for 15 minutes shielded from light. The labeled cells were washed with 500μL of FACS buffer and resuspended in 500μL of FACS buffer, strained through 40μm strainer prior to sorting on BD FACSAria Fusion sorter. For sorting of microglia, astrocytes, and cancer cells, cells were gated for size based on forward and side scatter, single cells, and Sytox Blue viability (Thermofisher, S34857). All myeloid cells ($CD45^+$ $CD11b^+$) and astrocytes ($CD45^-$, $ACSA2^+$) were sorted from control and metastatic mouse brains into 500μL of chilled FACS buffer. GFP^+ 231BR cells were sorted from metastatic brains into 500μL of FACS buffer.

scRNA-seq of $Foxn1^{nu/nu}$ cells

FACS isolated mouse microglia cells were centrifuged for 10 minutes at 300g and washed with 0.04% BSA in PBS. Cells were resuspended to achieve approximately 1,000 cells/μL. Final cell suspensions were counted on the Countess II automated cell counter to determine actual concentration for droplet generation. Cells were loaded onto the 10x Genomics Chromium Single Cell Gene Expression 3' v2 Chemistry kits for GEMs generation. Following

the Chromium Single Cell 3′ Reagents Kits version 2 user guide (CG00052 Rev B), cells were loaded to achieve approximately 10,000 cells for capture. Libraries were sequenced on the Illumina HiSeq 4000 platform to achieve an average of read depth of 50,000 mean reads per cell. Sequencing reads were aligned utilizing 10x Genomics Cell Ranger Count 3.0.2 to a dual indexed GRCh38 and mm10 reference genome.

Flow cytometry analysis of microglia from BCBM

For flow cytometry analysis of metastatic mouse brains, tissue was prepared as for FACS sorting, with the exception that 1mg/mL Collagenase D (Milipore Sigma #11213857001) was used for digestion instead of enzyme P provided in the Adult Brain Dissociation kit. After a single cell suspension was obtained, cells were stained with ZombieNIR viability dye (1:500, BioLegend) in 50µL of ice cold PBS for 15 minutes. Cells were washed with FACS buffer and blocked with anti-CD16/32 antibody diluted in FACS buffer for 15 minutes on ice. Next, cells were stained with fluorescent antibodies for 15 minutes on ice, protected from light. Cells were washed with 500µL FACS buffer and resuspended in 400µL FACS buffer, strained through a 40µm cell strainer and analyzed using BD Fortessa X20.

***In vitro* differentiation and early postnatal transplantation of iHPCs**

Differentiation of Hematopoietic Progenitor Cells from iPSCs (iHPCs) performed according to McQuade et al. (2018). Briefly, iPSCs were first passaged in mTeSR-E8 and on day 0, cells were transferred to Medium A from the STEMdiff Hematopoietic Kit (Stem Cell Technologies). On day three, flattened endothelial cell colonies were transferred to Medium B for seven additional days while iHPCs began to lift off the colonies. On day 10, non-adherent

CD43+ iHPCs were collected by removing medium and cells and at this point, d10-d11 iHPCs can be frozen in Bambanker (Fisher Scientific) for later transplantation. Cells used for early-postnatal iHPC transplantation were thawed in iPS-Microglia medium (DMEM/F12, 2X insulin-transferrin-selenite, 2X B27, 0.5X N2, 1X glutamax, 1X non-essential amino acids, 400 mM monothioglycerol, and 5 mg/mL human insulin freshly supplemented with 100ng/mL IL-34, 50ng/mL TGFb1, and 25 ng/mL M-CSF (Peprotech) according to [154]) and allowed to recover for 24 h. Early Postnatal Intracerebroventricular Transplantation of iHPCs was performed as described in [155]. Briefly P1 to P2 MITRG mice placed on ice for two-three min to induce hypothermic anesthesia. Free-hand transplantation was performed using a 30-gauge needle affixed to a 10μL Hamilton syringe, mice received 1μL of iHPCs suspended in sterile 1X DPBS at 31.25-62.5K cells/μL at each injection site (8 sites) totaling 250-500K cells/pup. Bilateral injections were performed at 2/5th of the distance from the lambda suture to each eye, injecting into the lateral ventricles at 3mm and into the overlying anterior cortex at 1mm, and into the posterior cortex in line with the forebrain injection sites, and perpendicular to lambda at a 45° angle. Transplanted pups were then returned to their home cages and weaned at P21.

Isolation of human xenotransplanted microglia

At 10 weeks old, MITRG mice were injected intracardially with 500,000 mCherry labeled 231BR cells as previously described. 25 days after intracardiac injection and following perfusion with ice cold PBS containing 5μg/ml actinomycin D (act D), whole metastatic brains were briefly imaged on a dissection microscope (Leica Biosystems, DMC 2900) for mCherry and GFP intensity. Half brains were then dissected, fixing the left hemisphere in 4%

PFA for histology and the right hemisphere was prepped for dissociation as described in [155] with modifications. The cerebellum was removed and the whole right hemisphere was stored briefly in RPMI 1640 containing 5μg/mL act D, 10μM triptolide, and 27.1ug/mL anisomycin. Tissue dissociation was then performed using the Tumor Dissociation kit, human (Miltenyi Biotec) and the gentleMACS OctoDissociator with heaters (Miltenyi Biotec) according to manufacturer guidelines with modifications. Briefly, tissue was cut into ~1mm pieces and placed into the C-tubes with the kit's enzymes, 5μg/mL act D, 10μM triptolide, and 27.1ug/mL anisomycin and samples were dissociated using the preprogrammed soft tumor protocol. Following enzymatic digestion, samples were strained through a 70μm filter and pelleted by centrifugation. Myelin and debris were removed by resuspending the pellet in 8mL 23% Percoll, overlaid with 2mL of 1X DPBS, spinning at 400xg for 25 minutes at 4°C, with acceleration and brake set to 0, and discarding the myelin band and supernatant.

MULTI-seq labeling and scRNA-seq of human microglia

For barcoding of cells from each individual mouse the MULTI-seq lipid- tagged indices for sample multiplexing for scRNA-seq protocol was followed [156]. Lipid anchor and co-anchor reagents were a generous gift from Zev Gartner, and barcode index oligos were purchased from Integrated DNA Technologies, Inc. Cells were resuspended and washed with 15 mL cold DPBS and pelleted by centrifugation (10 minutes, 400xg). The supernatant was discarded, and cells were resuspended in 180μL of DPBS. 20μL of 20μM Anchor:Barcode solution was added to a final concentration of 2μM, and incubated on ice for five minutes. Next 20μL of 20μM Co-Anchor solution was added, gently mixed and incubated for five minutes. After incubation 1mL 1% BSA in DPBS was added and cells were pelleted by centrifugation (five

minutes, 400xg). Finally, the supernatant was removed and washed a second time with 1mL 1% BSA in PBS and pelleted by centrifugation (5 minutes, 400xg). Next, mouse cell removal was performed by resuspending cell pellets in 160μL FACS buffer (0.5% BSA in 1X DPBS) + 40μL Mouse cell removal beads (Miltenyi Biotec) and incubated at 4°C for 15 minutes. Mouse and human cells were then separated using LS columns and the MidiMACs separator (Miltenyi Biotec) and the human cells were collected in the flow through. Human cells were pelleted via centrifugation (10 minutes, 400xg) and control samples and metastatic samples were then pooled separately. Cells were resuspended to ~1,000 cells per microliter in FACS buffer, according to counts performed on a hemocytometer.

ScRNA-seq of MITRG human microglia

Final cell suspensions were counted on the Countess II automated cell counter to determine actual concentration for droplet generation. Cells were loaded onto the 10x Genomics Chromium Single Cell Gene Expression 3' v3 Chemistry kits for GEMs generation. Following the Chromium Single Cell 3′ Reagents Kits version 3 user guide (CG000183 Rev C), cells were loaded to achieve approximately 10,000 cells for capture. MULTI-seq barcode libraries were prepared according to the MULTI-seq protocol [156]. Libraries were sequenced on the Illumina NovaSeq 6000 platform to achieve an average read depth of 50,000 mean reads per cell for 3' gene expression libraries. MULTI-seq barcode libraries were sequenced to achieve at least 5,000 reads per cell. Sequencing reads were aligned utilizing 10x Genomics Cell Ranger Count 3.1.0 to a dual indexed GRCh38 and mm10 reference genome. All libraries were aggregated using 10x Genomics Cell Ranger Aggr 3.1.0, to normalize the number of mean

reads per cells. MULTI-seq reads were processed according to the MULTI-seq protocol (https://github.com/chris-mcginnis-ucsf/MULTI-seq).

Microglia depletion study

Four-six week old *Csf1r*$^{\Delta FIRE/\Delta FIRE}$ (FIRE-KO) and *Csf1r*$^{FIRE/FIRE}$ (FIRE-WT) mice were injected intracranially in the right coronal suture with 100,000 enhanced GFP and luciferase labeled EO771 cells as previously described. To monitor brain tumor growth *in vivo*, mice were imaged for luciferase luminescence one day after injection, and every three days thereafter until endpoint. Imaged mice were anesthetized via isoflurane inhalant and administered 300µg D-Luciferin (Goldbio), intraperitoneally, in sterile DPBS. Following a 10-minute incubation, mice were imaged for bioluminescence for six minutes utilizing an IVIS Lumina III In Vivo Imaging System (Xenogen). Regions of interest were selected around each brain and average photon flux (total photons/s-cm^2) was recorded using Living Image analysis software (Perkin-Elmer) and average background flux subtracted. On day 14, mice were weighed, euthanized and dissected and the whole brains were removed and placed in a 24 well tissue culture plate submerged in ice cold PBS with D-Luciferin (1.5 mg/mL, Goldbio). After 10 minutes incubation, whole brains were removed from the solution and placed on a black plastic card and imaged for luminescence for 1 second. A region of interest was drawn around each brain and the total flux (ptotal photons/s-cm^2) was recorded for analysis.

Immune cell isolation and scRNA-seq of FIRE mice

FIRE-WT and FIRE-KO mice were injected intracranially with 100,000 EO771 cells and dissected after two weeks. For scRNA-seq, four FIRE-KO bearing visible brian tumors, two

FIRE-WT bearing visible brain tumors and two FIRE-WT without visible brain tumors were euthanized as previously described and brains were digested using our standard GentleMACS protocol for FACS isolation. After removal of myelin using debris removal solution (Miltenyi Biotec) cells from all 8 mice were pooled and stained for CD45 and viability. Single, live $CD45^+$ cells, including CD45 low microglia, were sorted into FACS buffer and subjected to 10X barcoding as previously described. Sequencing reads were then aligned utilizing 10x Genomics Cell Ranger Count 3.1.0 to a mm10 reference genome.

GSE139971 CITE-seq realignment

FASTQ files associated with GSE139971 HTO barcodes and mRNA samples were downloaded using 'fastq-dump --split-files --origfmt --gzip' and realigned using CITE-seq-Count 1.4.3 and Cell Ranger 3.0.2 respectively. HTO barcodes were assigned to cells using the procedure in Seurat v3 based on the umi count matrix output from CITE-seq-Count.

Souporcell genotyping

Genotyping was performed using Souporcell (Heaton et al, 2020) for GSE139971 and FIRE samples. For GSE139971, three genotype clusters were assigned and HTO barcodes were used to assign the genotypes to EO771, $Cx3cr1^{CreERT/+}$, or $Cx3cr1^{CreERT/+}$:$ROSA26i^{DTR/+}$. For FIRE samples, 2 genotype clusters were assigned and microglia presence, as determined by gene expression, was used to assign clusters to FIRE-WT (+microglia) and FIRE-KO (-microglia). Notably, using genotyping to label $Cx3cr1^{CreERT/+}$ and $Cx3cr1^{CreERT/+}$:$ROSA26i^{DTR/+}$ mice in this dataset had a high concordance with the expected antibody sample barcodes

(matching in 96.7% of cells), which supports our use of this method to label FIRE-WT and FIRE-KO cells.

Human/mouse cell assignment

Cells were aligned to a merged GRCh38/mm10 genome using Cell Ranger v3. Cells were then determined to be from mouse or human based on the frequency of reads aligning to the mouse genome with very low quality cells with <200 genes (nFeature_RNA) filtered before estimating. Cells were called as mouse for all cells above the top elbow in the mouse read mapping frequency plot (>0.875 for Foxn1$^{nu/nu}$ data; >0.95 for MITRG data), human for all cells below the bottom elbow (<0.05 for Foxn1$^{nu/nu}$ data; <0.1 for MITRG data), and any other cells were discarded as doublets or poor quality. Any counts for GRCh38 genes in the cells called as mouse were removed from the expression matrix and vice versa for mm10 genes in human cells.

Quality control metrics for scRNA-seq

Cells for the Foxn1$^{nu/nu}$ cell type identification analysis were filtered to have between 500 and 2000 genes (nFeature_RNA) and <10% mitochondrial genome reads (percent.mito) in any retained cell. Putative microglia/astrocyte doublet clusters with marker gene co-expression were removed from the Foxn1$^{nu/nu}$ microenvironment. This cell set was then used for subset myeloid and astrocyte analyses based on the cell type labels. Cells were further filtered for the myeloid analysis to have <5% percent.mito and low ribosomal expression (<10% of their transcriptome representing Rps and Rpl genes). An additional small cluster of putative microglia/astrocyte doublets was removed from the final astrocyte

analysis. Cells for the 231BR analysis in Foxn1$^{nu/nu}$ were filtered to have >2500 genes (nFeature_RNA), <60000 reads (nCount_RNA), and <10% percent.mito. Cells for the MITRG analysis were filtered to have <20% percent.mtio. Doublets and empty gems (Negative) were also removed from the MITRG analysis based on MULTI-Seq barcoding label assignment from the R package deMULTIplex. Cell cycle signatures (S.Score and G2M.Score, determined by CellCycleScoring in Seurat) were regressed from the data for the 231BR analysis as well as the MITRG analysis before clustering and dimensionality reduction. FIRE immune cells were first filtered to have >200 and <3500 genes (nFeature_RNA) and <7.5% percent.mito, low quality clusters were removed separately to conserve cell types with low gene expression (e.g. neutrophils), and doublets were removed based on Sourporcell labels [167]. GSE139971 samples were filtered to be singlets by both HTO-barcode and Souporcell assignment and only clusters that expressed CD45 (*Ptprc)* and were not assigned to the EO771 cluster by Souporcell genotype were kept for downstream analysis.

Clustering and differential expression

Main clustering and dimensionality reductions were performed in Seurat using the default Louvain and tSNE methods respectively. UMAP was used for dimensionality reductions in microglia subclustering analyses to better visualize global relationships. Some datasets were integrated using the mutual kNN algorithm adaptation in Seurat before these steps. Specifically, integration was performed on the Foxn1$^{nu/nu}$ full microenvironment and astrocyte analyses by sequencing batch (Con1:Met1, Con2:Con3, Met2:Met3) and the subclustering analyses for metastatic and control Foxn1$^{nu/nu}$ myeloid cells were also batch integrated. Integrated analyses used the “vst” selection method with nfeatures=2000 for

FindVariableFeatures and dims=1:30 for FindIntegrationAnchors and IntegrateData. Differential expression analyses were run on the RNA assay in Seurat with FindAllMarkers/FindMarkers using the Wilcoxon rank sum test and adjusted *P* values represent the Bonferroni corrected values for all single-cell analyses. For all samples except GSE139971, cell types and states were assigned to clusters manually based on gene expression profiles. GSE139971 cell type labels were determined by label transfer for FIRE immune cells using the standard pipeline in Seurat v3.

GO term analysis and gene scoring

GO term analyses were performed using the MouseMine [137] web portal with list input for M. musculus with the default background population for mouse analyses and using the Enrichr portal [102,103] with a gene list input. Gene inputs for each condition included only genes considered differentially expressed with a Bonferroni adjusted *P* value < 0.05 from the Wilcoxon rank sum test. Specific GO terms were then selected from the Gene Ontology Enrichment section for biological_process with Holm-Bonferroni adjusted *P* value < 0.05 in MouseMine or the GO Biological Process 2018 list in Enrichr with unadjusted *P* value < 0.05. All gene scoring on single-cell data was performed in Seurat using the AddModuleScore function with default parameters. MG-score gene list was taken directly as the Core MG list from Table S4 in [126]. M1 and M2 gene signatures were translated to mouse from Table S4 of [62] using the biomaRt package in R. Microglia subcluster profiles from Foxn1$^{nu/nu}$ mice were taken as top marker genes (logFC > 0.5) for each cluster compared to all other myeloid cells from mice with BCBM, and translated to human using the biomaRt package.

CHAPTER 5: Conclusions & future directions

Observing the natural world, identifying patterns in those observations, and generating new hypotheses based on said patterns is not only one of the earliest scientific pursuits, but a tenant of the scientific method[168]. In the few centuries since Aristotle described this method of 'inductive-deductive' reasoning [168], many tools have been created that facilitate the observation of previously unobservable natural phenomena. Single-cell RNA-seq (scRNA-seq), first achieved around 12 years ago[169], is a tool whose primary purpose is to unbiasedly assay the whole transcriptome of single-cells to give us insight into cell types, states, and their responses to stimuli. A key novelty of this technology is its ability to identify cell type heterogeneity from within a single tissue, though it has also improved our understanding of variation between tissues and biological conditions since it is not as sensitive as other methodologies to contamination from untargeted cell types.

The computational methods developed to quantify transcriptional differences between groups of cells can be broadly referred to as differential expression tests. When these methods work well, they can be used to identify novel biomarkers of a physiological phenomenon, which refers to either single or combined gene expression profiles that help discriminate the condition of interest from a homeostatic or separate physiological state. Biomarkers do not need to be mechanistic or functional, but they must be well correlated to a state or behavior of interest. Unfortunately, it is common to find that using a straightforward differential expression test in scRNA-seq results in thousands of statistically significant gene candidates, many of which are too noisily expressed to interpret or are better correlated to an uninteresting feature of the data than they are to the condition of

interest. In this work, we have developed multiple heuristics that improve the filtering, identification, and generalizations of biomarkers from scRNA-seq data and detail how the results from these methods contribute to our understanding of the pre-neoplastic and neoplastic breast epithelium. Further, we have shown how states of cells in the microenvironment around breast tumors can be used to predict tumor outcomes in the brain by fully characterizing the transcriptomic changes of microglia in response to breast cancer brain metastasis. Together, these studies highlight the multiple sources of heterogeneity in breast cancer and its associated metastases and give us insight into how this heterogeneity may be conserved across patients and model organisms.

The healthy breast epithelial hierarchy

In Chapter 2, we presented our work on a pilot scRNA-seq atlas of the healthy human breast epithelium from four reduction mammoplasty samples. From this dataset, we identified three conserved cell types (Basal, Luminal 1, and Luminal 2), and five conserved cell states (Basal, Myoepithelial, Luminal 1.1, Luminal 1.2, and Luminal 2). We then investigated potential lineage connections between the cell types and states using the pseudotime algorithm Monocle2, which required the development of a heuristic to identify conserved gene expression changes across these cell states in all four patients. This heuristic procedure started by running Monocle2 on each patient individually using markers of conserved cell states, then identified correlated gene modules across each patient's trajectory, and finally averaged these correlations to remove inconsistencies (e.g. remove genes that are positively correlated in one patient, but not correlated or negatively correlated in other patients) and generated a gene list for a combined ordering. By requiring

each ordering gene to be part of a correlated module, we increase the chances of these genes being meaningful pathway shifts and reduce the likelihood of noise driving the trajectory transitions. Additionally, by requiring consistency in gene module behaviors across patients, we reduce batch effects and ensure that our trajectory is driven by cell state gene expression changes rather than patient-specific gene expression differences. This procedure can be used on any dataset as an alternative to the Monocle2 proposed gene list selection method dpFeature[41].

From the resultant lineage trajectory, we were able to generate and support a few key hypotheses, as well as provide the gene list of smoothly transitioning breast epithelial state markers to the community for additional investigation. One observation is that basal and luminal lineages are connected by mesenchymal-like cell states, which is consistent with proposed markers of bipotent mammary stem cells[4,170,171]. A more novel observation is a connection between the hormone responsive (Luminal 2) and secretory (Luminal 1) luminal cells, with a possible terminal lineage marked by *KIT* and *ELF5*, previously characterized as "luminal progenitor" markers.[84] While it is interesting to propose these fundamental changes to our understanding of the luminal lineage using our scRNA-seq data, it is important to remember that pseudotemporal analysis with Monocle2 is not lineage tracing, and the trajectory is indicative of transcriptional gradients, but does not have any enforced directionality. Therefore, one can reasonably hypothesize bidirectionality or multiple progenitor cell types transitioning into the same terminal state, or even to reinterpret the trajectory as the effect of signaling gradients rather than lineage relationships if their biological understanding permits it. A way to refine this analysis for unbiased hypothesis generation in future studies would be to add RNA velocity, which overlays directionality to

a dimensionality reduction based on the ratio of spliced and unspliced RNA transcripts for each gene across the transcriptome.[43,44] By doing this, we can estimate where a cell is "going" or where it has arrived from since nascent, or unspliced RNA capture may indicate an increase in the transcription of a given gene, and vice versa for repression. Since the method is orthogonal to our Monocle2 analysis, meaning that it works off an entirely separate element of the data (in this case, the spliced or unspliced reads rather than the simple mapped gene expression profiles), it would facilitate new interpretations of the data without the need for new functional data.

We can also retrospectively alter our hypotheses in light of new functional studies, while preserving our scientific contribution. Recent studies in the mouse mammary gland have suggested that the hormone responsive (Luminal 2) and secretory (Luminal 1) lineages have entirely separate progenitor populations[172,173]. If we assume the mouse mammary gland and human breast have homologous progenitor populations, this means that our continuous luminal trajectory must represent a non-lineage driven gradient. Our *in situ* analysis of Luminal 1 and Luminal 2 cells suggest that they are not spatially restricted (i.e. not simply ductal and lobular), so it would need to be more complex than a pseudotime to pseudospace reinterpretation. A recent pre-print investigating the roles of BMI and parity in the human breast epithelium also using scRNA-seq may give hints as to other explanations for our observed plasticity. In Murrow et al, 2020, they find that BMI is associated with a decrease in the number of hormone responsive, Luminal 2 cells, and that parity decreases the level of hormone responses in a given Luminal 2 cell[51]. Our dataset controlled for parity, so this decreased hormone response gradient is not a likely explanation for our connected luminal lineages, but it is possible that BMI or menstrual cycle stage plays a similar role in

stimuli response and that we have captured this in our trajectory. In the mouse mammary gland, studies have also shown that cell cycle is altered during estrus and diestrus (menstrual cycle associated hormone fluctuations), and that this cycling phenotype is closely connected with the expression of lineage markers in hormone responsive luminal cells.[47,174] Thus, some of our captured differences between patients and our smooth luminal state transitions could be a consequence of natural menstrual cycle variation, and future studies with more patients and better annotated menstrual information could readily clarify this and identify which genes in our list are most related to these estrous differences. In summary, our trajectory and gene list provide a solid foundational understanding of breast epithelial cell state relationships in homeostasis, and as new functional literature appears, we can better determine what transcriptional gradient is driving each possible branch and transition point to gain a wholistic view of the highly plastic breast epithelium and hopefully, translate this information into a better understanding of the changes that occur in the cancer setting.

Biomarkers of micrometastasis in triple-negative breast cancer

In Chapter 3, we presented our work on scRNA-seq of matched tumor and micrometastatic cells from triple-negative breast cancer patient-derived xenografts. We used this dataset to investigate inter and intra-tumoral heterogeneity and identify differences between primary tumor and micrometastatic transcriptomes. A standard tobit test, controlled for patient differences, revealed 330 genes significantly differentially expressed between tumor and micrometastatic cells conserved across our three patient models, though we discovered that some of these genes were still heavily skewed towards patient subsets. Therefore, we added a second feature selection step to prioritize robust

biomarkers of micrometastasis by using a forward selection, stepwise logistic regression model that identified *LDHA*, *BHLHE40*, and *PHLDA2* as the most consistently predictive gene markers of tumor or micrometastatic status across individual cells. A model using only these three gene to predict our data labels was accurate ~75% of the time. This model also trended towards the misidentification of tumor cells as micrometastatic cells, which may suggest rare cells in the primary tumor are transcriptionally primed to metastasize. This observation held true when we used RNAscope to identify *PHLDA2* mRNA *in situ*, which found that almost all micrometastatic cells highly expressed *PHLDA2* while only rare tumor cells showed comparable levels of expression. We also validated that *PHLDA2*, and *BHLHE40* had prognostic utility for relapse-free survival (RFS) in a large cohort of publicly available data from breast cancer patients. From this, we discovered that stratifying patients for high *PHLDA2* and low *BHLHE40* in their primary tumor had a hazard ratio of 1.55 for RFS, suggesting that the combination of these two genes may be useful biomarkers of tumors with a high potential for metastatic invasion and seeding.

The major novelty of our approach was the addition of the forward selection, logistic regression model on top of the tobit differential expression test. We found that this forced us to prioritize genes with low dropout and helped us build a simple predictive model that performed quite well in label prediction. Forward selection also enforces uncorrelated gene sets, which indicates that *LDHA*, *PHLDA2*, and *BHLHE40* represent separate pathway markers. *PHLDA2* is less studied than *BHLHE40*, but they are suggested to have opposing effects on tumor invasiveness[98,99], while *LDHA* is primarily known for its role in glycolysis[175]. Based on *PHLDA2*'s other suggested role in improving xenograft engraftment, it may be part of an uncharacterized stress-response pathway, which would be consistent with other

observed stress-related gene expression profiles in micrometastatic cells (e.g. *HSPA8, SOD1*). Further, micrometastatic cells are far more epithelial-like transcriptionally than their primary tumor counterparts (i.e. they express more epithelial markers like *EPCAM, KRT14, KRT16*) and overall, EMT appears upregulated in primary tumors compared to micrometastatic cells, so the invasive role of *PHLDA2* may not be the major reason for its upregulation in distal micrometastases. Thus, one biological interpretation of our logistic regression model is that successful micrometastatic cells must downregulate glycolysis and invasiveness repressors and upregulate their stress responses to remain alive after intravasation and extravasation into the lung and lymph nodes. Importantly, in tandem with their downregulated glycolysis, our work showed that micrometastatic cells must upregulate OXPHOS and inhibiting their ability to do so decreases metastatic burden, though this transcriptional pattern was not specifically captured in our logistic regression model and may represent an important pitfall of our method. Namely, using forward selection, logistic regression on scRNA-seq data is slightly too conservative in terms of gene selection, so key pathways can be missed and this possibility should be separately investigated using the first-pass differential expression data.

While it is vital to follow up on the mechanistic roles of *PHLDA2* in micrometastasis with perturbation experiments, it is notable that this gene was not the only marker of micrometastasis we identified with prognostic utility from our initial tobit test. In fact, 15 of the top 20 genes shown to be upregulated in micrometastatic cells were found to significantly predict poor RFS in Basal like breast cancers based on KM Plotter data (result published, but not directly shown in chapter)[56]. A recent pre-print from Ma & Hernandez et al, developed a novel culture-transplant system using two of the patient-derived xenograft

models from our study (HCI002, HCI010), which allows for the perturbation of specific genes in these tumors to investigate their functional consequences[176]. In this pre-print, they also investigated the role of one of these top markers of micrometastasis, *NME1*, and found that its overexpression resulted in increased lung metastasis after orthotopic transplant, but no changes in primary tumor size in patient HCI010, indicating a function for this gene in either invasiveness or metastatic survival [176]. This culture system can and should be used to further investigate the function of our other micrometastatic markers (e.g. *PHLDA2*) as well as our tumor markers (e.g. *BHLHE40*), since it can help us deconvolve their pathway associations and roles in the metastatic cascade. For pathway analysis, we can perform either overexpression or knockdown of individual genes in our tumors cells, and utilize bulk RNA-sequencing to see how the 330 genes identified in our screen change as a result. To determine the role of a given gene in the cascade, we can also take advantage of circulating tumor cell (CTC) numbers by flow cytometry to determine whether intravasation improved, distant metastasis numbers by flow cytometry to determine whether extravasation improved, and use proliferation markers (e.g. *in situ* stains for Ki67) to determine whether outgrowth capabilities improved. By using mouse models with or without NK cells (e.g. NOD SCID vs NSG mice), we can also better understand which of our identified genes facilitated immune evasion in our described chapter, which utilized NOD SCID mice who retained their NK cells and had a far lower metastatic burden than the NSG mice used in the aforementioned pre-print. Overall, the core 330 differentially expressed genes provide a rich starting point for investigations into the metastatic cascade, and our logistic regression methodology offers another way of probing into the central features of micrometastasis. Our

unique methodology can also be extended for use on other scRNA-seq datasets, where we believe it will perform a similarly useful function.

Microglia responses to breast cancer brain metastasis

In Chapter 4, we discussed our work on the response of microglia to breast cancer brain metastasis (BCBM) using scRNA-seq on a series of mouse models to uncover the consequences of these responses. Using the Foxn1$^{nu/nu}$ immunocompromised mouse model with human 231BR triple-negative breast cancer cells, we identified a strong global M1-like, pro-inflammatory microglial response. We validated three key markers of microglia at the protein level, one for type I interferon response (BST2), and two for antigen presentation (CD74, MHC-II) in the Foxn1$^{nu/nu}$ model, as well as two immunocompetent mouse models of BCBM, 4T1-BALB/c and EO771-C57BL/6. Generally, this analysis validated that CD74 is upregulated specifically in tumor-proximal microglia, and that the major antigen presentation and type I interferon responses are conserved, but heterogeneously expressed, in all mice with BCBM. Delving deeper into our scRNA-seq data, we identified the microglia most likely to be tumor-proximal using an iterative scoring and subclustering method, and found that these microglia had five major states, namely proliferative (Cycling), cytokine and exosome secreting (Secretory), glycolytic (Glycolytic), interferon responsive (IFN responsive) , and antigen presenting (APC). To ensure these behaviors were not exclusive to mouse microglia, we next performed scRNA-seq on an immunocompromised, humanized mouse model of BCBM (231BR-MITRG) which allowed us to investigate responses of human microglia to human breast cancer cells. The human microglia from this model had fully analogous responses to BCBM as our Foxn1$^{nu/nu}$ model, and gene scoring demonstrated that

our mouse-derived gene signatures can be directly applied to human microglia to determine their state. Finally, we utilized an immunocompetent mouse model of BCBM with a stable genetic-depletion of microglia (EO771-FIRE-WT/FIRE-KO) to demonstrate that microglial depletion results in worse tumor outcomes (decreased tumor regression and increased morbidity), suggesting that the pro-inflammatory behaviors of microglia in BCBM have the anti-tumor role we had hypothesized.

A major question that still needs to be addressed in this study is the specific roles for each microglia state, as well as their temporal relationships during BCBM progression. Tumor associated macrophages (TAMs) are thought to change in character from tumor-suppressive to tumor-promoting throughout the course of tumor progression[64,177]. In our BCBM models, tumor progression is relatively fast (~three-four weeks), but since the disease progression is consistent, we could collect scRNA-seq data from microglia at a two-week timepoint to represent "early" responses. The tumor-proximal microglia would be quite rare at two weeks since tumors are much smaller on average, so microdissection may be necessary before sorting to enrich for microglia of interest. Additionally, we can investigate the *in situ* expression of our microglia state marker genes using RNAscope to both validate that our populations are tumor-proximal and to give a rough timeline for when they become enriched in the brain. A less robust way to investigate the temporal relationships of the microglia states would be to use pseudotime analysis, overlayed with RNA velocity to see the movement from one state to another. This would not necessarily be appropriate for addressing tumor education since all of the microglia are present in highly tumor burdened brains, but it could elucidate which states arise from other states within the tumor microenvironment (e.g. Secretory microglia may phagocytose breast cancer cells, and

transition into APC microglia). Using either new data or new analyses, we could also begin to address the functional roles of each microglia state. We have already named microglia based on what we expect their phenotypes and functions to be, and we have proposed some of the environmental factors that may drive these phenotypes (e.g. type I interferon or interferon gamma), but we have not validated that these transcriptome profiles correspond to the proposed functions. This follow up could be quite extensive, but a few simple options could be to co-culture a microglia cell line with 231BR cells to identify states that do not require the larger immune context (i.e. direct tumor influences), or to test antigen presentation *in vitro* using a T cell activation assay with microglia purified from tumor or non-tumor bearing brains.

On the topic of immune contexts, it would be interesting to investigate how the microglia states we identified shift in tumor rejection or tumor promoting microenvironments. As mentioned in our chapter, BALB/c mice are thought to have Th2, or tumor promoting immunity, while C57BL/6 mice are thought to have Th1, or tumor suppressive immunity[178]. Our protein expression data suggests that BALB/c mounts a weaker immune response to 4T1s than C57BL/6 mounts to EO771 since BALB/c mice have lower frequencies of microglia with MHC-II and BST2, and show no significant enrichment of CD74. This suggests that BALB/c mice may lack APC microglia (since CD74 is a primary marker of this population) which could result in an unproductive tumor-rejection response. It would be highly informative to collect the full immune repertoires of both the 4T1-BALB/c and EO771-C57BL/6 with scRNA-seq by sorting tumor bearing and naïve brains for CD45, which appeared effective in the EO771-FIRE model. This data may help us determine how immune infiltration influences the microglia states we see in our other BCBM models and

could also clarify which microglia states are enriched in productive and unproductive immune responses. This information will be important in patient contexts as well, since it is likely that human genetics plays a similar role in their tumor responsiveness, and patient stratification is key for effective immunotherapy.

Final remarks

Each of the projects investigated in this work were on a unique biological system (healthy breast epithelium, triple-negative breast tumor micrometastasis, and microglia in BCBM) but a few observations remained context independent. ScRNA-seq has allowed us to investigate cell state heterogeneity at a scale previously intractable, and what has appeared is actually a high-degree of conservation. Specifically, while cells have many substates during differentiation, as well as within and around tumors, the possible states are the mostly same across individuals, and only the proportion of cells in each state differ. This is not too dissimilar to our understandings of cell types and suggests that consortium projects that seek to define these cell states across many individuals, like the Human Cell Atlas, will be successful in creating a generalizable database from only a few hundred patient samples. We also demonstrated that a lot can be gained from scRNA-seq data by adding question-driven heuristics into a standard analysis pipeline. In the breast epithelium, this meant generalizing lineage trajectories across patients; in our PDX models of micrometastasis, this meant using a predictive model to find conserved and uncorrelated gene markers of micrometastatic cells; and in microglia, this meant using pseudo-bulk data to predict tumor-proximity to assess heterogeneity during BCBM. These methodologies led to the creation of novel, systems-level hypotheses, which are consistent with our current understanding of biology

but remain able to be meaningfully reinterpreted as new understandings arise. Even with careful attention to the literature, it is highly unlikely that all of the hypotheses presented here will hold true but our hope is that the observations and methodologies we have presented here will have enough clarity and accuracy to serve as a foundation on which even better hypotheses can be built in the future.

REFERENCES

1. Visvader, J. E. & Stingl, J. Mammary stem cells and the differentiation hierarchy: Current status and perspectives. *Genes Dev.* **28**, 1143–1158 (2014).

2. Paine, I. S. & Lewis, M. T. The Terminal End Bud: the Little Engine that Could. *J. Mammary Gland Biol. Neoplasia* **22**, 93–108 (2017).

3. Dontu, G. & Ince, T. A. Of Mice and Women: A Comparative Tissue Biology Perspective of Breast Stem Cells and Differentiation. *J. Mammary Gland Biol. Neoplasia* **20**, 51–62 (2015).

4. Fu, N. Y., Nolan, E., Lindeman, G. J. & Visvader, J. E. Stem cells and the differentiation hierarchy in mammary gland development. *Physiol. Rev.* **100**, 489–523 (2020).

5. Dai, X. *et al.* Breast cancer intrinsic subtype classification, clinical use and future trends. *Am. J. Cancer Res.* **5**, 2929–2943 (2015).

6. Prat, A. & Perou, C. M. Deconstructing the molecular portraits of breast cancer. *Mol. Oncol.* **5**, 5–23 (2011).

7. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10869–10874 (2001).

8. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **533**, 747–752 (2000).

9. Rodilla, V. & Fre, S. Cellular plasticity of mammary epithelial cells underlies heterogeneity of breast cancer. *Biomedicines* **6**, 9–12 (2018).

10. Anstine, L. J. & Keri, R. A new view of the mammary epithelial hierarchy and its implications for breast cancer initiation and metastasis. *J. Cancer Metastasis Treat.* **2019**, (2019).

11. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* **12**, 68–86 (2010).

12. Skibinski, A. & Kuperwasser, C. The origin of breast tumor heterogeneity. *Oncogene* **34**, 5309–5316 (2015).

13. Sauter, G., Lee, J., Bartlett, J. M. S., Slamon, D. J. & Press, M. F. Guidelines for Human Epidermal Growth Factor Receptor 2 Testing: Biologic and Methodologic Considerations. *J. Clin. Oncol.* **27**, 1323–1333 (2009).

14. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

15. Hu, Z. *et al.* The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* **7**, 1–12 (2006).

16. Soysal, S. D., Tzankov, A. & Muenst, S. E. Role of the Tumor Microenvironment in Breast Cancer. *Pathobiology* **82**, 142–152 (2015).

17. Paget, S. THE DISTRIBUTION OF SECONDARY GROWTHS IN CANCER OF THE BREAST. *Lancet* **133**, 571–573 (1889).

18. Wei, S. & Siegal, G. P. Metastatic organotropism: An intrinsic property of breast cancer molecular subtypes. *Adv. Anat. Pathol.* **24**, 78–81 (2017).

19. Cejalvo, J. M. *et al.* Intrinsic subtypes and gene expression profiles in primary and metastatic breast cancer. *Cancer Res.* **77**, 2213–2221 (2017).

20. Chen, X. *et al.* TNBCtype: A subtyping tool for triple-negative breast cancer. *Cancer Inform.* **11**, 147–156 (2012).

21. Streets, A. M. *et al.* Microfluidic single-cell whole-transcriptome sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 7048–7053 (2014).

22. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

23. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, (2017).

24. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, (2019).

25. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).

26. Thi, H. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* 1–32 (2020).

27. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**, 281-291.e9 (2019).

28. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* **8**, 329-

337.e4 (2019).

29. DePasquale, E. A. K. *et al.* DoubletDecon: Deconvoluting Doublets from Single-Cell RNA-Sequencing Data. *Cell Rep.* **29**, 1718-1727.e8 (2019).

30. O'Flanagan, C. H. *et al.* Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *Genome Biol.* **20**, 1–13 (2019).

31. Van Den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).

32. Richardson, G. M., Lannigan, J. & Macara, I. G. Does FACS perturb gene expression? *Cytom. Part A* **87**, 166–175 (2015).

33. Sun, S., Zhu, J., Ma, Y. & Zhou, X. Accuracy, Robustness and Scalability of Dimensionality Reduction Methods for Single Cell RNAseq Analysis. *bioRxiv* 1–21 (2019) doi:10.1101/641142.

34. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* (2008).

35. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).

36. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, 1–12 (2008).

37. Carbon, S. *et al.* The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

38. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (80-.).* **352**, 189–196 (2016).

39. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

40. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).

41. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

42. Herring, C. A., Chen, B., McKinley, E. T. & Lau, K. S. Single-Cell Computational Strategies for Lineage Reconstruction in Tissue Systems. *Cmgh* **5**, 539–548 (2018).

43. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

44. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).

45. Giraddi, R. R. *et al.* Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development. *Cell Rep.* **24**, 1653-1666.e7 (2018).

46. Bach, K. *et al.* Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat. Commun.* **8**, (2017).

47. Pal, B. *et al.* Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. *Nat. Commun.* **8**, 1–13 (2017).

48. Sun, H. *et al.* Single-cell RNA-Seq reveals cell heterogeneity and hierarchy within mouse mammary epithelia. *J. Biol. Chem.* **293**, 8315–8329 (2018).

49. Wuidart, A. *et al.* Early lineage segregation of multipotent embryonic mammary gland progenitors. *Nat. Cell Biol.* **20**, 666–676 (2018).

50. Nguyen, Q. H. *et al.* Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.* **9**, 1–12 (2018).

51. Murrow, L. M. *et al.* Pregnancy and obesity modify the epithelial composition and hormone signaling state of the human breast. *bioRxiv* (2020) doi:https://doi.org/10.1101/430611.

52. Fortner, R. T. *et al.* Parity, breastfeeding, and breast cancer risk by hormone receptor status and molecular phenotype: Results from the Nurses' Health Studies. *Breast Cancer Res.* **21**, 1–9 (2019).

53. Gao, R. *et al.* Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. *Nat. Commun.* **8**, 228 (2017).

54. Chung, W. *et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* **8**, 1–12 (2017).

55. Karaayvaz, M. *et al.* Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.* **9**, (2018).

56. Davis, R. T. *et al.* Transcriptional diversity and bioenergetic shift in human breast cancer metastasis revealed by single-cell RNA sequencing. *Nat. Cell Biol.* **22**, 310–320 (2020).

57. Kim, C. *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* **173**, 879-893.e13 (2018).

58. Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).

59. Lawson, D. A. *et al.* Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* **526**, 131–135 (2015).

60. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).

61. Martinez, F. O. & Gordon, S. The M1 and M2 paradigm of macrophage activation: Time for reassessment. *F1000Prime Rep.* **6**, 1–13 (2014).

62. Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174**, 1293-1308.e36 (2018).

63. Qian, J. *et al.* A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Res.* (2020) doi:10.1038/s41422-020-0355-0.

64. Williams, C. B., Yeh, E. S. & Soloff, A. C. Tumor-associated macrophages: Unwitting accomplices in breast cancer malignancy. *npj Breast Cancer* **2**, (2016).

65. Veglia, F., Perego, M. & Gabrilovich, D. Myeloid-derived suppressor cells coming of age

review-article. *Nat. Immunol.* **19**, 108–119 (2018).

66. Alshetaiwi, H. *et al.* Defining the emergence of myeloid-derived suppressor cells in breast cancer using single-cell transcriptomics. *Sci. Immunol.* **5**, (2020).

67. Meng, S. *et al.* Distribution and prognostic value of tumor-infiltrating T cells in breast cancer. *Mol. Med. Rep.* **18**, 4247–4258 (2018).

68. Vikas, P., Borcherding, N. & Zhang, W. The clinical promise of immunotherapy in triple-negative breast cancer. *Cancer Manag. Res.* **10**, 6823–6833 (2018).

69. Teng, M. W. L., Ngiow, S. F., Ribas, A. & Smyth, M. J. Classifying cancers basedon T-cell infiltration and PD-L1. *Cancer Res.* **75**, 2139–2145 (2015).

70. Savas, P. *et al.* Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat. Med.* **24**, 986–993 (2018).

71. Waldman, A. D., Fritz, J. M. & Lenardo, M. J. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat. Rev. Immunol.* (2020) doi:10.1038/s41577-020-0306-5.

72. Bartoschek, M. *et al.* Spatially and functionally distinct subclasses of breast cancer-associated fibroblasts revealed by single cell RNA sequencing. *Nat. Commun.* **9**, (2018).

73. Roswall, P. *et al.* Microenvironmental control of breast cancer subtype elicited through paracrine platelet-derived growth factor-CC signaling. *Nat. Med.* **24**, 463–473 (2018).

74. Brechbuhl, H. M. *et al.* Fibroblast subtypes regulate responsiveness of luminal breast cancer to estrogen. *Clin. Cancer Res.* **23**, 1710–1721 (2017).

75. Pelon, F. *et al.* Cancer-associated fibroblast heterogeneity in axillary lymph nodes drives metastases in breast cancer through complementary mechanisms. *Nat. Commun.* **11**, (2020).

76. Costa, A. *et al.* Fibroblast Heterogeneity and Immunosuppressive Environment in Human Breast Cancer. *Cancer Cell* **33**, 463-479.e10 (2018).

77. Kieffer, Y. *et al.* Single-Cell Analysis Reveals Fibroblast Clusters Linked to Immunotherapy Resistance in Cancer. *Cancer Discov.* **10**, 1330–1351 (2020).

78. Shackleton, M. *et al.* Generation of a functional mammary gland from a single stem cell. *Nature* **439**, 84–88 (2006).

79. Stingl, J. *et al.* Purification and unique properties of mammary epithelial stem cells. *Nature* **439**, 993–997 (2006).

80. Shehata, M. *et al.* Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. *Breast Cancer Res.* **14**, 1–19 (2012).

81. Stingl, J., Eaves, C. J., Zandieh, I. & Emerman, J. T. Characterization of bipotent mammary epithelial progenitor cells in normal adult human breast tissue. *Breast Cancer Res. Treat.* **67**, 93–109 (2001).

82. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

83. Gudjonsson, T., Adriance, M. C., Sternlicht, M. D., Petersen, O. W. & Bissell, M. J. Myoepithelial cells: their origin and function in breast morphogenesis and neoplasia. *J. Mammary Gland Biol. Neoplasia* **10**, 261–272 (2005).

84. Regan, J. L. *et al.* C-Kit is required for growth and survival of the cells of origin of Brca1-mutation-associated breast cancer. *Oncogene* **31**, 869–883 (2012).

85. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotech* **33**, 495–502 (2015).

86. Bianchini, G., Balko, J. M., Mayer, I. A., Sanders, M. E. & Gianni, L. Triple-negative breast cancer: Challenges and opportunities of a heterogeneous disease. *Nat. Rev. Clin. Oncol.* **13**, 674–690 (2016).

87. Weigelt, B., Peterse, J. L. & Van't Veer, L. J. Breast cancer metastasis: Markers and models. *Nat. Rev. Cancer* **5**, 591–602 (2005).

88. Oskarsson, T., Batlle, E. & Massagué, J. Metastatic stem cells: Sources, niches, and vital pathways. *Cell Stem Cell* **14**, 306–321 (2014).

89. Hermann, P. C. *et al.* Distinct Populations of Cancer Stem Cells Determine Tumor Growth and Metastatic Activity in Human Pancreatic Cancer. *Cell Stem Cell* **1**, 313–323 (2007).

90. Hunter, K. W., Crawford, N. P. S. & Alsarraj, J. Mechanisms of metastasis. *Breast Cancer Res.* **10**, 1–10 (2008).

91. DeRose, Y. S. *et al.* Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nat Med* **17**, 1514–1520 (2011).

92. Bruna, A. *et al.* Resource A Biobank of Breast Cancer Explants with Preserved Intra-tumor Heterogeneity to Screen Anticancer Resource A Biobank of Breast Cancer

Explants with Preserved Intra-tumor Heterogeneity to Screen Anticancer Compounds. *Cell* 1–15 (2016) doi:10.1016/j.cell.2016.08.041.

93. Lambert, A. W., Pattabiraman, D. R. & Weinberg, R. A. Emerging Biological Principles of Metastasis. *Cell* **168**, 670–691 (2017).

94. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–6 (2014).

95. Ikwegbue, P. C., Masamba, P., Oyinloye, B. E. & Kappo, A. P. Roles of heat shock proteins in apoptosis, oxidative stress, human inflammatory diseases, and cancer. *Pharmaceuticals* **11**, 1–18 (2018).

96. Karantza, V. Keratins in health and cancer: More than mere epithelial cell markers. *Oncogene* **30**, 127–138 (2011).

97. Győrffy, B. *et al.* An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res. Treat.* **123**, 725–731 (2010).

98. Asanoma, K. *et al.* Regulation of the Mechanism of TWIST1 Transcription by BHLHE40 and BHLHE41 in Cancer Cells . *Mol. Cell. Biol.* **35**, 4096–4109 (2015).

99. Moon, H. G. *et al.* Prognostic and functional importance of the engraftment-associated genes in the patient-derived xenograft models of triple-negative breast cancers. *Breast Cancer Res. Treat.* **154**, 13–22 (2015).

100. Liu, D. *et al.* Prognostic significance of serum lactate dehydrogenase in patients with breast cancer: A meta-analysis. *Cancer Manag. Res.* **11**, 3611–3619 (2019).

101. Lawson, D. A. *et al.* Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* **526**, 131–135 (2015).

102. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

103. Kuleshov, M. V *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90-7 (2016).

104. Witzel, I., Oliveira-Ferrer, L., Pantel, K., Müller, V. & Wikman, H. Breast cancer brain metastases: Biology and new clinical perspectives. *Breast Cancer Res.* **18**, 1–9 (2016).

105. Ostrom, Q. T., Wright, C. H. & Barnholtz-Sloan, J. S. *Brain metastases: epidemiology*. *Handbook of Clinical Neurology* vol. 149 (Elsevier B.V., 2018).

106. Niikura, N. *et al.* Treatment outcomes and prognostic factors for patients with brain metastases from breast cancer of each subtype: a multicenter retrospective analysis. *Breast Cancer Res. Treat.* **147**, 103–112 (2014).

107. Brufsky, A. M. *et al.* Central nervous system metastases in patients with HER2-positive metastatic breast cancer: Incidence, treatment, and survival in patients from registHER. *Clin. Cancer Res.* **17**, 4834–4843 (2011).

108. Rostami, R., Mittal, S., Rostami, P., Tavassoli, F. & Jabbari, B. Brain metastasis in breast cancer: a comprehensive literature review. *Journal of Neuro-Oncology* (2016) doi:10.1007/s11060-016-2075-3.

109. Martin, A. M. *et al.* Immunotherapy and Symptomatic Radiation Necrosis in Patients With Brain Metastases Treated With Stereotactic Radiation. *JAMA Oncol.* **4**, 1123–1124

(2018).

110. Deeken, J. F. & Löscher, W. The blood-brain barrier and cancer: Transporters, treatment, and trojan horses. *Clinical Cancer Research* vol. 13 1663–1674 (2007).

111. Tosoni, A., Ermani, M. & Brandes, A. A. The pathogenesis and treatment of brain metastases: a comprehensive review. *Crit. Rev. Oncol. Hematol.* **52**, 199–215 (2004).

112. Hanisch, U. K. & Kettenmann, H. Microglia: Active sensor and versatile effector cells in the normal and pathologic brain. *Nature Neuroscience* (2007) doi:10.1038/nn1997.

113. Wolf, S. A., Boddeke, H. W. G. M. & Kettenmann, H. Microglia in Physiology and Disease. *Annu. Rev. Physiol.* **79**, 619–643 (2017).

114. Hammond, T. R., Robinton, D. & Stevens, B. Microglia and the Brain: Complementary Partners in Development and Disease. *Annu. Rev. Cell Dev. Biol.* **34**, 523–544 (2018).

115. Quail, D. F. & Joyce, J. A. The Microenvironmental Landscape of Brain Tumors. *Cancer Cell* **31**, 326–341 (2017).

116. Goldmann, T. *et al.* Origin, fate and dynamics of macrophages at central nervous system interfaces. *Nat. Immunol.* (2016) doi:10.1038/ni.3423.

117. Mrdjen, D. *et al.* High-Dimensional Single-Cell Mapping of Central Nervous System Immune Cells Reveals Distinct Myeloid Subsets in Health, Aging, and Disease. *Immunity* **48**, 380-395.e6 (2018).

118. Jordão, M. J. C. *et al.* Single-cell profiling identifies myeloid cell subsets with distinct fates during neuroinflammation. *Science (80-.).* (2019) doi:10.1126/science.aat7554.

119. Duchnowska, R. *et al.* Immune response in breast cancer brain metastases and their microenvironment: The role of the PD-1/PD-L axis. *Breast Cancer Res.* **18**, (2016).

120. Coniglio, S. J. *et al.* Microglial Stimulation of Glioblastoma Invasion Involves Epidermal Growth Factor Receptor (EGFR) and Colony Stimulating Factor 1 Receptor (CSF-1R) Signaling. *Mol. Med.* (2012) doi:10.2119/molmed.2011.00217.

121. Pyonteck, S. M. *et al.* CSF-1R inhibition alters macrophage polarization and blocks glioma progression. *Nat. Med.* **19**, 1264–1272 (2013).

122. Quail, D. F. *et al.* The tumor microenvironment underlies acquired resistance to CSF-1R inhibition in gliomas. *Science (80-.).* (2016) doi:10.1126/science.aad3018.

123. Yan, D. *et al.* Inhibition of colony stimulating factor-1 receptor abrogates microenvironment-mediated therapeutic resistance in gliomas. *Oncogene* (2017) doi:10.1038/onc.2017.261.

124. Qiao, S., Qian, Y., Xu, G., Luo, Q. & Zhang, Z. Long-term characterization of activated microglia/macrophages facilitating the development of experimental brain metastasis through intravital microscopic imaging. *J. Neuroinflammation* (2019) doi:10.1186/s12974-018-1389-9.

125. Guldner, I. H. *et al.* CNS-Native Myeloid Cells Drive Immune Suppression in the Brain Metastatic Niche through Cxcl10. *Cell* 1–15 (2020) doi:10.1016/j.cell.2020.09.064.

126. Bowman, R. L. *et al.* Macrophage Ontogeny Underlies Differences in Tumor-Specific Education in Brain Malignancies. *Cell Rep.* **17**, 2445–2459 (2016).

127. Rojo, R. *et al.* Deletion of a Csf1r enhancer selectively impacts CSF1R expression and

development of tissue macrophage populations. *Nat. Commun.* **10**, (2019).

128. Kettenmann, H., Hanisch, U.-K., Noda, M. & Verkhratsky, A. Physiology of Microglia. *Physiol. Rev.* (2011) doi:10.1152/physrev.00011.2010.

129. Bos, P. D. *et al.* Genes that mediate breast cancer metastasis to the brain. *Nature* (2009) doi:10.1038/nature08021.

130. Lorger, M. & Felding-Habermann, B. Capturing changes in the brain microenvironment during initial steps of breast cancer brain metastasis. *Am. J. Pathol.* **176**, 2958–2971 (2010).

131. Kienast, Y. *et al.* Real-time imaging reveals the single steps of brain metastasis formation. *Nat. Med.* **16**, 116–122 (2010).

132. Valiente, M. *et al.* Serpins promote cancer cell survival and vascular Co-option in brain metastasis. *Cell* (2014) doi:10.1016/j.cell.2014.01.040.

133. Chen, Q. *et al.* Carcinoma-astrocyte gap junctions promote brain metastasis by cGAMP transfer. *Nature* **533**, 493–498 (2016).

134. Priego, N. *et al.* STAT3 labels a subpopulation of reactive astrocytes required for brain metastasis. *Nat. Med.* **24**, 1481 (2018).

135. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

136. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat.*

Biotechnol. **36**, 411–420 (2018).

137. Motenko, H., Neuhauser, S. B., O'Keefe, M. & Richardson, J. E. MouseMine: a new data warehouse for MGI. *Mamm. Genome* **26**, 325–330 (2015).

138. Zlotnik, A. & Yoshie, O. Chemokines: A new classification system and their role in immunity. *Immunity* (2000) doi:10.1016/S1074-7613(00)80165-X.

139. Griffith, J. W., Sokol, C. L. & Luster, A. D. Chemokines and Chemokine Receptors: Positioning Cells for Host Defense and Immunity. *Annu. Rev. Immunol.* **32**, 659–702 (2014).

140. Taggart, D. *et al.* Anti–PD-1/anti–CTLA-4 efficacy in melanoma brain metastases depends on extracranial disease and augmentation of CD8 T cell trafficking. *Proc. Natl. Acad. Sci.* (2018) doi:10.1073/pnas.1714089115.

141. Contreras-Zárate, M. J. *et al.* Estradiol induces BDNF/TrkB signaling in triple-negative breast cancer to promote brain metastases. *Oncogene* **38**, (2019).

142. Li, Q. *et al.* Developmental heterogeneity of microglia and brain myeloid cells revealed by deep single-cell RNA sequencing. *bioRxiv* 406363 (2018) doi:10.1101/406363.

143. Blasius, A. L. *et al.* Bone Marrow Stromal Cell Antigen 2 Is a Specific Marker of Type I IFN-Producing Cells in the Naive Mouse, but a Promiscuous Cell Surface Antigen following IFN Stimulation. *J. Immunol.* (2006) doi:10.4049/jimmunol.177.5.3260.

144. Neil, S. J. D., Zang, T. & Bieniasz, P. D. Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. *Nature* (2008) doi:10.1038/nature06553.

145. Ting, J. P. Y. & Trowsdale, J. Genetic control of MHC class II expression. *Cell* (2002) doi:10.1016/S0092-8674(02)00696-7.

146. Schröder, B. The multifaceted roles of the invariant chain CD74 - More than just a chaperone. *Biochimica et Biophysica Acta - Molecular Cell Research* (2016) doi:10.1016/j.bbamcr.2016.03.026.

147. Butovsky, O. *et al.* Identification of a unique TGF-β-dependent molecular and functional signature in microglia. *Nat. Neurosci.* (2014) doi:10.1038/nn.3599.

148. Gosselin, D. *et al.* An environment-dependent transcriptional network specifies human microglia identity. *Science (80-.).* (2017) doi:10.1126/science.aal3222.

149. Bennett, M. L. *et al.* New tools for studying microglia in the mouse and human CNS. *Proc. Natl. Acad. Sci. U. S. A.* (2016) doi:10.1073/pnas.1525528113.

150. Watanabe, H., Numata, K., Ito, T., Takagi, K. & Matsukawa, A. Innate immune response in Th1- and Th2-dominant mouse strains. *Shock* (2004) doi:10.1097/01.shk.0000142249.08135.e9.

151. Mills, E. L. *et al.* Succinate Dehydrogenase Supports Metabolic Repurposing of Mitochondria to Drive Inflammatory Macrophages. *Cell* **167**, 457-470.e13 (2016).

152. Lauro, C. & Limatola, C. Metabolic Reprograming of Microglia in the Regulation of the Innate Inflammatory Response. *Front. Immunol.* **11**, 1–8 (2020).

153. Rongvaux, A. *et al.* Development and function of human innate immune cells in a humanized mouse model. *Nat. Biotechnol.* (2014) doi:10.1038/nbt.2858.

154. McQuade, A. *et al.* Development and validation of a simplified method to generate human microglia from pluripotent stem cells. *Mol. Neurodegener.* (2018) doi:10.1186/s13024-018-0297-x.

155. Hasselmann, J. *et al.* Development of a Chimeric Model to Study and Manipulate Human Microglia In Vivo. *Neuron* (2019) doi:10.1016/j.neuron.2019.07.002.

156. McGinnis, C. S. *et al.* MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* (2019) doi:10.1038/s41592-019-0433-8.

157. Elmore, M. R. P. *et al.* Replacement of microglia in the aged brain reverses cognitive, synaptic, and neuronal deficits in mice. *Aging Cell* (2018) doi:10.1111/acel.12832.

158. Spangenberg, E. *et al.* Sustained microglial depletion with CSF1R inhibitor impairs parenchymal plaque development in an Alzheimer's disease model. *Nat. Commun.* (2019) doi:10.1038/s41467-019-11674-z.

159. Hammond, T. R. *et al.* Single-Cell RNA Sequencing of Microglia throughout the Mouse Lifespan and in the Injured Brain Reveals Complex Cell-State Changes. *Immunity* **50**, 253-271.e6 (2019).

160. Keren-Shaul, H. *et al.* A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169**, 1276-1290.e17 (2017).

161. Mathys, H. *et al.* Temporal Tracking of Microglia Activation in Neurodegeneration at Single-Cell Resolution. *Cell Rep.* **21**, 366–380 (2017).

162. Ries, C. H. *et al.* Targeting tumor-associated macrophages with anti-CSF-1R antibody reveals a strategy for cancer therapy. *Cancer Cell* **25**, 846–859 (2014).

163. Bruttger, J. *et al.* Genetic Cell Ablation Reveals Clusters of Local Self-Renewing Microglia in the Mammalian Central Nervous System. *Immunity* **43**, 92–106 (2015).

164. Han, J., Harris, R. A. & Zhang, X. M. An updated assessment of microglia depletion: Current concepts and future directions. *Mol. Brain* **10**, 1–8 (2017).

165. Ley, K. The second touch hypothesis: T cell activation, homing and polarization. *F1000Research* **3**, 1–15 (2014).

166. Campbell, J. P., Merkel, A. R., Masood-Campbell, S. K., Elefteriou, F. & Sterling, J. A. Models of Bone Metastasis. *J. Vis. Exp.* (2012) doi:10.3791/4260.

167. Heaton, H. *et al.* Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020).

168. Andersen, H. & Hepburn, B. Scientific Method. in *The Stanford Encyclopedia of Philosophy* (ed. Zalta, E. N.) (Metaphysics Research Lab, Stanford University, 2020).

169. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).

170. Wang, D. *et al.* Identification of multipotent mammary stemcells by protein C receptor expression. *Nature* **517**, 81–84 (2015).

171. Colacino, J. A. *et al.* Heterogeneity of Human Breast Stem and Progenitor Cells as Revealed by Transcriptional Profiling. *Stem Cell Reports* **10**, 1596–1609 (2018).

172. Wang, C., Christin, J. R., Oktay, M. H. & Guo, W. Lineage-Biased Stem Cells Maintain Estrogen-Receptor-Positive and -Negative Mouse Mammary Luminal Lineages. *Cell*

Rep. **18**, 2825–2835 (2017).

173. Van Keymeulen, A. *et al.* Lineage-Restricted Mammary Stem Cells Sustain the Development, Homeostasis, and Regeneration of the Estrogen Receptor Positive Lineage. *Cell Rep.* **20**, 1525–1532 (2017).

174. Shehata, M. *et al.* Proliferative heterogeneity of murine epithelial cells in the adult mammary gland. *Commun. Biol.* **1**, 1–10 (2018).

175. Mishra, D. & Banerjee, D. Lactate dehydrogenases as metabolic links between tumor and stroma in the tumor microenvironment. *Cancers (Basel).* **11**, (2019).

176. Ma, D. *et al.* Patient-derived xenograft culture-transplant system for investigation of human breast cancer metastasis. *bioRxiv* 2020.06.25.172056 (2020) doi:10.1101/2020.06.25.172056.

177. Lin, Y., Xu, J. & Lan, H. Tumor-associated macrophages in tumor metastasis: Biological roles and clinical therapeutic applications. *J. Hematol. Oncol.* **12**, 1–16 (2019).

178. Watanabe, H., Numata, K., Ito, T., Takagi, K. & Matsukawa, A. Innate immune response in Th1- and Th2-dominant mouse strains. *Shock* **22**, 460–466 (2004).

APPENDIX A: Combined ordering genes for human breast epithelial trajectory inference

Table provides list of genes used as input for a combined Monocle2 trajectory across human breast epithelial cell states (Related to **Fig 2.5A**). Gene module clusters (based on correlations) and the most common cell state association for each set of module genes are also included for reference.

Gene	Module	Top cell state associations for module	Gene	Module	Top cell state associations for module
TPM2	1	Basal/Myoepithelial	TUBB6	8	Basal/Myoepithelial
TP63	1	Basal/Myoepithelial	SFN	8	Basal/Myoepithelial
TAGLN	1	Basal/Myoepithelial	PRMT1	8	Basal/Myoepithelial
SPARCL1	1	Basal/Myoepithelial	MRPS6	8	Basal/Myoepithelial
SPARC	1	Basal/Myoepithelial	KRT5	8	Basal/Myoepithelial
S100A2	1	Basal/Myoepithelial	ITGB1	8	Basal/Myoepithelial
MYLK	1	Basal/Myoepithelial	IL24	8	Basal/Myoepithelial
MYL9	1	Basal/Myoepithelial	IL20	8	Basal/Myoepithelial
MT2A	1	Basal/Myoepithelial	GSTP1	8	Basal/Myoepithelial
MT1X	1	Basal/Myoepithelial	GSTO1	8	Basal/Myoepithelial
MT1E	1	Basal/Myoepithelial	GAS6	8	Basal/Myoepithelial
MT1A	1	Basal/Myoepithelial	FHL2	8	Basal/Myoepithelial
MRGPRX3	1	Basal/Myoepithelial	FBXO2	8	Basal/Myoepithelial
KRT17	1	Basal/Myoepithelial	DRAP1	8	Basal/Myoepithelial
KRT14	1	Basal/Myoepithelial	CD82	8	Basal/Myoepithelial
IGFBP7	1	Basal/Myoepithelial	CAV1	8	Basal/Myoepithelial
IGFBP6	1	Basal/Myoepithelial	AREG	8	Basal/Myoepithelial
IGFBP4	1	Basal/Myoepithelial	ACTN1	8	Basal/Myoepithelial
IGFBP2	1	Basal/Myoepithelial	TM4SF1	3	Mixed Luminal
ID1	1	Basal/Myoepithelial	MUC1	3	Mixed Luminal
FEZ1	1	Basal/Myoepithelial	EPCAM	3	Mixed Luminal
DST	1	Basal/Myoepithelial	DAPP1	3	Mixed Luminal
CALML3	1	Basal/Myoepithelial	CIB1	3	Mixed Luminal
CALD1	1	Basal/Myoepithelial	ATP1B1	3	Mixed Luminal
BHLHE41	1	Basal/Myoepithelial	POLR2L	7	Unknown/Unclassified
APOE	1	Basal/Myoepithelial	ATP5G1	7	Unknown/Unclassified

AKR1B1	1	Basal/Myoepithelial	CYCS	5	Unknown
TPT1-AS1	9	Luminal 1.1	TSPAN1	4	Luminal 2
SERPINB7	9	Luminal 1.1	TNFSF10	4	Luminal 2
SERPINB4	9	Luminal 1.1	TFPI	4	Luminal 2
SERPINB3	9	Luminal 1.1	STC2	4	Luminal 2
SAA2	9	Luminal 1.1	S100P	4	Luminal 2
S100A8	9	Luminal 1.1	PTGR1	4	Luminal 2
RCAN1	9	Luminal 1.1	ORM1	4	Luminal 2
RARRES1	9	Luminal 1.1	LIMCH1	4	Luminal 2
PROM1	9	Luminal 1.1	HSPB1	4	Luminal 2
PHLDA1	9	Luminal 1.1	GOLM1	4	Luminal 2
OVOS2	9	Luminal 1.1	EFHD1	4	Luminal 2
OLFM4	9	Luminal 1.1	EDN1	4	Luminal 2
NDRG2	9	Luminal 1.1	DNAJC12	4	Luminal 2
MESP1	9	Luminal 1.1	CD99	4	Luminal 2
MALL	9	Luminal 1.1	C8orf4	4	Luminal 2
LTF	9	Luminal 1.1	ANKRD30A	4	Luminal 2
GLRX	9	Luminal 1.1	AGR2	4	Luminal 2
GABRP	9	Luminal 1.1	SMS	2	Luminal 1.2
FDCSP	9	Luminal 1.1	HSPA1A	2	Luminal 1.2
CXCL17	9	Luminal 1.1	CCND1	2	Luminal 1.2
ANKRD36C	9	Luminal 1.1	C4orf48	2	Luminal 1.2
ALDH1A3	9	Luminal 1.1			

APPENDIX B: Gene signatures for BCBM-R microglia subpopulations

Table provides marker genes for BCBM-R microglia subpopulations (excluding Cycling) from Foxn1$^{nu/nu}$ data compared to all other myeloid cells from mice with BCBM, translated to their human equivalents using biomaRt in R. Related to **Fig 4.5.1**.

APC		IFN Responsive	Secretory	Glycolytic		
ISG15	ID2	ISG15	CD52	MT-ATP6	CXCL16	TMSB4Y
MT-CO1	BST2	IFI44L	RHOC	ISG15	RPL26	RPS18
CD52	P2RY14	TOR3A	PRDX1	MT-ND1	RPL35A	AIF1
VCAM1	CD300LF	GBP2	CSF1	ENO1	UQCR11	COX7A2
GBP2	CCL2	IFI44	RGS1	RPL22	SNRPG	RPS4X
FCGR3A	ITGB2	FCGR1B	CH25H	RPL11	RPL23	EIF3E
FCGR3B	SLFN12	FCGR1A	ATF3	CD52	CST7	PABPC1
GBP6	SLFN12L	MNDA	SLC15A3	UQCRHL	RPS15	RPL8
MNDA	CD86	IFIT2	RNF121	RPS8	CCL5	RPS14
IFI44L	ICAM1	IFIT3	RAB7B	GNG5	CCL7	CXCL13
IFI44	CCL4	IFITM2	CTSD	GBP2	HMOX1	ATP5MF
TOR3A	CCL4L2	IFITM1	LDHA	AKR1A1	TSPO	
SLAMF8	CD33	IFITM3	HCAR2	PRDX1	RPS19	
PRDX1	SIGLEC6	TRIM5	HCAR3	UQCRH	SPP1	
MS4A6A	ZBP1	IRF7	TNFRSF12A	RPS13	EEF2	
MS4A6E	CXCL10	UBE2L6	CADM1	EEF1G	RPL32	
SRGN	C19orf38	IFIT1B	ALDOA	FAU	SELENOW	
IFIT1B	APOE	HCAR2	LGALS3	EIF5AL1	RPL38	
IL18BP	LGALS3BP	HCAR3	CD63	PGAM1	TMSB10	
IFIT2	CXCL9	PHF11	PKM	EIF3F	RPS3A	
PRDX5	STAT1	STAT2	CD9	RPS3	RPL22L1	
CH25H	CD40	MX1	RPS2	RPLP2	RPL24	
PSAP	SLFN5	LGALS3BP	C3AR1	ATP5F1C	RPL19	
LDHA	TLR2	CXCL10	CST7	COX8A	RTCB	
IRF7	ARL5C	PARP14	MT1G	PRDX5	CXCL9	
IFITM2	RPL19	SP140	PLEK	LDHA	COX7C	
IFITM1	RSAD2	CCL4	CCRL2	IRF7	RPL36	
IFITM3	NAAA	CCL4L2	CCL15-CCL14	IFITM2	COTL1	
IFIT3	SAMHD1	ZBP1	CCL15	IFITM1	RPL18A	
CD69	CCL18	IFI35	CCL23	IFITM3	RPS21	
CD63	CCL3	XAF1	CSTB	SERF2	COX6B1	
RPS2	CCL3L1	CCL18	ANKH	NACA	COPS9	

B2M	CCL3L3	CCL3	CD14	ATP5F1B	RPL34
COX6A2	USP18	CCL3L1	FTL	RPL21	RPL28
ALDOA	USP41	CCL3L3	CAPG	COX6A2	HINT1
PSME1	CCRL2	SLFN5	APOE	SLC25A3	RPL31
ITGAX	IL1B	USP18	CCL18	RPS2	RPS28
SERPINA3	NFE2L2	USP41	CCL3	CD63	GPI
PKM	TAP2	CCL2	CCL3L1	RPL4	ZBP1
LAG3	IL2RG	EIF2AK2	CCL3L3	PFDN5	EIF5A
CTSC	TAP1	TSPO	NCEH1	PSME2	APOE
NPC2	HLA-DQB1	CMTM8	OSM	ELOB	COX7A2L
LGALS3	HLA-DQB2	RTP4	IL1B	NPC2	COX4I1
STAT2	LPL	SLFN12	TLR2	RPL6	RPL29
PSME2	HLA-DMB	SLFN12L	HMOX1	PSME1	RPL37
HCAR2	TAPBP	STAT1	C5AR1	TPT1	LGALS3BP
HCAR3	CRLF2	CCL8	CCL4	ALDOA	UQCRQ
BCL2A1	C4A	BST2	CCL4L2	B2M	RPL3
PHF11	C4B	DHX58	EIF4A1	RPS17	RPS27A
CSTB	CD36	HERC6	MFSD12	RPS25	ATP5F1E
CCL8	CSF2RA	IRGM	SPP1	ATP5MG	C19orf38
MIF	HLA-DOA	LY6E	GNAS	GATM	COX7B
CXCL16	FGL2	FGL2	RPL32	LGALS3	RPL12
SLC11A1	CD83		CXCL16	PKM	EIF3H
CXCL13	CD72		LGALS1	RPL18	CD72
SELENOW	CD274		PLAUR	TBCA	RPS6
GRN	CD74		PLD3	FTL	PSMB1
RPS19	ASS1		ID2	BST2	ASS1
TSPO	CYBB		MIF	RPL14	CD74
SPP1	TNF		GADD45B	RPS5	RPL7
HMOX1	PIM1		CTSZ	AXL	ATOX1
PARP14	CTSB		PMP22	RPS16	COX6C
CCL7	IRGM		RPL12	EIF3K	PSMB8
CCL5	CDKN1A		RPL35	EEF1B2	RPL35
CST7	PSMB8		TNF	UBA52	RPL30
RPL32	HLA-DMA		LPL	RPL13	RACK1
RPS5	PGK1		CD83	RPS9	RPL36A
CTSZ	FTL		CTSB	ATP5MGL	PGK1
C3			IER3	CCL8	TXN
AXL			SERPINE1	MIF	HLA-DQB1
IL1RN			FAM20C	CSTB	HLA-DQB2