In Situ Gene Expression Profiling of Heterogeneous Cancer Cell States in Breast and Lung Carcinomas

---

A Dissertation

Presented to
the faculty of the School of Engineering and Applied Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

---

by

Shambhavi Singh

August 2020

APPROVAL SHEET

This dissertation is in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biomedical Engineering at the University of Virginia

Shambhavi Singh, Author

This dissertation has been read and approved by the examining committee:

Dr. Kevin A. Janes, Dissertation Advisor
Department of Biomedical Engineering

Dr. Gustavo K. Rohde, Committee Chair
Department of Biomedical Engineering

Dr. Francine E. Garrett-Bakelman, Committee Member
Department of Biochemistry and Molecular Genetics

Dr. Amy H. Bouton, Committee Member
Department of Microbiology, Immunology, and Cancer Biology

Dr. Matthew J. Lazzara, Committee Member
Departments of Chemical Engineering & Biomedical Engineering

Accepted for the School of Engineering and Applied Science:

Dr. Craig H. Benson, Dean
School of Engineering and Applied Science

Abstract

For the diagnosis and treatment of cancers, it is often assumed that all cells in a tumor are identical. However, solid tumors are composed of cells that differ in cell-type, genotype, and phenotype. Individual cancer cells in tumors regulate their behavior in response to complex internal and external cues. Together, these differences result in heterogeneous cancer cell states that influence tumor growth, metastatic progression, and treatment response. Characterizing the nature and prevalence of heterogeneous cancer cell states is fundamental to understanding why patients diagnosed with the same disease often have variable outcomes.

In this dissertation, we present experimental and bioinformatics approaches to measure heterogenous cancer cell states in breast and lung carcinomas. We coupled laser capture microdissection with sequencing measurements to obtain transcriptomic data from groups of 10 cancer cells in their native context within tumors. This profiling method has improved measurement sensitivity compared to existing single-cell transcriptomic methods, enabling us to deeply interrogate cancer cell transcriptomes. Analyzing 10-cell transcriptomes with an abundance-based dispersion metric, we identified heterogeneously expressed genes that represent different cancer cell states.

To identify early differences between cells that may influence patient outcomes and treatment responses, we profiled five biopsy samples from patients with luminal breast cancer. We detected thousands of heterogeneously expressed genes in individual tumors that comprise many pathways relating to proliferation, immune response, and stress tolerance. Moreover, we identified a recurrent set of genes that are heterogeneously expressed in multiple breast tumors. Genes in this set suggest that breast cancer cells sporadically activate pathways that are known to drive other types of cancer.

To systematically measure the influence of heterotypic interactions on cancer cells, we profiled 3D cell culture and murine models of small cell lung cancer (SCLC). We profiled SCLC cells in isolated 3D cultures and metastatic liver colonies to decode the influence of heterogeneous tumor microenvironments on cancer cell states. We observed a shift in the plasticity of SCLC cells upon liver colonization, and identified an expanded set of heterogenous states that expressed markers of multiple cell-types.

In this dissertation, we interrogated heterogenous cancer cell states within isolated cells, solid tumors, and metastases. The findings presented here provide novel insight into the transcriptional landscapes of breast cancer and lung cancer cells, towards the goal of understanding differential outcomes for patients with these diseases.

Acknowledgments

Having never considered a career as an engineer, it has truly been my great fortune to obtain my doctoral training at the Department of Biomedical Engineering at UVA. It has been the most educational, rigorous, collaborative, and supportive environment, and I would like to dedicate this dissertation to the many people who have contributed to my academic and personal growth over my graduate career.

First, I would like to thank my advisor, Dr. Kevin Janes for being the quintessential mentor and possessing so many invaluable qualities that a list would be impossible. You have taught me to bring rigor and careful consideration to every aspect of science, be it experimental design or statistical analysis, reading or writing. Thank you for the care and time you have put into my development, and I hope I can emulate your many virtues in my own career.

I would also like to thank my dissertation committee: Dr. Gustavo Rohde, Dr. Matthew Lazarra, Dr. Francine-Garrett Bakelman, and Dr. Amy Bouton. Thank you for your evaluation of my work and bringing your diverse perspectives to the work I was doing; I have benefitted greatly from all of your feedback. I would also like to thank Dr. Kristen Atkins for her pathology expertise and all our coffee chats about cellular morphology, science communication, and translational research. A big thank you also to Dr. Jennifer Harvey for enabling our access to clinical samples and Kathy Repich for helping me navigate the complicated IRB waters.

A lot of the presented in this dissertation was enriched greatly by the teamwork between Dylan Schaff, Matthew Sutcliffe, and myself aka #thedreamteam. I'm glad we were able to commiserate about the LCM making scary sounds or the cryostat exploding or R forgetting its own functions. Together, we always managed to come up with solutions!

I have been very grateful for the fantastic environment in the Janes lab overall – it takes a special group of people to make long, difficult days in the lab as fun as they have been! Thank you to the past members Sameer, Millie, and Chris who taught me so much when I first started and have continued to provide support and advice. To Andrew, Taylor, Wisam, and Anthony – I'm excited to watch you guys progress in your careers, with expert advice form Bishal and Cameron. Lixin thanks for training me when I started, and Cheryl, thank you so much for keeping the lab going as smoothly as it does. Please invite me for happy hours - I will not be far!

I could not imagine graduate school without Liz and TK – thank you for your dependable support, wisdom, and deep understanding through this shared process. I'm glad we were never more than a few feet away, either in MR5 or at 1800 JPA. Life in Charlottesville is also incomplete without my most thoughtful and loving friends, Casey Morrison and Liz Hoang.

None of this would have been possible without my fiercest supporters – my family. Thank you to my parents who embody unconditional love and have always taught me that my dreams have no limits. And to my family in Charlottesville, my fiancé Clayton (and our dog, Lola) who provide unwavering love and encouragement to me at the times I need it the most. I am so lucky to have you all as my family, and I love you.

Table of Contents

Table of Figures

Table of Tables

1 Introduction

1.1 Different scales and types of tumor heterogeneity

Tumors initiate when a single cell acquires genetic lesions that confer growth advantages (1–3). Through many generations of cell division, the mutant cancer cell grows into a multicellular tumor made up of millions of cells. Across human tumors, there is variation in the exact genetic lesions that initiate tumorigenesis (4–7). Further, tumors arise from a variety of cell types, in nearly every organ of the body (8). Tumors therefore vary in their genetics, lineages, and microenvironments (9,10).

1.1.1 Inter-tumor heterogeneity

Inter-tumor heterogeneity refers to differences in tumors from different patients (**Figure 1.1A**). Tumors vary in tissue and lineage of origin, resulting in broad tumor types like breast cancer or lung cancer. These broad tumor types are further subcategorized based on molecular characterizations. For example, patients with breast cancer are diagnosed to have one of four subtypes, Luminal A, Luminal B, HER2-enriched, and basal-like. Extensive study of inter-tumor differences between breast cancer patients has identified distinct prognostic features for the four subtypes (11,12). The first two subtypes, Luminal A and B, are defined by tumor cell expression of the hormone receptors, estrogen receptor (ER) and/or progesterone receptor (PR). The third subtype is HER2-amplified tumors, defined by overexpression of the human epidermal growth factor receptor-2 or HER2. Basal-like breast tumors are defined as “triple-negative” by a lack of expression of ER, PR, and HER2 (13).

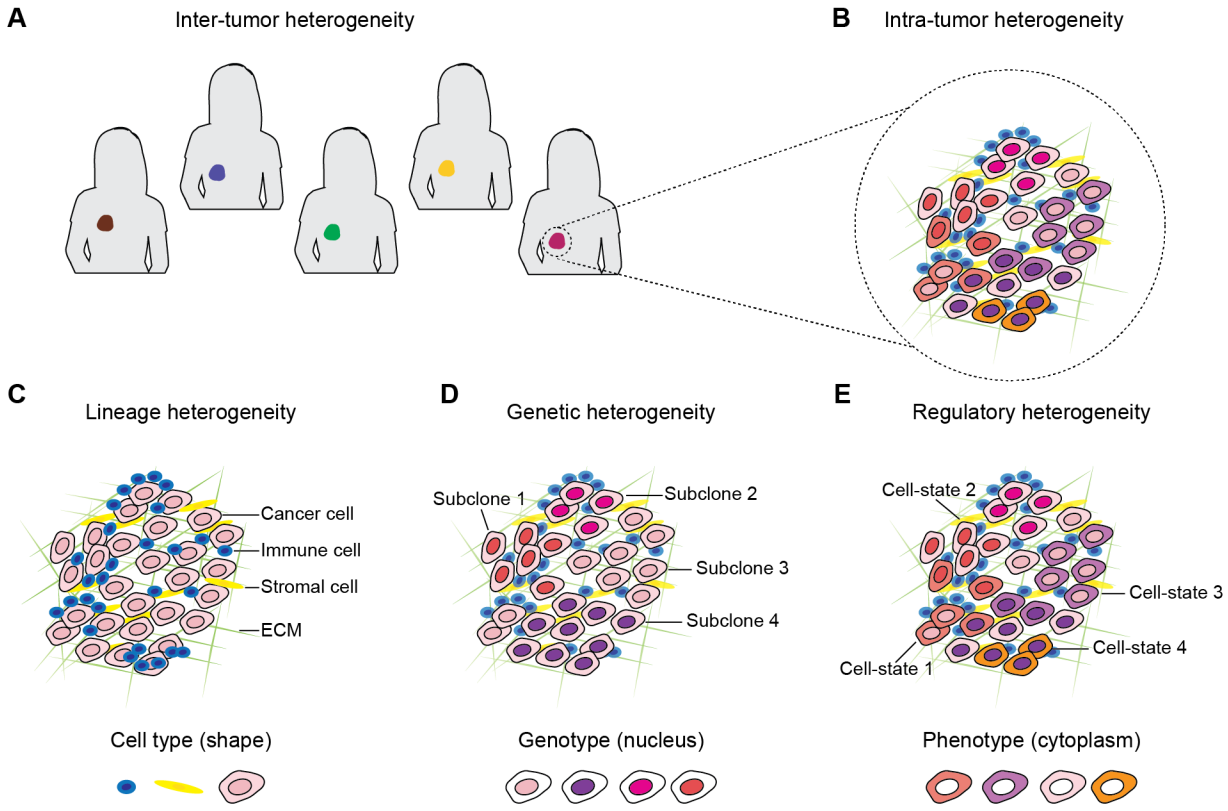

Figure 1.1 Different scales and types of tumor heterogeneity
(A) Differences between individual patients' tumors display inter-tumor heterogeneity when evaluated in bulk.
(B) A simplified illustration of the complete tumor microenvironment. Within individual tumors, single cells display heterogeneity in lineage, genetics, and regulatory states
(C) Cells of different developmental lineages comprise individual tumors. Lineage differences are illustrated by cell shape and color: mutant tumor cells (pink), stromal cells (yellow), immune cells (blue). The extracellular matrix is depicted in green.
(D) Different genetic subclones of cancer cells exist in a single tumor. Colors of nuclei indicate genotype differences of subclones
(E) Cancer cells also exist in diverse regulatory cell-states. Colors of the cytoplasm indicate phenotype variations between cancer cells.

Tumor subcategorization has also led to targeted therapies for these subtypes such as targeted anti-estrogen therapy for luminal tumors and anti-HER2 therapy for HER2-amplified tumors (14,15). However, many patients within a single subtype still have unpredictable clinical outcomes and incomplete responses to therapy, indicating further variability amongst tumors (16,17). Measurements of inter-tumor heterogeneity are usually made in bulk and assume that all cells within individual tumors are identical, ignoring heterogeneity within tumors (18). Differences at the level of individual cells are categorized as intra-tumor heterogeneity and are associated with variable prognosis and treatment response in many tumor types (19–21).

1.1.2 Intra-tumor heterogeneity

Intra-tumor heterogeneity refers to many types of variations between the cells that comprise a tumor (**Figure 1.1B**). As tumors progress, mutant cancer cells interact with their neighboring normal cells in the tissue and with immune cells that respond to the tumor as it grows (22–24). The complete tumor microenvironment comprises the tumor cells, supporting tissue-resident stromal cells, a variety of immune cells, and the tissue extracellular matrix (25,26). Within single tumors there are many permutations of interactions and microenvironments that can influence cellular genotypes and phenotypes. The following subsections define the different categories of intra-tumor heterogeneity and their influence on tumor growth and progression.

1.1.2.1 Lineage heterogeneity between cell types in tumors

A solid tumor is comprised of cells that have heterogeneous developmental lineages, such as cancer cells that arise from local cells in the tissue and immune cells that derive from the hematopoietic system (**Figure 1.1C**). Differences in cell type (here used interchangeably with cell lineage) arise throughout the body during normal development, as progenitor cells

undergo differentiation to produce cell types with specialized functions (27) . For example, airways in the normal lung are lined with club cells that secrete antimicrobial peptides, goblet cells that secrete mucus, stromal fibroblasts that create extracellular matrix, and stem cells that regenerate the other cell types (28–33). When tumors arise in the lung, different non-malignant stromal cells remain present alongside the malignant cancer cells, and can even support tumor growth (34). Tumors also invoke inflammatory responses which cause lymphocytes to migrate and infiltrate the tumor, increasing the lineage heterogeneity within tumors (35,36).

Measurements of lineage heterogeneity in tumors have resulted in important insights for patient treatments (35,37). In certain lung cancers, increased numbers of CD8+ effector T-cell lymphocytes are associated with improved outcomes for patients (37,38). Further characterization of T-cells across multiple tumor types have identified similar effector T-cells with tumor suppressive effects, catalyzing the development of several novel strategies for cancer immunotherapy like adoptive cell transfer and checkpoint blockade (39–43). For personalized medicine, cataloging the nature and proportions of lymphocytes within individual tumors continues to be an important avenue to stratify patients for response to immunotherapy agents (44,45).

Stromal cells that normally reside in the tissue can play diverse tumor supportive roles, like cancer-associated fibroblasts (CAFs) that have been a focus of study in breast tumors (46–48). In normal breast tissue, fibroblasts secrete and turnover the ECM and assist in epithelial cell differentiation during puberty (46). In tumors, CAFs participate in paracrine signaling via secreted ligands like HGF that support growth of the cancer cells, as well as chemokines like CXCL12 that promote invasion by causing tumor cells to take on a migratory phenotype (9,49,50). CAFs were also found to induce treatment resistance to both chemotherapies and anti-estrogen therapies when co-cultured with MCF7 breast cancers cells (51). In human breast cancers, heterogeneity within CAFs has been identified. A subset of CAFs that express CD146 can promote resistance to anti-estrogen treatment by reducing cancer cell dependence on

signaling through the estrogen receptor (52). These multifaceted roles for CAFs demonstrate that heterotypic interactions between different cell-types in a tumor can modulate treatment resistance and tumor outcomes.

This section introduced lineage heterogeneity in tumors and how heterogenous cell lineages can influence tumor progression and patient outcome. The following two subsections detail heterogeneities that arise within a cell type, between the cancer cells within tumors.

1.1.2.2 Genetic heterogeneity between cancer cells

Malignant transformation of normal cells to cancer occurs through multiple DNA modifications such as mutations, deletions, and amplifications (4,53,54). As cancer cells continue to replicate, individual cells develop different genetic lesions, resulting in genetic heterogeneity within tumors (55). On a single-cell level, genetic changes are irreversible and mutations are passed on to daughter cells, leading to subclones within tumors that have different genotypes (**Figure 1.1D**) (56–58). Single-cell genetic methods have been developed to provide insight into the evolution of intra-tumor genetic variations as tumors initiate, progress, and ultimately metastasize (19,59–61).

Studies of genetic heterogeneity in individual tumors of breast cancer have substantiated a punctuated clonal evolution model (62,63). In this model, bursts of large-scale genomic changes occur early in cancer cells resulting in copy number alterations (CNAs) that cause gains and losses of various genes (59). Many ER positive breast tumors show recurrent early copy number gains in chromosome arms 1q and 8q, which are largely retained in clones throughout a tumor (60,64). Further diversity arises within subclones through individual point mutations within genes that accumulate over time and create subbranches within tumors (60). Tracking evolutionary dynamics of genetic heterogeneity has shed light on the initiation of the metastatic process, demonstrating that it can often begin early during the establishment of a primary tumor (65).

In breast cancers, rare mutations present in <1% of cancer cells have been identified that modulate tumor evolution and treatment response (19,60). But for the remaining majority of mutations, relating specific DNA changes in cancer cells to their resultant phenotypes remains a challenge. The overall mutational burden within individual tumors can vary drastically between tumor types, with some tumors like triple-negative breast cancer having 13-fold higher mutation rates than other tumor types (60). In tumors with high mutational burdens, growth driving genetic changes are often accompanied by a large swathe of passenger mutations whose functional consequences are unclear (66–69). It becomes necessary to measure the repertoire of functional states of cancer cells to decipher the biological implications of genetic lesions.

1.1.2.3 Regulatory heterogeneity between cancer cells

Due to differences in the regulation of gene and protein expression, cancer cells display regulatory heterogeneity that causes them to have variable phenotypes (**Figure 1.1E**) (70). In this context, a regulatory cell state is defined as a set pattern of co-expressed genes and proteins that coordinate specific cellular phenotypes (71–74). Regulatory variations occur as cells transition between different expression states in reversible and context dependent ways (75,76). The different phases of the cell cycle are examples of cell states; cells express a consistent patterns of genes and proteins in the G2/M phases to enable proliferation, which are turned off when cells are in the G1/S phases (77). Expression patterns of proliferative genes have been used diagnostically in breast cancer to stratify patients for chemotherapy, and recent studies have identified that proportions of cells in G2/M cell-states are prognostic in melanoma and glioblastoma samples (78,79).

Regulatory heterogeneity can arise between genetically identical cells leading to different cell-states (**Figure 1.2**) (54–57). This occurs as cancer cells integrate both internal (e.g. genetic) cues and external (e.g. microenvironmental) cues to modulate their cellular states (80,81). In large tumors, regulatory variations arise in cancer cells due to spatial differences in

nutrient availability. Cells in the interior of tumors face a more hypoxic environment than cells on the periphery, which are closer to vasculature. Hypoxia-induced factors respond to the change in oxygen availability and pH to induce regulatory changes that switch cellular metabolism to hypoxia-tolerant pathways (82–85). Conceptually, such non-genetic regulatory changes allow cancer cells to rapidly respond to stimuli and survive longer while they undergo the necessary genetic changes.

Paracrine and juxtracrine interactions with other cell types in the tumor microenvironment also influence the regulatory states of cancer cells (86,87). For example, cancer cells that comprise the leading edge on the periphery of tumors interact with heterotypic cell-types and have different functional roles compared to cells in the interior of a tumor (88). Multiple experimental models of breast cancer have demonstrated that cancer cell migration involves cooperation with other cell types (89–91). In these models, non-malignant stromal fibroblasts and macrophages lead malignant cancer cells outside the basement membrane into surrounding tissue. The molecular composition of the ECM and its biophysical characteristics also regulate leading edge cancer cells to activate cytoskeleton and matrix remodeling pathways, allowing the tumor to degrade the basement membrane and invade into the tissue parenchyma (92–95).

Some mechanisms of regulatory variation create very drastic differences in cell-state, such that cancer cells begin to represent lineages other than the ones from which the tumor arose (96). This hijacking of cellular regulatory mechanisms to enable specific, long-lived cell-states is called “lineage plasticity”. A well-studied example of this is epithelial-to-mesenchymal transition, where epithelial carcinoma cells de-differentiate into a mesenchymal state that does not require anchorage and is able to migrate through vasculature (97–99). More recent studies have identified new forms of lineage plasticity especially in tumors where gene functions of *TP53* and/or *RB1* are lost. In combination with treatment, these genomic changes enable epithelial cells to take on neuroendocrine lineages that are resistant to many treatments

(100,101). Treatment induced reprogramming into a mesenchymal-like state has also been identified across multiple tumors, where cancer cells can adopt a “persistor” cell-state that is treatment resistant (102–104).

Intra-tumor regulatory heterogeneity has many causes and results in differential rates of growth for tumors, creating uncertainty in existing models for patient prognosis (105–107). Several studies have shown that intra-tumor heterogeneity is further amplified upon treatment, as diverse single cells within a tumor react heterogeneously to interventions (19,20,104,108,109). Therefore, pre- and post-treatment induced heterogeneity together contribute to uncertainty in prognosis and differential responses to therapy. Cancer cell regulatory heterogeneities, resulting from complex integration of internal and external cues are the main focus of this dissertation.

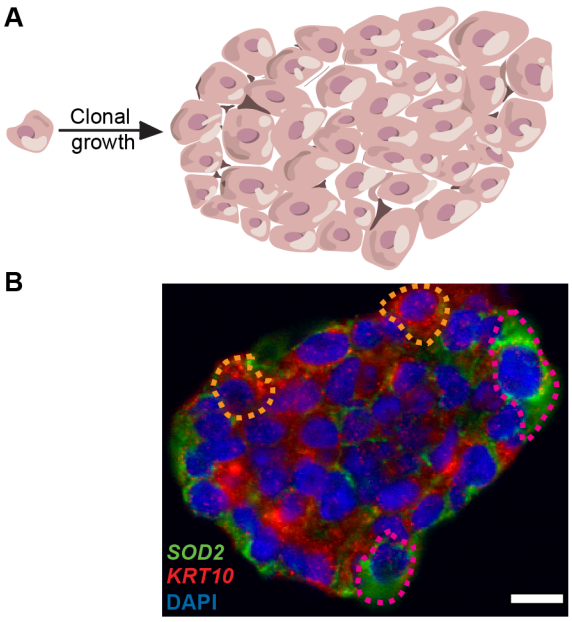

Figure 1.2 Genetically identical cells display functional heterogeneity
(A) Schematic of 3D cell culture model of clonal multicellular spheroid growth from one single cell. **(B)** Gene expression of *SOD2* and *KRT10* in MCF10A-5E spheroids as detected by RNA FISH measurements. All cells in the 3D spheroid are clonally derived from a single cell as in **(A).** Dashed lines represent states in two different expression states, cell state #1 (orange) or cell state #2 (magenta). Scale bar is 10μm.
RNA FISH image was provided by Kevin Janes.

1.2 Insights from single-cell gene-expression profiling of intra-tumor heterogeneity

Regulatory variations in cells arise as differences in the expression of mRNA transcripts and protein molecules in the cell. In this dissertation, we focus on transcriptomic measurements of total mRNA expression in cells to infer regulatory states. In the last decade, single-cell RNA-sequencing (scRNA-seq) methods have become highly accessible, resulting in an ever-increasing number of single-cell studies of intra-tumor heterogeneity (110–112). In this section, we will discuss some emergent themes from scRNA-seq studies of intra-tumor heterogeneity in human tumors (113).

1.2.1 Shared roles for tumor infiltrating lymphocytes across multiple cancers

Gene-expression profiling provides an unbiased method to assess both cell-type and cell-state heterogeneity in tumors. Creating catalogs of cell populations that express different lineage markers is a primary goal of many scRNA-seq studies (113,114). Due to the presence of pre-defined lineage markers, tumor infiltrating lymphocytes (TILs) are easily identified in scRNA-seq datasets of multiple tumor types. Advances in immunotherapy have also catalyzed measurements of the proportions and diversity of TILs in solid tumors. To focus on TILs, three out of seven scRNA-seq studies of breast tumors pre-selected CD45+ immune cells and omitted cancer cells prior to expression profiling (19,115–120). One of these studies noted that expression profiles of immune cells in tumors were similar to those found in normal tissues, but showed increased diversity upon interacting with cancer cells (116). This study also detected a unique tumor-associated immune cell-type, myeloid-derived suppressor cells, that are immunosuppressive and tumor promoting (116).

TIL composition has been cataloged in many other tumor types like melanoma, lung cancer, and liver cancer (78,120–123). Studies across multiple tumor types have identified

shared phenotypes of TILs that are related to tumor progression and treatment response (65,112). An immunosuppressive environment facilitated by higher proportions of T-cells in the "exhausted" state compared to an "activated state" has been associated with poor patient prognosis in multiple tumor types (78,120–123). scRNA-seq studies have demonstrated that TILs display similarities to normal immune cells, and that TILs across different patients and tumor types are also similar (65,112). These studies indicate that there may be convergent cell-states for TILs that arise across tumors and present additional strategies for immunotherapy.

1.2.2 Cancer cells display marked inter-tumor and intra-tumor heterogeneity

scRNA-seq studies of human tumors have uncovered marked expression differences between and within individual tumors. Compared to the expression patterns of TILs, malignant cancer cells across different patients show fewer shared transcriptional programs (124). When scRNA-seq data are clustered, TILs cluster by cell type with intermingling of different patients' tumors while cancer cells separate predominantly by patient (19,78,117). A striking example of this was observed in squamous tumors of the head and neck, where T cells from 10 patients co-clustered together, but cancer cells formed 10 separate clusters (125). Thus far, most scRNA-seq studies of human tumors have focused on advanced tumors, which show marked inter- and intra-tumor genetic variations (78,124,125). Patient-specific genetic variations uniquely amplify the repertoire of regulatory variations within cancer cells (see Section 1.1). Therefore, due to the compounded effect of genetic and regulatory variations between and within tumors, finding convergent cancer cell-states remains a challenge.

Deep explorations within single subtypes of cancer have provided insights into some rare but recurrent states of cancer cells. In multiple studies of melanoma tumors, a rare population of cancer cells that express high levels of the kinase *AXL* have been shown to be associated with resistance to targeted therapies (20,78). In contrast, melanoma cells that

express high levels of the *MITF* transcription factors display sensitivity to the same targeted therapies (126). These findings have important implications for patient stratification and combination therapeutic options to prevent disease recurrence in melanoma. However, the *AXL/MITF* cell-states have not been generalizable to other tumor types beside melanomas, indicating that tumor type might be an important determinate of cancer cell-states (65,124). A recent study evaluating chemoresistance in triple-negative breast cancer identified sporadic pre-treatment expression of cancer related genes like *MYC and COL1A1* that may cause intrinsic resistance (19). However, the expression patterns were highly patient specific and non-generalizable, suggesting that these approaches may need to be reevaluated for profiling epithelial tumors. There remains a need for focused studies of cancer cells within tumor subtypes to identify generalizable cancer cell states that could be diagnostically and therapeutically actionable.

1.2.3 Diversity and convergence of regulatory heterogeneities in tumors and metastasis: open questions

Regulatory state variations of individual cancer cells remain difficult to measure and interpret (36,70,78,116,120,127,128). Most studies of intra-tumor heterogeneity have focused on late-stage tumors that are advanced or have metastasized (78,124,125). However, the dynamics of cell-state changes of cancer cells in late stage tumors are complicated, and have not been generalizable (124). It is unclear when regulatory state variations first arise in cancer cells or if there are convergent trajectories that cancer cells take in early tumors. Studies of genetic heterogeneity have shown that early variations create differences in metastatic potential and chemoresistance (19,57). Similarly, regulatory heterogeneities in early stages are likely to provide insight towards understanding disease progression and eventually, patient outcomes (63). It remains necessary to study cancer cell regulatory states in early tumors, and systematically measure how different microenvironments modulate cancer cell phenotypes.

Regulatory variations between cells have been documented in cell culture (**Figure 1.2**), and in the setting of non-malignant cells as well, indicating that single-cell heterogeneity exists through cell-autonomous processes (20,76,129). However, it is not clear to what extent regulatory variations depend on the specific cell being examined, nor is it clear how regulatory states of the same cell would change in response to different microenvironments. The influence of genetic lesions, cell-type, and microenvironmental stimuli are often convolved when measuring cells in tissue. Separating the contributions of these different influences is critical to understanding the process of metastasis where cells have to adapt to new environments and heterotypic interactions.

Thus far in this chapter we introduced many influences that alter the regulatory states of tumor cells. Combined with stochastic variation and dynamic interactions with other cell types, there are seemingly infinite variations that could be measured amongst cells (75). This creates a significant challenge in interpreting the biological significance of any measurements of cancer cell heterogeneity (124). If certain regulatory states confer growth advantages, we would expect them to be selected for across multiple cells in multiple tumors. Therefore, understanding the biological implications of cancer cell variations requires integrating information over all scales of tumor heterogeneity. Identifying regulatory mechanisms that converge across patients would yield novel ways to stratify patients at the time of diagnosis and new strategies for combination therapies (130).

1.3 Methods for single-cell transcriptomics and shared challenges

Since the first single-cell transcriptomic measurements were reported in 2009, many iterations of scRNA-seq methods have emerged (110,131–134). Different methods optimize different aspects of the core steps of the scRNA-seq workflow: cell isolation, RNA extraction, reverse transcription of RNA to cDNA, and cDNA amplification to generate starting material for

next-generation sequencing. Broadly, scRNA-seq methods can be divided into plate-based and droplet-based methods that prioritize sensitivity and throughput, respectively (111,135). The next two subsections describe the leading methods in these two categories and their critical features. The last subsection discusses the overarching challenges that persist across all scRNA-seq methods.

1.3.1 Plate-based methods

In plate-based methods, single cells are deposited into individual wells of microplates (or chambers of microfluidic chips) for cell lysis and mRNA extraction (134,135). SMART-seq2 is the leading protocol of choice for plate-based methods and has been used in ~35% of studies investigating intra-tumor heterogeneity in human tumors (112,113,136). For this method, individual single cells are isolated through flow sorting and deposited into 96-well plates. Flow sorting can be combined with antibody-labelling (fluorescence activated cell sorting, FACS) to selectively profile specific populations of cells. Once individual cells are lysed, poly-adenylated (polyA) mRNA is reverse transcribed using an oligo-dT primer to obtain the first-strand of cDNA. SMART-seq2 is named for its “Switching mechanism at 5’ end of RNA template” (SMART) reverse transcription, a critical technological advance that couples full-length cDNA generation with the incorporation of global primer for PCR (132,136,137). The SMART method relies on the intrinsic properties of certain reverse transcriptase enzymes like the Moloney murine leukemia virus (MMLV) reverse transcriptase. The MMLV reverse transcriptase consistently adds a string of cytosines to the 3’ end of transcribed cDNA. When a complimentary template switching oligonucleotide (TSO) is added that has a string of guanines to base pair with the 3’ CCCs, the MMLV is able to switch templates, and complete the 5’ end of the mRNA and add a global PCR primer that can be incorporated into the TSO. This allows for second-strand synthesis to be initiated automatically without the need for additional tailing of the cDNA molecule, enabling

amplification of full-length cDNA molecules. This improved cDNA generation method results in improved, full-length mRNA detection by SMART-seq2, with an estimated ~40% coverage rate of the transcriptome (the highest end of the range for scRNA-seq methods) (137). Full-length mRNA capture has several downstream analytical advantages. The first advantage is more accurate read alignments, as full-length measurements reduce the dependency on the 3'-end of genes that are often poorly annotated and difficult to map (136). Further, reads that span the full-length of a transcript allow detection of novel isoforms and splice-variants that would be impossible from 3'-end counting alone. These features result in SMART-seq2 having amongst the highest gene detection sensitivity across scRNA-seq methods, with most studies reporting 4-9,000 genes detected per single-cell (132,134).

Several other plate-based methods diverge from SMART-seq2 in cell multiplexing and cDNA amplification. Massively Parallel Single Cell Sequencing or MARS-seq is one method that achieves increased throughput by multiplexing and sorting cells in to 384-well plates (138). MARS-seq foregoes TSOs to utilize a modified oligo-dT primer that additionally contains nucleotide sequences to serve as cellular and molecular barcodes (132,134,138). These barcodes are combinations of oligonucleotides that can be generated to uniquely label individual cells, allowing cells to be multiplexed at the earliest steps of the protocol. In addition to cellular barcodes, further combinations of oligonucleotides can be used to create Unique Molecular Identifiers (UMIs) which can be added to individual mRNA molecules during reverse transcription. MARS-seq also employs in vitro transcription (IVT) to amplify cDNA, as opposed to PCR based amplification in SMART-seq2. For IVT, the oligo-dT primer contains a promoter sequence for T7 RNA polymerase. After cDNA synthesis, the T7 enzyme transcribes multiple antisense RNA copies from the cDNA template. After multiple rounds of IVT, the RNA is once again reverse-transcribed to cDNA for sequencing (110,111). The combination of UMIs and IVT minimize the technical variation added by non-linear PCR amplification (134). After sequencing and alignment, gene expression is estimated by counting unique UMIs instead of total

sequenced reads to avoid counting PCR duplicates. While PCR based technical variation and counting noise is reduced in MARS-seq, this method only tends to detect 500-5000 genes per cell, demonstrating reduced sensitivity compared to SMART-seq2 (132,134).

1.3.2 Droplet based methods

In droplet-based methods, individual cells are encapsulated into liquid droplet emulsions for cell lysis and mRNA extraction. For these methods, two separate flows of liquids are created – one flow contains beads with lysis buffer and reverse transcription reagents and the second flow contains cells in a limiting single-cell dilution (111,132). These flows are combined and then emulsified by adding oil drops at a specific frequency to create nanoliter droplets that each contain one cell and one reagent bead. DROP-seq and inDROP were the pioneering methods developed for droplet-based scRNA-seq (139,140). Recently, 10X Genomics has developed a commercial device and kit for droplet-based scRNA-seq that is becoming increasingly popular (113,141). 10X Genomics' Chromium method uses a gel bead that contains oligonucleotides with other reaction components aimed at maximized multiplexing. The oligo-dT primer is extended to contain molecular barcodes to identify each cell as well as UMIs to identify individual mRNAs. The gel beads also contain TSOs containing sequencing adapters to be added with template switching reverse transcription. As individual cells and mRNAs are barcoded within the emulsion, all droplets can be broken together after mRNA capture and reverse transcribed and amplified in one reaction (139). These multiplexed reactions greatly reduce the reagent costs per cell and simplify the workflow, resulting in up to 100-fold increases in throughput compared to plate-based methods (111,135).

The high throughput of droplet-based methods allows for rapid cataloging of vast numbers of cells in a highly accessible, one-pot method. However, the efficiency of droplet-based methods remains low and only ~50% of input cells are captured, and within those cells

only ~13% of mRNAs are measured (139,141). To be cost effective, droplet-based scRNA-seq libraries are usually sequenced at lower depth, exacerbating measurement losses (135). This results in markedly low detection sensitivity, with reported averages of 500-3000 genes/cell (132,134).

1.3.3 Challenges and limitations for single-cell transcriptomics

Several optimizations to scRNA-seq workflows have greatly improved throughput and detection sensitivity of different methods. However, there remain a common set of challenges that are discussed below.

1.3.3.1 Challenges in isolating tumor cells in their native context

The first step of all scRNA-seq processes tends to be the most disruptive: isolating single cells from a multicellular tumor. It is nearly impossible to select a single cell within a solid tumor without disrupting its native context. Current experimental procedures involve both mechanical and enzymatic cell dissociation, followed by sorting single cells either using flow cytometry (which further involves cell labeling), or microfluidic devices (136,142) . These experimental steps have been documented to invoke stress and injury responses in cells that cause artefactual changes to gene expression (143). Further, once cells are dissociated, all information regarding their spatial position in the tumor and microenvironment is lost, providing no context for the eventual gene-expression differences measured. Therefore, it is necessary to combine single-cell transcriptomic measurements with cell isolation procedures that retain cells in their native context and minimize disruption prior to measurements (129,144,145).

1.3.3.2 Limited dynamic range of scRNA-seq measurements

scRNA-seq measurements are technically fraught due to the labile chemical structure of RNA and the picogram quantities obtained from single cells (146,147). These factors severely restrict the dynamic range of the transcriptome capture by single-cell measurements. It has been estimated that the majority of the transcriptome is at fewer than 50 mRNA copies in single-cells (148). Additionally, the conversion efficiency of RNA to complementary DNA during reverse transcription is estimated to be between 10-40%, resulting in subsampling of most transcripts (148,149). Further subsampling occurs during next-generation sequencing, as only a limited number of reads can be obtained per cell. Together this results in most scRNA-seq measurements only capturing the highest expressed genes in any single cell, limiting the inferences that can be made about regulatory variations (150).

1.3.3.3 Analytical challenges of scRNA-seq data

The major advantage of scRNA-seq is its high throughput, especially the droplet-based methods that enable simultaneous measurements from thousands of cells. The tradeoff is shallower read-depths which increases technical variance in measurement of lowly expressed genes. Several low-expressed genes appear to “dropout” in scRNA-seq data, meaning they are undetected in cells due to subsampling rather than true lack of expression (149,151). This results in large, high-dimensional datasets with a large number of zeros, creating left-censored distributions with difficult statistical properties (149,151). Several analytical advances to reduce dimensionality have aided in visualization and clustering of scRNA-seq datasets, of which uniform manifold approximation and projection (UMAP) is the latest to gain popularity (152). A non-linear dimensional reduction technique, UMAP is used extensively to cluster whole transcriptomes of single-cells to identify cell-type and cell-state subpopulations. This method is relatively computationally efficient even for very large datasets, and provides a rapid way to gain insights from high-dimensional datasets.

A greater challenge is identifying the specific genes that give rise to cell-state clusters. Analytical methods for differential expression in bulk RNA-seq data fail for scRNA-seq data because they do not account for technical variation arising from drop-outs and zero-inflated datasets (152–156). This is particularly problematic for low abundance genes for which statistical assumptions of lognormal distributions fail (149). In scRNA-seq data, the variance associated with a transcript has an inverse relationship with its overall abundance in the population, therefore, noise models have to incorporate estimates of global abundance of genes in the population of cells measured (154). One way to incorporate these technical aspects is to model a transcript's abundance as a mixture of drop-out and amplified components, as is done in SCDE, a scRNA-seq analysis method (154). SCDE models drop-outs as a Poisson process and the amplified component of a given transcript as a negative binomial process. Together, this enables the generation of a transcriptome-wide expectation model for all genes measured in a given set of samples (153,154). Differentially expressed genes can then be identified as those genes whose expression variances are not explained simply by technical variation.

Accurate noise models are especially critical for scRNA-seq data due to a lack of technical controls. Since every cell is measured uniquely, scRNA-seq measurements convolve both technical and biological variation. For analysis, the ability to measure one cell at a time is also a drawback and extracting biological significance of single-cell variations remains challenging.

1.4 Identifying regulatory heterogeneities through stochastic profiling

To address the limitations of current approaches for measuring intra-tumor heterogeneity, our lab has previously devised a "stochastic profiling" approach. Stochastic profiling combines *in situ* cell isolation by laser-capture microdissection with gene expression measurements (129,157). In this approach, small pools of 10 cells are sampled to mitigate the

technical losses at the single-cell input level. Repeated transcriptomic measurements of 10-cell pools are made to obtain expression distributions on a gene-by-gene basis. Technical controls are incorporated into the stochastic profiling by pooling a larger number of cells and splitting down into 10-cell equivalents after RNA extraction. Expression distributions of genes are then evaluated to identify heterogeneous transcripts that differ from a null distribution of lognormal biological variability (**Figure 1.3**). Filtered transcripts that display increased biological variability are then clustered to identify co-fluctuating genes that comprise regulated transcriptional programs (129).

Despite 10-cell pooling, stochastic profiling can determine single-cell variations in gene expression through fluctuation analyses. Further, 10-cell measurements can be mathematically deconvolved to estimate the underlying single-cell expression frequencies for heterogeneously expressed genes (158). Experimental validation of stochastic profiling has confirmed frequency predictions for genes that are heterogeneously expressed in as few as 2-5% of the population (157,158). Stochastic profiling has been applied to measure *in vitro* 3D cultures of breast epithelial cells to identify multiple cancer-related cell states (76,129). However, the original experimental pipeline for stochastic profiling is not compatible with next-generation sequencing methods. In this dissertation we will discuss updates to stochastic profiling to obtain sequencing-based transcriptomic measurements from groups of 10-cells isolated from tumors and tissues.

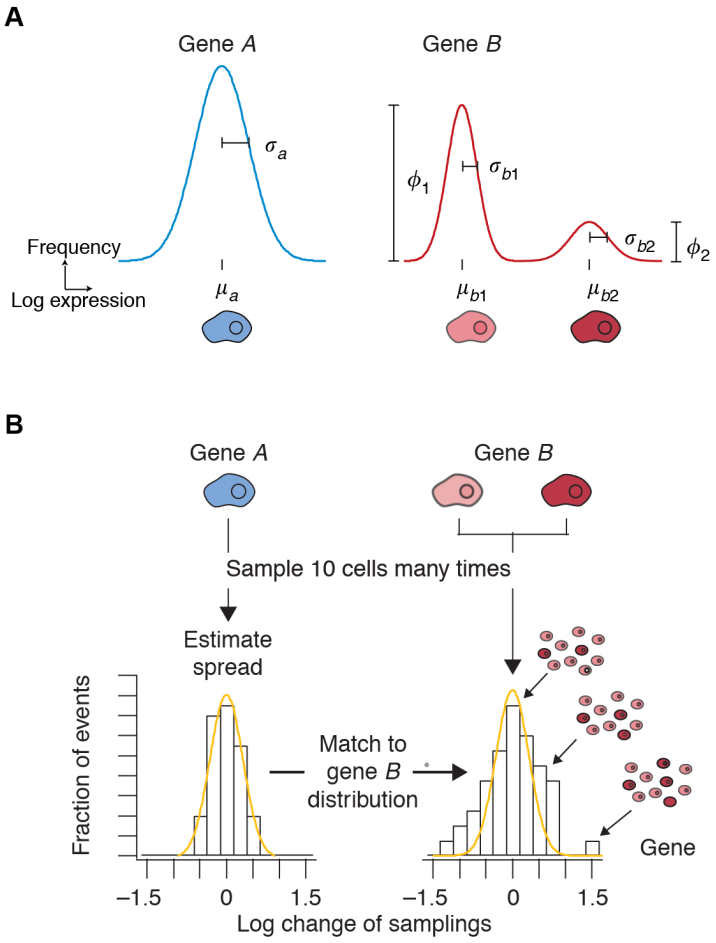

Figure 1.3 Stochastic profiling identifies heterogeneously expressed genes in 10-cell pools
(A) Schematic of lognormal variation (Gene A) and regulatory heterogeneity (Gene B) that underlie stochastic profiling.
(B) Repeated 10-cell samplings are used to compare expression distributions arising from lognormal noisy measurements (Gene A) to expression distributions with high variance due to biological heterogeneity (Gene B).
Reprinted with permission from (129).

1.5 Cancer types studied in this dissertation

In this dissertation, we focused on epithelial tumors to identify recurrent cell states arising across multiple tumors (Chapter 3) or multiple microenvironmental settings (Chapter 4). We investigate these questions in two different tumors: human luminal breast cancers and a murine model of small cell lung cancer. Luminal breast cancers and small cell lung cancers differ in many aspects of tumor heterogeneity, as summarized below.

1.5.1 Luminal breast carcinoma

Breast carcinomas arise from the glandular epithelium in the breast. The luminal subtypes comprise 70% of all breast tumors and are diagnosed by their expression of hormone receptors for estrogen and progesterone (14,159). Due to screening mammography, most luminal tumors are detected early and surgically resected. Patients are additionally treated with adjuvant anti-estrogen therapy, with or without chemotherapy. However, 30-40% of luminal breast cancers are inherently resistant to anti-estrogen treatments through mechanisms that are still being investigated (160,161).

Many facets of the breast tumor microenvironment have been associated with altering tumor growth and treatment response. Studies of the breast tumor microenvironment have been pivotal in identifying tumor supportive roles for CAFs and macrophages (section 1.1.2). However, studies of cell-type heterogeneity in luminal cancers have not yielded new therapies. Due to an overall lack of TILs compared to other tumor types, luminal breast tumors are typically considered “immune cold” and poor candidates for immunotherapy, although new studies with scRNA-seq are challenging this view (116,162,163).

Extensive molecular characterizations of luminal tumors have identified inter-tumor differences in genetic variation and gene expression. Mutations in the kinase *PIK3CA* are observed in 29-45% of tumors, followed by *TP53* mutations in 10-29% tumors, and lower

frequencies of mutations in several other genes (5). Mutations in *PIK3CA* also appear in luminal tumors after treatment resistance, which has motivated many trials to test combinations of anti-estrogen and anti-PI3K therapies (161,164). While beneficial for metastatic disease, PI3K therapies have shown lack of efficacy in early tumors, possibly due to inter-tumor differences in mutations (164). Stratification based on gene-expression profiling has been clinically actionable in luminal tumors, subdividing them into Luminal A (low proliferation) and Luminal B (high proliferation) (17,18). This subdivision has been useful in stratifying patients for chemotherapy, and gene-expression signatures for predicting chemotherapy response are used clinically. Targeting proliferation in luminal cancers has translated into the direct inhibition of cell cycle activators CDK4/6, which has significantly improved survival in patients with metastatic disease (166).

However, despite early diagnosis, genomic characterizations, and targeted therapies, greater than 30% of patients with luminal tumors suffer disease recurrence (16,17). Luminal tumors show marked inter-tumor heterogeneity in both genomic mutations and gene-expression, but there are few studies exploring intra-tumor heterogeneity in these tumors (11, 117, 118). Single-cell heterogeneities between cancer cells in early stages of luminal breast tumors may explain treatment resistance and variable rates of disease recurrence. To identify regulatory variations in luminal cancer cells, we leveraged their early detection to measure intra-tumor heterogeneity in patient biopsies at the time of diagnosis (Chapter 3).

1.5.2 Small cell lung carcinoma

Small cell lung carcinoma (SCLC) is a deadly form of lung cancer that arises from pulmonary neuroendocrine cells (PNECs) that line the respiratory airway (167,168). PNECs arise from epithelial lineages but retain stem-like capacity to differentiate into different cell-types in the lung in response to injury and inflammation (169,170). In contrast to luminal breast

tumors, SCLCs are usually detected late and have often metastasized by the time of diagnosis (6). Patients are treated with chemotherapy, but the vast majority acquire resistance to therapy over time resulting in a very low overall survival rate of ~5% for these tumors (6,171).

High rates of acquired resistance to chemotherapy have prompted studies into heterogeneity of SCLC, uncovering that SCLC cancer cells in the same tumor can display heterogeneous expression of neuroendocrine markers (172,173). SCLC cells display lineage plasticity, retaining the ability of PNECs to differentiate into multiple cell-types (173). Re-programming of PNECs is driven by interactions with other cell types during injury and inflammatory processes, indicating an important role for the microenvironment in regulating SCLC cell states (170,174). Despite a lack of studies observing TILs in SCLC, immune checkpoint blockade has been attempted in SCLC because of their exceedingly poor prognosis. Immunotherapy trials thus far have had very limited success, indicating the need for better tools to stratify SCLC patients for these treatments (175).

SCLC tumors have stereotyped genetic lesions, and greater than 80% of tumors have loss of function of the genes *TP53* and *RB1 (6)*. Several other oncogenes driving growth pathways are mutated at ~20% frequency in SCLC patients (171). To target common proliferative effects of these mutations in SCLC, trials have tested inhibitors for Aurora kinases (which coordinate cell division), with promising early results (171,176). Recently, inter-tumor heterogeneity in gene expression has identified four major subtypes of SCLC. These are defined by transcriptional programs regulated by specific transcription factors: *ASLC1* (SCLC-A), *NEUROD1* (SCLC-N), *POU2F3* (SCLC-P), and *YAP1* (SCLC-Y) (175). However, single tumors display expression of more than one of these classifying transcription factors and their prognostic value remain unclear (175,177). Given their inherent plasticity, it also remains unknown to what extent individual SCLC cells follow these transcriptional programs within single tumors (109,173,178).

The phenotypic plasticity of SCLC in response to heterotypic interactions motivated a systematic analysis of cancer cell regulatory heterogeneity. SCLC tumors are frequently unresectable, making comparative measurements between tumors and metastases in human samples difficult. Therefore, to examine microenvironment influences on tumor heterogeneity we turned to a genetically engineered murine model (GEMM) of SCLC (Chapter 4) (179). Using a syngeneic model allowed us to retain the interactions between cancer cells and immune cells *in vivo* and study the influence of these interactions on cancer cell regulatory heterogeneity.

1.6 Overview of this dissertation

In this dissertation we developed novel approaches to measure cancer cell regulatory heterogeneity and identify recurrent variations in breast and lung carcinoma cells.

In this chapter, we provided an introduction to how regulatory variations between cancer cells arise and how they influence cellular phenotypes and tumor progression. In Chapter 2, we present a method for *in situ* transcriptomic measurements which addresses the challenges in measuring tumor cell heterogeneity outlined in preceding sections. We present an experimental pipeline to obtain transcriptomic measurements from spatially resolved 10-cell pools, microdissected *in situ* (10cRNA-seq). 10cRNA-seq is readily applicable to tissues and tumors obtained from both clinical samples and murine models.

In Chapter 3, we characterize the gene-regulatory heterogeneities of cancer cells in early stage breast cancer biopsies. Combining 10cRNA-seq measurements with stochastic profiling analysis, we uncover thousands of heterogeneously expressed genes in individual cases of luminal breast cancer. We identified a recurrent set of genes shared by multiple tumors that are known drivers for other cancer types, but not identified as driver genes in breast tumors.

In Chapter 4, we study the microenvironmental modulation of cancer cell heterogeneity in a murine model of small cell lung cancer. We detail regulatory variations intrinsic to small cell

lung cancer cells, and how these variations dramatically expand in the context of liver colonization and heterotypic cell interactions. Upon liver colonization *in vivo*, we observe the ability of cancer cells to de-differentiate into cell states that express markers of multiple lineages.

Finally, in Chapter 5, we bring together the findings from human and murine models of epithelial tumors to discuss the implications of shared regulatory variations and the future directions of research that emerge from them. We also discuss future applications of the tools and analytical methods developed in this dissertation.

The work presented in this dissertation identifies shared themes in cancer cell heterogeneity across epithelial tumors and discerns the influence of complex microenvironments on cancer cell heterogeneities. We developed experimental and analytical approaches to characterize the diversity in cancer cell states and identify points of convergence. More broadly, this dissertation expands our understanding of cancer cell heterogeneity, providing insight into tumor progression, and ultimately, differential outcomes for patients diagnosed with the same disease.

2 In situ 10-cell RNA Sequencing in Tissue and Tumor Biopsy Samples

2.1 Foreword

Single-cell RNA-seq methods described in Chapter 1 classify new and existing cell types very effectively, but alternative approaches are needed to quantify the individual regulatory states of cells in their native tissue context. In this Chapter, we combined the tissue preservation and single-cell resolution of laser capture with an improved preamplification procedure enabling RNA sequencing of 10 microdissected cells. This in situ 10-cell RNA sequencing (10cRNA-seq) can exploit fluorescent reporters of cell type in genetically engineered mice and is compatible with freshly cryoembedded clinical biopsies from patients. Through recombinant RNA spike-ins, we estimate dropout-free technical reliability as low as ~250 copies and a 50% detection sensitivity of ~45 copies per 10-cell reaction. By using small pools of microdissected cells, 10cRNA-seq improves technical per-cell reliability and sensitivity beyond existing approaches for single-cell RNA sequencing (scRNA-seq). Detection of low-abundance transcripts by 10cRNA-seq is comparable to random 10-cell groups of scRNA-seq data, suggesting no loss of gene recovery when cells are isolated in situ. Combined with existing approaches to deconvolve small pools of cells, 10cRNA-seq offers a reliable, unbiased, and sensitive way to measure cell-state heterogeneity in tissues and tumors (Chapters 3 and 4).

This work was published in *Scientific Reports* in March 2019 with me as co-first author (180). I have adapted the text and figures for this chapter in accordance with Springer Nature publishing policies.

2.2 Introduction

Tumors are complex mixtures of cells that are heterogeneous in their genetics, lineage, and microenvironment (21,181). Whole-tumor profiles of genes and transcript abundances yield inter-tumor differences that are clinically important for patient prognosis, but these cellular profiles are population averages (11,18,182,183). The tumor microenvironment contains several different cell types that vary among cases (25,47,184–187). At the single-cell level, cancer cells are heterogeneous and genetic subclones evolve as the disease progresses (58,59). Tumor cells also display non-genetic heterogeneity and can switch between regulatory states in a reversible and context-dependent manner (20,75,104). Together, these variations dictate phenotypic differences such as proliferative index, metastatic potential, and response to therapy (20,54,76,78,188,189).

Assessing intra-tumor heterogeneity of gene regulation requires precise transcriptomic measurements of a very small number of cells isolated from within the tumor context. The current methods for single-cell RNA sequencing (scRNA-seq) are powerful in their ability to profile thousands of individual cells and identify differences in genotype or lineage in a mixed population. However, the first step of most large-scale scRNA-seq methods is some form of tissue dissociation and single-cell isolation, which can alter transcriptional profiles and confound downstream analyses (143,190). Approaches such as laser-capture microdissection (LCM) can obtain samples for RNA-seq (144,191–193), but they usually require so many cells for reliable measurement that single-cell variation is obscured (**Figure 2.1**). Dissociation-based scRNA-seq methods also struggle with technical variability, including "dropout" of medium-to-low abundance transcripts that yield zero aligned reads (146,151,155,194). The 3–40% conversion efficiency (140,146,149,194,195) of RNA to amplifiable cDNA is problematic given estimates that 90% of the transcriptome is expressed at 50 copies or fewer per cell (148). While valid for the most consistently expressed genes and markers within a sample, scRNA-seq data miss a

large proportion of the transcriptome (148,196). Measuring single-cell expression profiles in situ is even more challenging because of losses incurred during biomolecule extraction as well as non-mRNA contaminants, which can be considerable in stroma-rich specimens. Collectively, these hurdles make it difficult to measure tumor- cell regulatory heterogeneities reliably and evaluate their functional consequences.

Multiple studies have reported a pronounced improvement in gene detection and technical reproducibility when using 10–30 cells of starting material rather than one cell (129,144,149,157,197–199). The increased cellular RNA offsets losses incurred during reverse transcription, enabling more reliable downstream amplification. The gains are irrespective of amplification strategy and detection platform, and they are more dramatic than when increasing the starting material another tenfold to 100 cells. Previously, we combined the technical advantages of 10-cell pooling with the in-situ fidelity of LCM to devise a random-sampling method called “stochastic profiling” (129,157).The method identifies single-cell regulatory heterogeneities by analyzing the statistical fluctuations of transcriptomes measured repeatedly as 10- cell pools microdissected from a cell lineage (70,129). Pooling increases gene detection and technical reproducibility; repeated sampling is used to extract the single-cell information that is retained in pools of 15 cells or smaller (**Figure 2.1**). Genes with bimodal regulatory states create skewed deviations from a null model of biological and technical noise, which parameterize the underlying population-level distribution more accurately than single-cell measurements (158,198,200). By applying stochastic profiling to spatially organized breast-epithelial spheroids and gene panels measured by quantitative PCR or microarray, we uncovered multiple regulatory states relevant to 3D organization and stress responses (76,201,202). However, this early work did not stringently evaluate the importance of sample integrity for primary tissues from animals or patients, nor did it involve probe-free measures of 10-cell data like RNA sequencing.

Here, we report improvements in sample handling, amplification, and detection that enable RNA sequencing of 10-cell pools isolated from tissue and tumor biopsies by LCM and its extensions. We find that cryoembedding of freshly isolated tissue pieces is crucial to preserve the localization of genetically encoded fluorophores in engineered mice used for fluorescence-guided LCM. By incorporating ERCC spike-ins at non-disruptive input amounts in the amplification, we calibrate sensitivity and provide a standard reference to compare with other scRNA-seq methods (150). Sample tagging and fragmentation (tagmentation) is accomplished by Tn5 transposase, which is compatible with the revised procedure as well as with past 10-cell amplifications (203). We sequence archival samples that had previously been measured by BeadChip microarray to provide a side-by-side comparison of transcriptomic platforms with limiting material (129,204). Applying 10-cell RNA sequencing (10cRNA-seq) to various mouse and human cell types isolated by LCM, we obtain substantially better exonic alignments, and increases in gene coverage are consistent with the single-cell sensitivity of prevailing scRNA-seq methods. The realization of 10cRNA-seq by LCM creates new opportunities for stochastic profiling and other unmixing approaches to deconvolve single-cell regulatory states in situ (158,198).

2.3 Results

Methods for profiling small quantities of cellular RNA have evolved considerably over the past decade, but they all involve the same fundamental steps: 1) cell isolation, 2) RNA extraction, 3) reverse transcription, 4) preamplification, and 5) detection (205). The original protocol for in situ 10-cell profiling combines LCM for cell isolation followed by proteinase K digestion for RNA extraction (157). The extracted material undergoes an abbreviated high-temperature reverse transcription with oligo$(dT)_{24}$, and cDNA is carefully preamplified by poly(A)

PCR that generates sufficient 3’ ends (~500 bp in size) for microarray labeling and hybridization (**Figure 2.2**) (157,206).

Unsurprisingly, the earliest steps in the procedure are the most critical for achieving the maximum amount of amplifiable starting material. To avoid losses, steps 1–4 (cell isolation through preamplification) are normally performed without intermediate purification. Therefore, buffers and reagents must be carefully tested and titrated to be mutually compatible throughout the “one-pot” protocol. Since description of the procedure, multiple commercial providers merged or were acquired, leading to the discontinuation of multiple RNAse inhibitors, the Taq polymerase, and the BeadChip microarrays. The collective disruptions in sourcing prompted a modernization of 10-cell profiling toward RNA-seq of primary material at a biopsy scale, including how tissue–tumor samples were handled before the start of the procedure (**Figure 2.2**).

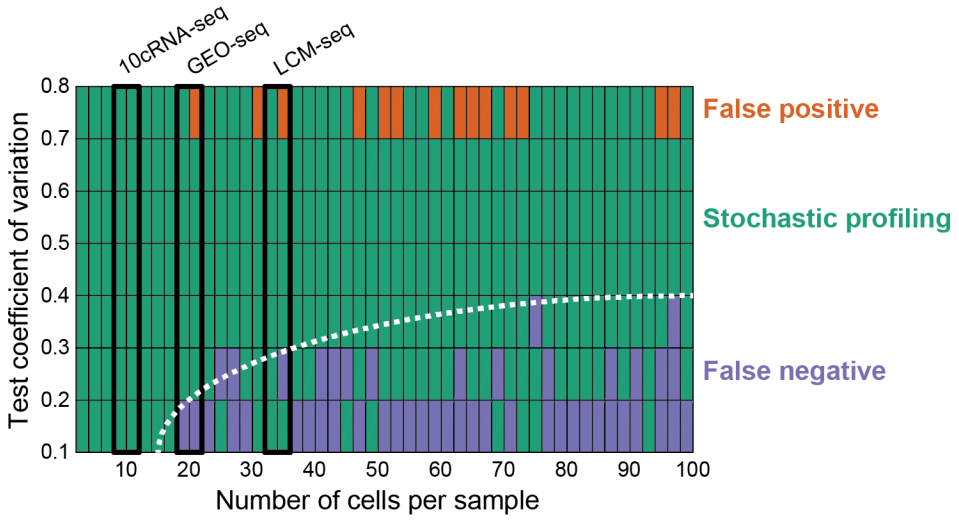

Figure 2.1 Population averaging obscures single-cell regulatory heterogeneities in pools of more than ~15 cells.
Monte Carlo simulations (157) of stochastic-profiling experiments are shown for 25 random samples, an expression fraction of 50%, a reference coefficient of variation of 0.3, and a fold difference in regulatory states of 5. False positives (orange) arise when a one-state gene with high variance relative to the reference distribution is incorrectly scored as having two regulatory states. False negatives (blue) arise when a two-state gene with low variance relative to the reference distribution is incorrectly as scored as having one regulatory state. Effective stochastic profiling (green) occurs when two-state genes are correctly scored as heterogeneously regulated. Cell input requirements for 10cRNA-seq are shown compared to applications of GEO-seq (191) and LCM-seq (193). Note the increases in false negatives for larger test variances observed with larger number of cells per sample.

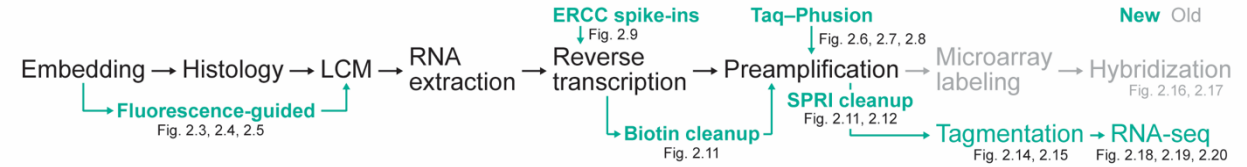

Figure 2.2 A revised transcriptomic pipeline for in situ 10-cell RNA sequencing.
Substantive changes are indicated in green and gray.

2.3.1 Protein localization for LCM requires fresh cryoembedding

To minimize extra handling steps that could degrade RNA, in situ profiling of clinical samples is ordinarily performed with rapid histological stains (**Figure 2.2**) (129,205,207,208). LCM can also be guided by fluorescence in place of histology when using cells or animals engineered to encode genetic labels (209,210). However, new challenges arise when seeking to preserve localization and brightness of encoded fluorophores during single-cell isolation and RNA extraction. Compared to polysome-bound mRNAs, fluorescent proteins diffuse much more readily, and chromophores may be damaged by the fixation and dehydration steps needed to preserve RNA integrity. Fluorescent-protein structure is preserved by chemical fixatives, but covalent crosslinking of biomolecules is unsuitable for extracting RNA from tissue. Fluorescence-guided profiling therefore entails a competing set of tradeoffs that must be balanced for optimal performance.

We reasoned that the greatest flexibility would be afforded by reporter mice expressing tandem-dimer Tomato (tdT)—a bright, high molecular-weight derivative of DsRed (211). Key handling parameters were evaluated using Cspg4-CreER;Trp53$^{F/F}$;Nf1$^{F/F}$;Rosa26-LSL-tdT mice, a model of malignant glioma (212). In these animals, administration of tamoxifen elicits sparse labeling of oligodendrocyte precursor cells (OPCs) in the brain, enabling fluorescence retention to be assessed in single cells. Extensive optimization of cryosectioning and wicking conditions was required to preclude fluorophore diffusion while ensuring reliable LCM pickup (see Methods). We found that an accelerated 70-95-100% ethanol series maintained tdT fluorescence and localization of labeled cells through xylene clearing and dehydration (**Figure 2.3A**). Separately, using freshly embedded tissue from a “mosaic analysis of double markers” (MADM) animal that labels various brain lineages with EGFP, tdT, or both, we confirmed that EGFP fluorescence was also acceptably retained with the 70-95-100% ethanol series (**Figure 2.4**) (213,214). Although EGFP diffusion was noticeably greater compared to tdT owing to its

smaller size (~28 kDa vs. ~54 kDa), we could nonetheless reliably identify the cell bodies of single EGFP-positive cells for LCM. Surprisingly, we found that fresh-tissue embedding was critically important for preserving single-cell localization and brightness. Snap- freezing before cryoembedding caused considerable loss and delocalization of tdT fluorescence, even when prefrozen material was rapidly embedded in dry ice-isopentane (–40˚C) (**Figure 2.3B,C**). Brightfield images of these cryosections also showed considerable tissue damage compared to freshly embedded material (**Figure 2.5**). For mechanically challenging tissues in which embedding support is important for cryosectioning, we conclude that fresh-tissue embedding is essential for maximum biomolecular retention and integrity.

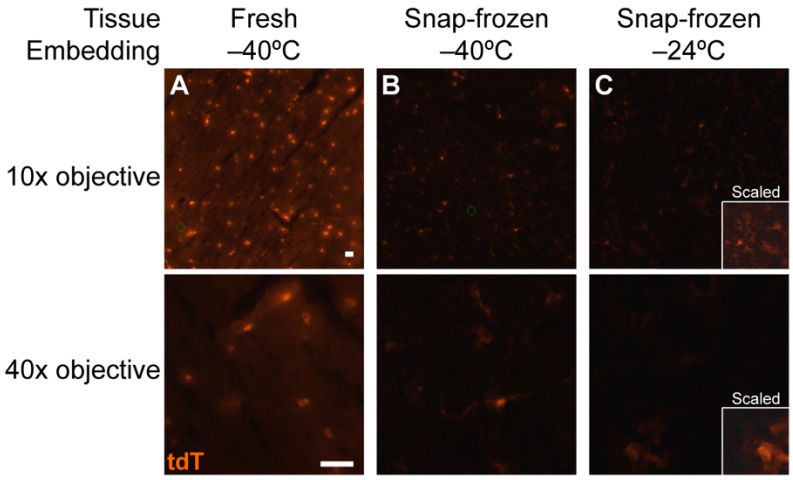

Figure 2.3 Fresh cryoembedding preserves tandem-dimer Tomato (tdT) fluorescence and localization better than snap-frozen alternatives.
Brain samples from *Cspg4-CreER;Trp53$^{F/F}$;Nf1$^{F/F}$;Rosa26-LSL-tdT* animals were either
(A) Freshly cryoembedded in Neg-50 medium with dry ice-isopentane (–40ºC)
(B) Snap-frozen in dry ice-isopentane and then cryoembedded
(C) Snap-frozen and slowly cryoembedded in a cryostat (–24ºC).
Low- and high-magnification images were captured with the factory-installed color camera on the Arcturus XT LCM instrument. Images were exposure matched and are displayed with a gamma compression of 0.67. Insets have been rescaled to emphasize tdT diffusion away from the cell body. Scale bar is 25 μm. Brightfield images from the same sections are shown in **Figure 2.5**.

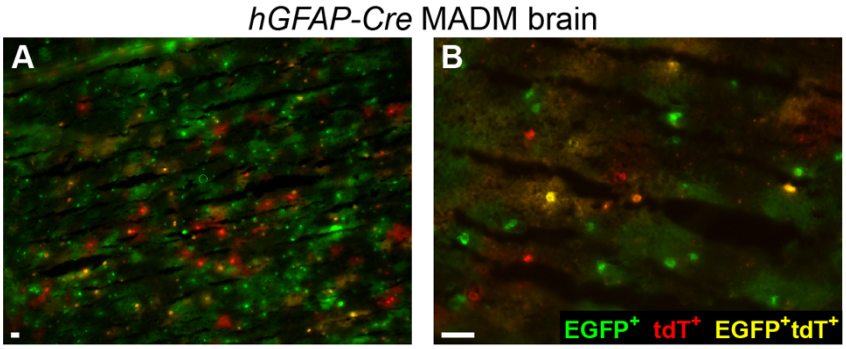

Figure 2.4 Fresh cryoembedding and 70-95-100% ethanol dehydration retains sufficient EGFP fluorescence and localization to identify single cells alongside tdT.
Tissue preparation was performed with a mosaic analysis of double markers (MADM) animal expressing Cre under control of the hGFAP promoter to label multiple brain lineages with EGFP (green), tdT (red), or both (yellow).
(A) Low- and **(B)** high-magnification images were captured with the factory-installed color camera on the Arcturus XT LCM instrument. Red and green spectral channels were separated, false colored, and merged to generate final images. Scale bar is 25 μm.

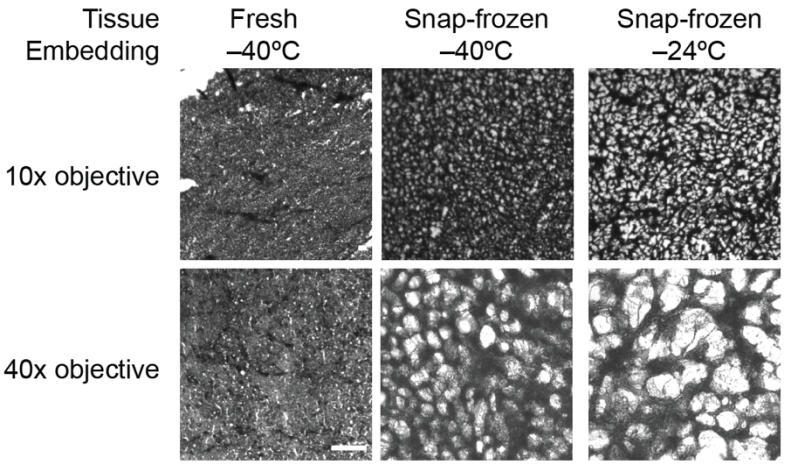

Figure 2.5 Fresh cryoembedding preserves tissue integrity better than snap-frozen alternatives.
Brain samples from *Cspg4-CreER;Trp53$^{F/F}$;Nf1$^{F/F}$;Rosa26-LSL-tdT* animals were either
(A) Freshly cryoembedded in Neg-50 medium with dry ice-isopentane (–40ºC)
(**B)** Snap-frozen in dry ice-isopentane and then cryoembedded, or
(C) Snap-frozen and slowly cryoembedded in a cryostat (–24ºC).
Low- and high-magnification images were captured with the factory-installed color camera on the Arcturus XT LCM instrument and converted to grayscale. Scale bar is 25 μm. Images of tdT fluorescence from the same sections are shown in **Figure 2.3**.

2.3.2 Improving poly(A) preamplification for modern RNA-seq

Previously, in situ 10-cell profiling was optimized for quantification by BeadChip microarray, but microarrays have been supplanted by RNA-seq for unbiased measures of the transcriptome (**Figure 2.2**) (215). An advantage of RNA-seq is that nucleic acids are detected regardless of origin, enabling use of exogenous RNA standards to calibrate sensitivity and quantitative accuracy when spiked into a biological sample (216–218). The versatility of RNA-seq is also a caveat, because all nucleic acids in a sample will be sequenced, including unwanted preamplification byproducts and contaminating DNA from mitochondria or the nucleus (219–221). In the original scRNA-seq report that used a variant of poly(A) PCR, only 37 ± 9% of sequenced reads aligned to RefSeq transcripts, and exonic alignment rates below 50% remain common (131,134). Therefore, we focused improvements to poly(A) preamplification towards ensuring that most sequencing reads aligned to the 3' ends of cellular mRNAs.

In poly(A) PCR, cDNA is 3’ adenylated and then preamplified with a universal T_{24}-containing primer called AL1 (206). We previously found that the amount of AL1 strongly influenced overall sensitivity of gene detection, with improvements noted at concentrations as high as 25 μM (157). Excess AL1 also drives nonspecific amplification of low molecular-weight primer concatemers, which do not influence gene measurements by quantitative PCR or microarray but create overwhelming contamination for RNA-seq (222). To improve poly(A) PCR, we screened a range of commercial Taq and proofreading polymerases along with empirical blends of those that maximized the intended ~500 bp cDNA products relative to nonspecific concatemer. We obtained a better-than-additive preamplification by combining Taq and Phusion polymerases (see Methods). An equal mixture of the two enzymes dramatically increased the yield of ~500 bp preamplification products relative to nonspecific concatemer (**Figure 2.6A**, lower). The empirical blend also significantly improved the preamplification of both high-abundance (GAPDH) and low-abundance (PARN) targets as measured by quantitative PCR

(**Figure 2.6A**, upper). The two-enzyme blend further enabled a 10-fold decrease in AL1 primer concentration without detectable loss in preamplification efficiency (**Figure 2.6B**). The Taq-Phusion combination was superior for a primary breast-cancer biopsy (**Figure 2.6**) as well as two murine tissue sources: a murine small-cell lung cancer line derived from $Trp53^{\Delta/\Delta}Rb^{\Delta/\Delta}$ lung epithelium and tdT- labeled OPCs (**Figure 2.7** and **Figure 2.8**), illustrating its generality (223). The enzyme modification created a viable starting point for combining poly(A) PCR preamplification with RNA-seq.

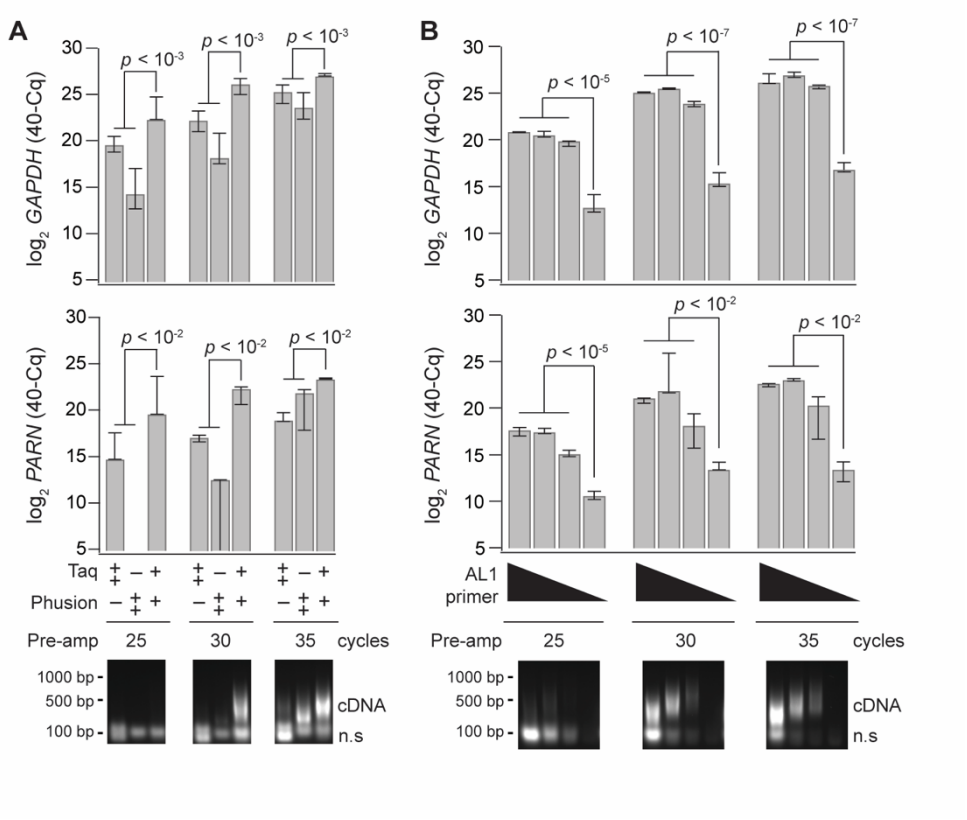

Figure 2.6 A blend of Taq–Phusion polymerases improves selective poly(A) amplification of cDNA and reduces AL1 primer requirements.

Cells were obtained by LCM from a human breast biopsy and split into 10-cell equivalent amplification replicates.

(A) Poly(A) PCR was performed with 15 µg of AL1 primer with Taq alone (10 units), Phusion alone (4 units) or Taq/Phusion combination (3.75 units/1.5 units).

(B) Poly(A) PCR was performed with either 25, 5, 2.5 or 0.5 µg of AL1 primer and the Taq–Phusion blend from (A). Above—Relative abundance for the indicated genes and preamplification conditions was measured by quantitative PCR (qPCR). Data are shown as the median inverse quantification cycle (40–Cq) ± range from n = 3 amplification replicates and were analyzed by two-way (A) or one-way (B) ANOVA with replication. Below—Preamplifications were analyzed by agarose gel electrophoresis to separate poly(A)-amplified cDNA from nonspecific, low molecular-weight concatemer (n.s.). Qualitatively similar results were obtained separately three times. Lanes were cropped by poly(A) PCR cycles for display but were electrophoresed on the same agarose gel and processed identically.

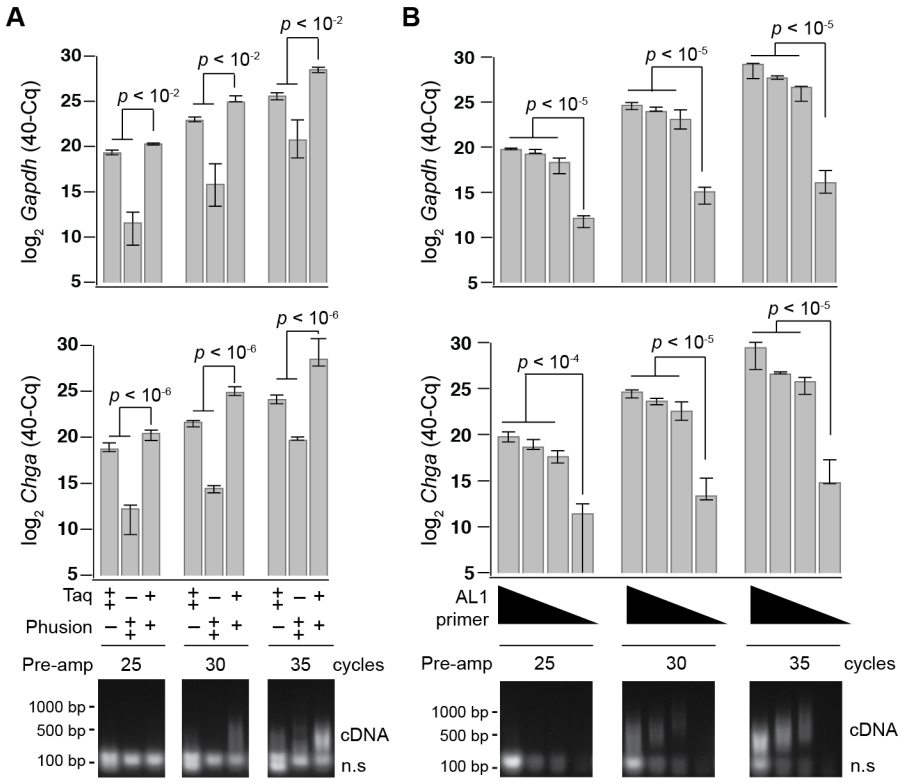

Figure 2.7 Improvements with the Taq–Phusion polymerases blend generalize to murine small-cell lung cancer cells.

Cells were obtained by LCM and split into 10-cell equivalent amplification replicates.

(A) Poly(A) PCR was performed with 25 µg of AL1 primer with Taq alone (10 units), Phusion alone (4 units) or Taq/Phusion combination (3.75 units/1.5 units).

(B) Poly(A) PCR was performed with either 25, 5, 2.5 or 0.5 µg of AL1 primer and the Taq–Phusion blend from (A). Above—Relative abundance for the indicated genes and preamplification conditions was measured by quantitative PCR (qPCR). Data are shown as the median inverse quantification cycle (40–Cq) ± range from n = 3 amplification replicates and were analyzed by two-way (A) or one-way (B) ANOVA with replication. Below—Preamplifications were analyzed by agarose gel electrophoresis to separate poly(A)-amplified cDNA from nonspecific, low molecular-weight concatemer (n.s.). Lanes were cropped by poly(A) PCR cycles for display but were electrophoresed on the same agarose gel and processed identically.

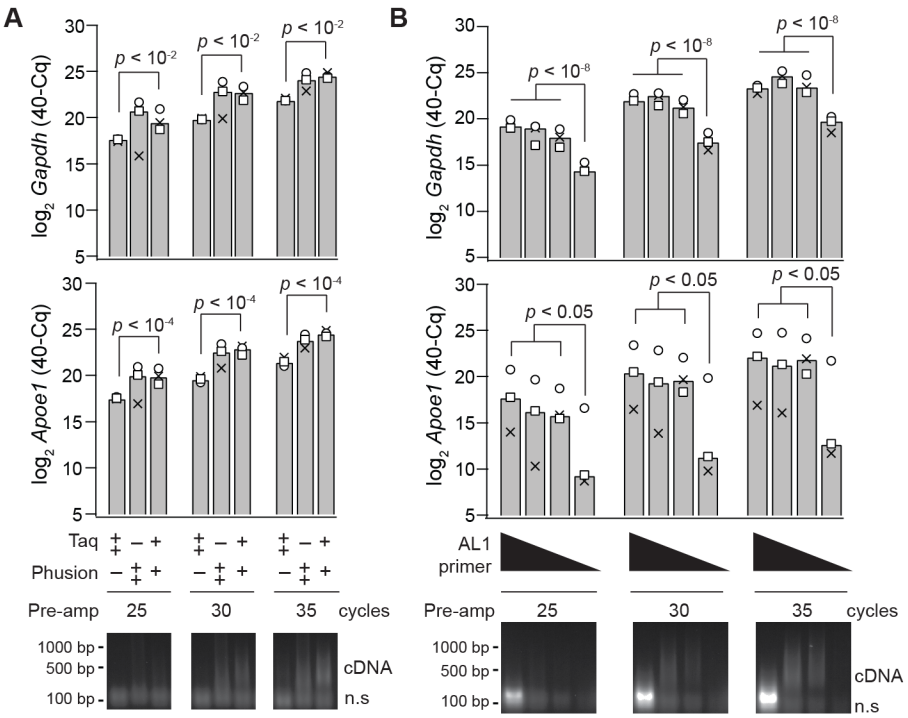

Figure 2.8 Improvements with the Taq–Phusion polymerases blend generalize to murine tdT-labeled oligodendrocyte precursor cells.

Cells were obtained by fluorescence-guided LCM and split into 10-cell equivalent amplification replicates.

(A) Poly(A) PCR was performed with 25 µg of AL1 primer with Taq alone (10 units), Phusion alone (4 units) or Taq/Phusion combination (3.75 units/1.5 units).

(B) Poly(A) PCR was performed with either 25, 5, 2.5 or 1 µg of AL1 primer and the Taq–Phusion blend from (A). Above—Relative abundance for the indicated genes and preamplification conditions was measured by quantitative PCR (qPCR). Data are shown as the median inverse quantification cycle (40–Cq) ± range from n = 3 replicates collected over 3 separate LCM acquisitions (markers) and were analyzed by two-way ANOVA with replication. Below—Preamplifications were analyzed by agarose gel electrophoresis to separate poly(A)-amplified cDNA from nonspecific, low molecular-weight concatemer (n.s.). Lanes were cropped by poly(A) PCR cycles for display but were electrophoresed on the same agarose gel and processed identically.

Sensitivity, accuracy, and precision of the updated poly(A) PCR approach were assessed using recombinant RNA spike-ins as internal positive controls (218). A dilution of ERCC spike-ins was defined that did not detectably perturb the measured abundance of endogenous transcripts in RNA equivalents from 10 microdissected cells (**Figure 2.9A**). After poly(A) PCR of the spike-in dilution plus 100pg RNA (~10 cells), we measured the relative abundance of individual spike-ins, using quantitative PCR (qPCR) to eliminate RNA-seq read depth as a complicating factor. Purified qPCR end products served as an absolute reference of each spike-in for cross-comparison (see Methods). We observed good linearity across 22 spike-ins spanning an abundance of ~10^4 (**Figure 2.9B**). Deviations, technical noise, and dropouts all increased considerably for spike-ins below ~250 copies per reaction, consistent with previous reports (146). This collective measurement uncertainty restricts interpretation of single-cell data to highly expressed transcripts, but 10-cell pooling reduces the threshold to ~25 copies on average per cell. With poly(A) PCR, we did not observe qualitative dropout in more than 50% of technical replicates for spike-ins as dilute as four copies per reaction (ERCC85; **Figure 2.9B**), indicating good sensitivity. RNA spike-ins do not mimic the characteristics of endogenous transcripts extracted from cells, but they can provide a common reference to benchmark preamplification methods for RNA-seq (150). These experiments indicated that the improved poly(A) preamplification was sufficiently reliable for unbiased profiling of 10-cell transcriptomes.

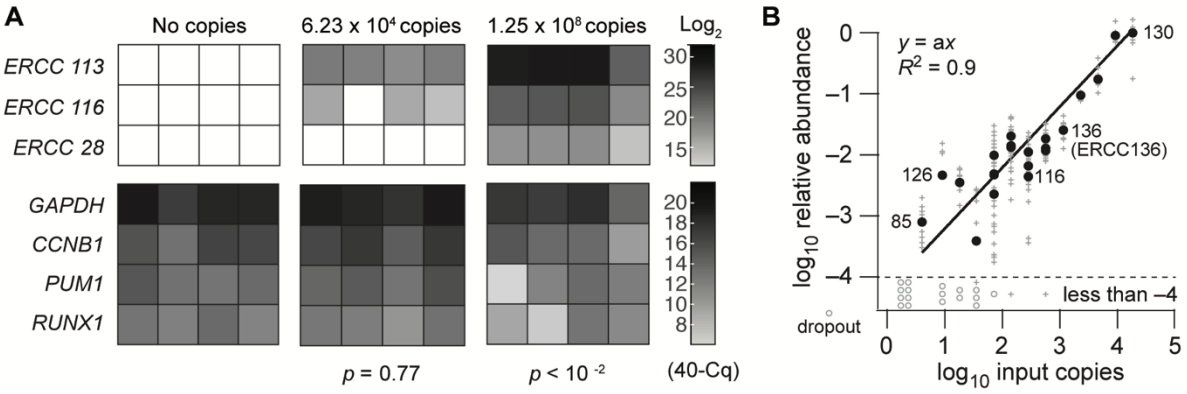

Figure 2.9 Optimized ERCC spike-in dilutions assess poly(A) PCR sensitivity and dynamic range without suppressing cDNA amplification of endogenous transcripts.
(A) 100 pg RNA was supplemented with ERCC Mix 1 at the indicated dilutions and amplified via optimized poly(A) PCR. ERCC and endogenous gene abundances were measured by qPCR, and data are shown in grayscale as the inverse quantification cycle (40–Cq) from n = 4 amplification replicates. Negative effects of the ERCC spike-ins on endogenous genes (lower) were assess by two-way ANOVA with replication.
(B) ERCC Mix 1 (6.23 x 10^4 copies) was spiked into 100 pg RNA and amplified via optimized poly (A) PCR. Proportional abundance of ERCC standards was estimated with a seven-log dilution series from purified qPCR end products. Data are shown as the median 40–Cq (black) for 22 ERCC spike-in standards from n = 8 amplification replicates (gray) with undetected "dropouts" reported below (circles).

For RNA extraction from the LCM cap, an optimized digestion buffer is used containing proteinase K to release mRNAs from precipitated ribosomes (129). Proteinase K also digests nucleosomes, which may cause elution of contaminating genomic DNA. In past and current analyses of human LCM samples preamplified ± reverse transcription, we never found genomic copies of genes amplified within ~0.4% of measured mRNA transcripts (ΔCq ≥ 8 for 16 genes measured in four human cell types, **Figure 2.10**). For mouse tissues, however, genomic copies were more prevalent and variable, with some genes measured as abundantly without reverse transcription as with it (**Figure 2.11A** and **Figure 2.10**). Gel electrophoresis showed weak-but-detectable bands above the desired ~500 bp product in preamplifications without reverse transcription, implying nonspecific amplification (**Figure 2.11A**, lower). Concerned that the murine genome could compete with the amplification of cDNA, we appended an intermediate purification following reverse transcription with 5'-biotin-modified oligo$(dT)_{24}$. Biotinylated cDNA was purified on streptavidin-conjugated magnetic beads, which could be separated from contaminants in the LCM extract and used as a starting template for poly(A) preamplification. Addition of the biotin cleanup step mildly improved the amplification of cDNAs and, importantly, eliminated the confounding abundance of murine genomic DNA (**Figure 2.11B**). We recommend biotinylated oligo$(dT)_{24}$ and bead purification for mouse samples considering the recurrent challenges with genomic DNA (**Figure 2.10** and see Discussion).

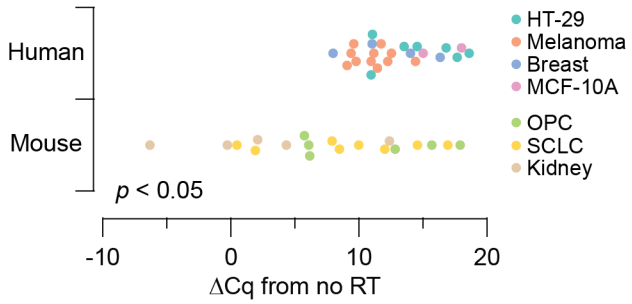

Figure 2.10 Prevalence of genomic DNA contamination during poly(A) amplification of mouse tissue.
Differences in quantification cycles between LCM samples ± reverse transcription (ΔCq from no RT) are shown for various genes in HT-29 cells (human colon adenocarcinoma), primary human melanoma and breast cancer, and MCF-10A cells (human breast epithelial) compared to mouse oligodendrocyte precursor cells (OPC), mouse small-cell lung cancer cells (SCLC), and mouse kidney cells isolated by LCM. Human-mouse differences were assessed by rank-sum test.

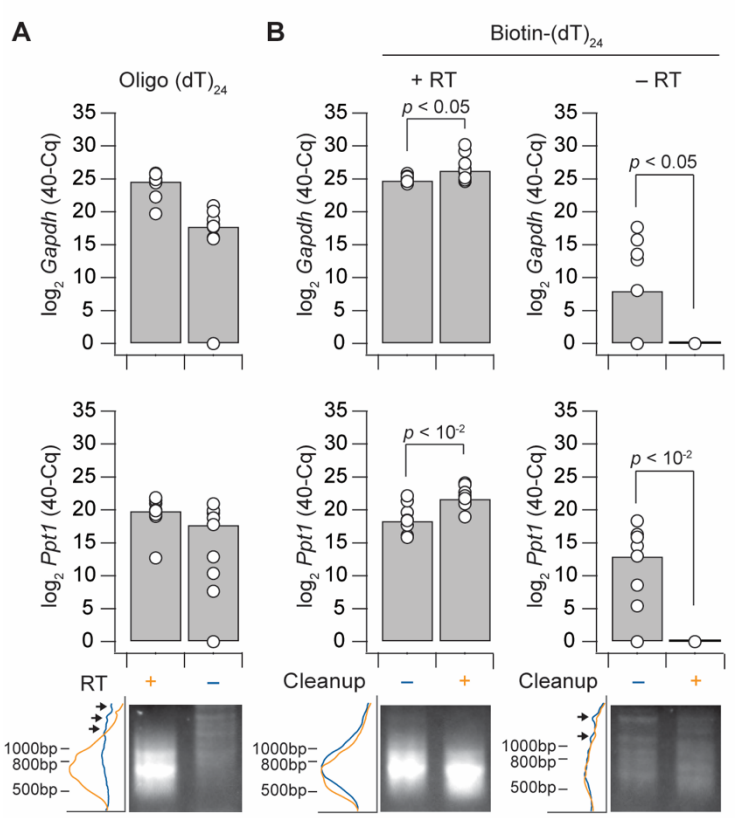

Figure 2.11 Poly(A) amplification of murine sequences without reverse transcription is eliminated with 5'-biotin-modified oligo(dT)$_{24}$ and streptavidin bead cleanup.
(A) Reverse transcription-free preamplification of genomic DNA confounds accurate quantification of some mRNAs.
(B) Bead cleanup eliminates nonspecific preamplification of genomic DNA. Above—Data are shown as the median inverse quantification cycle (40–Cq, gray) of n = 3 independent experiments (three amplification replicates per experiment). Differences with and without bead cleanup were assessed by Wilcoxon rank sum test in MATLAB. Below—Preamplifications were analyzed by agarose gel electrophoresis to separate poly(A)-amplified cDNA from nonspecific, low molecular-weight concatemer (n.s.) and genomic amplification. Electrophoretic traces were analyzed by densitometry to the left of the image, with genomic amplicons highlighted (arrows). Lanes were cropped by the indicated conditions for display but were electrophoresed on the same agarose gel and processed identically.

Poly(A) PCR samples are kept dilute to avoid saturating the preamplification, but aliquots can be carefully reamplified up to microgram scale for microarray hybridization (129,157). In preparing libraries for sequencing, we pursued tagmentation using Tn5 transposase because addition of sequencing adapters is sterically impossible within the ~40 bp distal ends of a PCR amplicon (224). The steric restrictions of Tn5 were advantageous for pruning away the long, A-repetitive universal primer from poly(A) amplicons that would otherwise be wastefully sequenced. Commercial Tn5 tagmentation kits (Nextera XT) require 1000-fold less material than past microarray hybridizations, prompting reevaluation of how the 10-cell libraries were prepared. We retained the mid-logarithmic reamplification approach described previously but substituted paramagnetic Solid Phase Reversible Immobilization (SPRI) beads for library purification (225). Two rounds of purification with 70% (vol/vol) SPRI beads eliminated ~99% of primer dimers and concatemers in 10-cell reamplifications from various sources (**Figure 2.12** and **Figure 2.13**). Reamplified samples yielding at least 200 ng of purified product (**Figure 2.14**) were tagmented at 1-ng scale according to the Nextera XT protocol. Although poly(A) amplicon sizes are centered at ~500 bp (**Figure 2.12A**), we found that the higher SPRI bead ratio recommended for 300–500 bp inputs (180% [vol/vol] beads) was essential for purification of tagmented libraries (**Figure 2.15**). Under these conditions, both new and archival poly(A) PCR preamplifications are compatible with RNA sequencing.

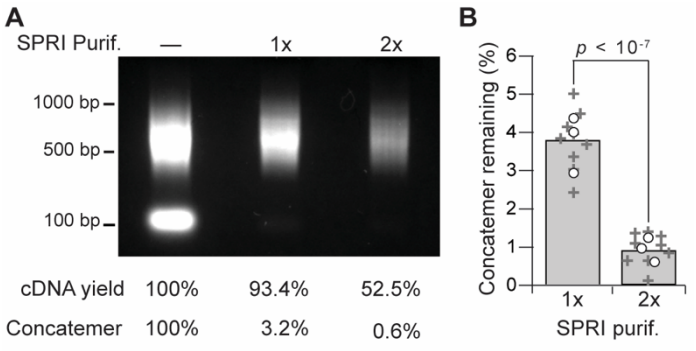

Figure 2.12 Iterative SPRI bead purification eliminates low molecular-weight contaminants before tagmentation.
(A) Poly(A) PCR reamplifications (41) of 10-cell human breast cancer samples were analyzed by gel electrophoresis without purification or after one (1x) or two (2x) rounds of purification with 70% (vol/vol) SPRI beads.
(B) Contaminating low molecular-weight concatemers are significantly reduced after two rounds of SPRI bead purification. Data are shown as the mean (gray) of n = 3 independent reamplifications (circles) each purified three times (+). Differences were assessed by two-way ANOVA with replication.

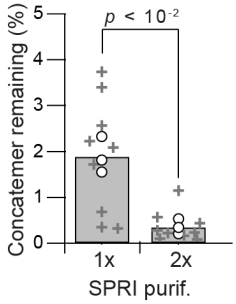

Figure 2.13 Two rounds of SPRI bead purification reduce low molecular-weight contaminants from 10-cell reamplifications of mouse small-cell lung cancer cells.
Data are shown as the mean (gray) of n = 3 independent reamplifications (circles) each purified three times (+). Differences were assessed by two-way ANOVA with replication.

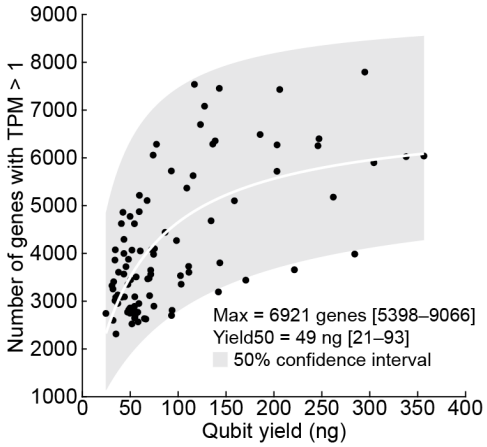

Figure 2.14 Maximal gene-detection sensitivity requires an SPRI bead yield of at-least 200 ng poly(A) cDNA.
Low-coverage RNA sequencing of mouse oligodendrocyte precursor cells or transformed derivatives (n = 96) was used to relate gene-detection sensitivity to SPRI bead yield quantified by Qubit fluorescence through a hyperbolic function with the indicated parameters (Max, Yield50).

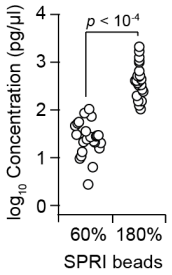

Figure 2.15 Higher SPRI bead ratio is essential for purification of tagmented libraries.
TapeStation concentrations of sequencing libraries following tagmentation and purification with either 60% or 180% [vol/vol] beads. Differences in library concentrations were assessed by paired two-tailed t test.

2.3.3 Paired comparison of 10-cell transcriptomics by BeadChip microarray and RNA-seq

Poly(A) PCR provides an abundant source of material for transcript quantification, creating an opportunity to revisit 10-cell samples profiled earlier on BeadChip microarrays. In the original application of stochastic profiling, 10-cell samples were locally microdissected from 3D spheroids of a clonal human breast-epithelial cell line (129). We sequenced 18 biological replicates from this study (6.6 ± 2.3 million reads) along with three 10-cell pool-and-split controls that assessed technical variability (129,149). Technical correlation was as high within pool-and-split replicates measured by RNA-seq as when the same replicates were measured by microarray (R ~ 0.9; **Figure 2.16B,C,D,F–H)**. For both platforms, undetectable genes in one technical replicate were quantified up to ~10^2 = 100 transcripts per million (TPM) or ~$10^{3.3}$ = 2000 BeadChip fluorescence intensity in another replicate. Among detected genes with at-least one technical replicate yielding zero measured TPM, we found that RNA-seq correlated with BeadChip intensity across replicates (R ~ 0.4, p ~ 0; **Figure 2.17A**). The concordance between the two platforms strongly argues that transcript losses are authentic dropout events, not artifacts of RNA-seq read depth or BeadChip detection sensitivity (154). Combining the reliable detection limits of 100 TPM (**Figure 2.16B,C,F**) and ~250 ERCC copies/reaction (**Figure 2.9B**), we predict (250 copies/reaction)/(10 cells/reaction x 100 TPM) = 250,000 mRNA copies per cell, consistent with published estimates (197).

When 10-cell transcript representation was compared, we found that RNA-seq TPM and BeadChip microarray intensities were correlated (R ~ 0.6; **Figure 2.16A,E,I**), albeit not as strongly as reported elsewhere (204,226). Some genes yielded background fluorescence on microarrays but moderate- to-high TPM, likely due to BeadChip probe sequences absent from the amplicons generated by poly(A) PCR. Among genes with a median TPM > 1000 by RNA-seq, we identified 27 BeadChip probes exhibiting a median fluorescence less than $10^{2.5}$. The median distance of the 27 probes from the 3’ end of the corresponding gene was 845 bases

(IQR: 492–1392 bases), upstream of the distal ~500 bp 3' ends amplified by poly(A) PCR. The probe-independent nature of RNA-seq reinforces one of its critical advantages for 10-cell transcriptomics.

We also evaluated quantitative concordance of the 18 10-cell samples measured both by BeadChip microarray and RNA-seq. The variance of 7713 genes was twice their mean value measured on each platform, suggesting significant biological variation across the 18 samples ($p < 0.01$). For biologically variable genes, the median sample-by-sample Pearson correlation between BeadChip microarray and RNA-seq was 0.42 (interquartile range: 0.16–0.63), with 599 transcripts showing $R \geq 0.8$ (**Figure 2.17B**). Considering a median TPM of 17 (interquartile range: 4–49) for the 10-cell data analysed, these cross-platform correlations fall within the range reported for TCGA microarrays and RNA-seq ($R \sim 0.4–0.9$) (226). Our retrospective analysis indicates that 10cRNA- seq data corroborate BeadChip microarrays and provide broader access to 3' mRNA ends not represented on oligonucleotide probe sets.

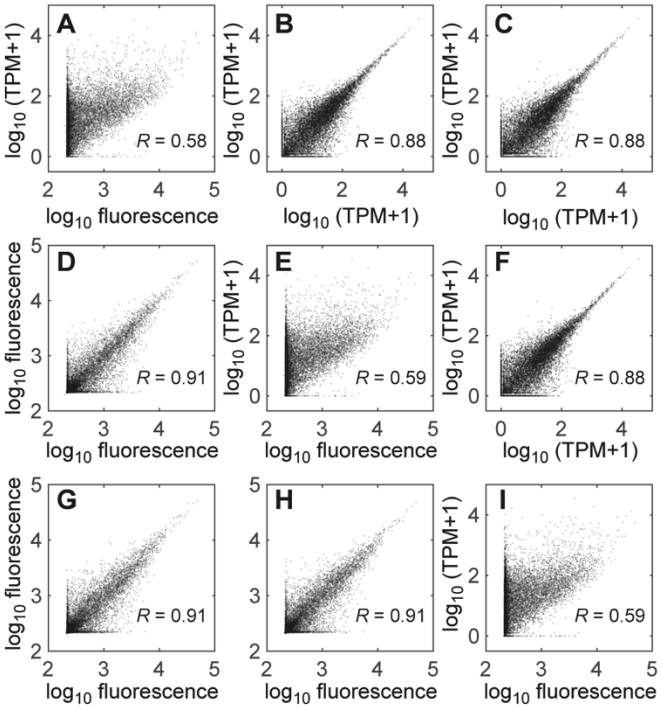

Figure 2.16 Paired comparison of 10-cell transcriptomes profiled by BeadChip microarray and 10cRNA-seq.

(A–I) Three pool-and-split 10-cell replicates from before (41) were reamplified, purified, and tagmented for RNA-seq.

Inter-replicate correlations among BeadChip microarray triplicates **(D,G,H)** and 10cRNA-seq triplicates **(B,C,F)** as well as intra-replicate correlations between platforms **(A,E,I)** are shown together with the log-scaled Pearson correlation (R).

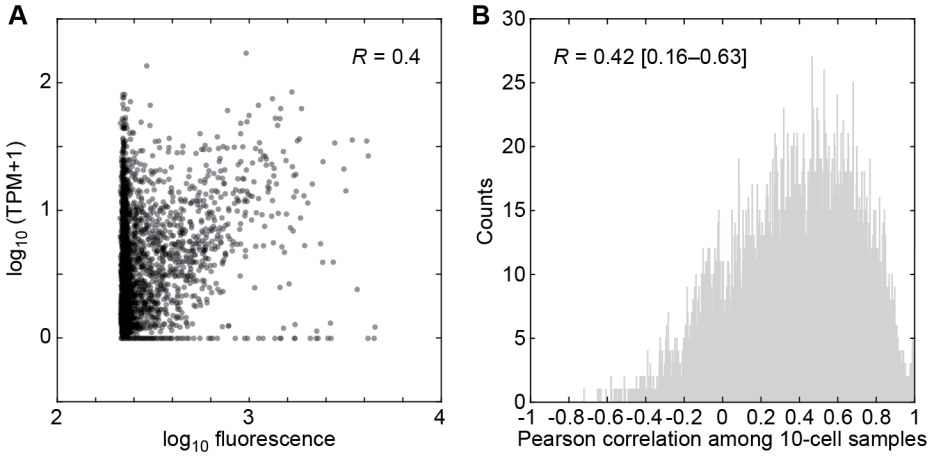

Figure 2.17 Significant technical and biological covariation between BeadChip microarray and 10cRNA-seq.
(A) Variably detected genes remain correlated between transcriptomic platforms. Genes with at-least one pool-and-split TPM = 0 (*n* = 3256 genes) were plotted versus BeadChip fluorescence. **(B)** Significant sample-to-sample correlations between independent 10-cell pools (*n* = 18) measured by BeadChip microarray and RNA-seq. The median log-scaled Pearson correlation (*R*) is shown with the interquartile range in brackets for 7713 genes.

2.3.4 Advantages of 10cRNA-seq for diverse mouse and human cell types

Last, we aggregated the intermediate revisions to 10-cell transcriptomic profiling (**Figure 2.2**) and asked whether there were more-overarching benefits to sequencing small pools versus single cells. Different methods for scRNA-seq have already been rigorously compared by multiple groups (134,150). Since a 10-cell approach could be adopted by many of these approaches, we focused instead on the data quality from published scRNA-seq datasets of various types relative to similar cells profiled by our 10cRNA-seq approach, including biological replicates and pool-and-split controls. We identified two scRNA-seq datasets for murine OPCs, two for murine lung neuroendocrine cells, two for human breast cancer, and one for MCF-10A cells (**Table 2.1**) (19,227–231). All raw data were identically processed and aligned to the transcriptome with RSEM (232). Using transcriptome references stringently emphasized exonic read alignments, and the RSEM model for expectation maximization enabled the degeneracy of 3'-end sequences to contribute to transcript quantification. Data quality was gauged by the percentage of reads aligned, and sensitivity was assessed by the number of Ensembl genes with an estimated TPM greater than one.

For the mouse cell types, we observed significant increases in gene detection between 10cRNA-seq and certain scRNA-seq datasets (**Figure 2.18A**). OPCs isolated by fluorescence-guided LCM showed increased gene detection with 10cRNA-seq compared to scRNA-seq of OPCs purified by fluorescence-activated cell sorting (GSE75330). Gene detection in the sorted OPCs was poorer than when OPCs were collected randomly in a cell atlas of the mouse cortex (GSE60361), emphasizing the stresses caused by non-LCM methods of enrichment. We were unable to detect a significant increase in gene detection between small-cell lung cancer cells profiled by 10cRNA-seq and single neuroendocrine cells randomly dissociated from the mouse airway and profiled by plate- based scRNA-seq. However, neuroendocrine cells are so rare in this tissue that plate-based scRNA-seq was very underpowered (n = 5 cells). When droplet-

based scRNA-seq was used to increase statistical power to n = 92 cells, there was a significant reduction in gene counts compared to 10cRNA-seq profiling the equivalent of 120 cells (n = 12 10-cell replicates). Results were similar but even more striking for human cell types (**Figure 2.18C**). 10cRNA-seq of MCF-10A cells and primary breast cancer cells showed high alignment rates and routinely detected more than 10,000 Ensembl genes, the upper limit for any single cell profiled by three different scRNA-seq methods (19,230,231). In cases where gene sensitivities were comparable, we noted dramatically improved alignment rates for 10cRNA-seq (**Figure 2.18B,D**), reinforcing the efficiency of data collection by adopting a 10-cell approach.

The increased detection of transcripts in 10cRNA-seq data could arise from the accumulation of sporadic gene-expression events among single cells in the 10-cell pool. 10cRNA-seq collects 10-cell pools that are histologically indistinguishable by LCM, but it does not control for noisy transcriptional bursting or differences in cell-cycle phase. To evaluate whether the 10cRNA-seq detection statistics were consistent with those from scRNA-seq data, we randomly combined similar single-cell transcriptomes into 10-cell groups, modelling dropouts as a binomial probability for RNA-to-cDNA conversion (see Methods). We aggregated 48 random 10-cell assemblies within each of the six scRNA-seq datasets and noted a significant increase in gene counts that was comparable to 10cRNA-seq data (**Figure 2.19**). On a per-cell basis, 10cRNA-seq matches the gene-recovery sensitivity of scRNA-seq and may be preferable when isolating single cells in situ is critical.

Table 2.1 Characteristics of published RNA-seq datasets analyzed in this study.
Listed are accession numbers, tissue sources, method of transcript capture, read type and length, as well as read depth in millions (median and range).

Study	Tissue	Method	Read length	Read depth x 10^6
GSE75330	Mouse OPC	Full-length	50bp single-end	1.34 (0.6-8.4)
GSE60361	Mouse OPC	Full-length	50bp single-end	1.85 (0.6-8.4)
GSE103354 Plate-based	Mouse lung	Full-length	75bp single-end	0.32 (0.06-0.6)
GSE103354 Droplet	Mouse lung	3'-end	75bp single-end	0.07 (0.001-0.84)
GSE66357	MCF10A cells	3'-end*	75bp paired-end**	0.17 (0.004-1.8)
GSE113197	Human breast	Full length	100bp paired-end	1.5 (0.00001- 4.4)
PRJNA396019	Human breast	3'-end***	75bp and 100bp paired-end	0.88 (0.002-20.6)

*RNA printing
**Asymmetric paired reads: 25bp barcode sequence read 1 and 60bp sequence read 2 (231)
***Single-nucleus RNA-seq

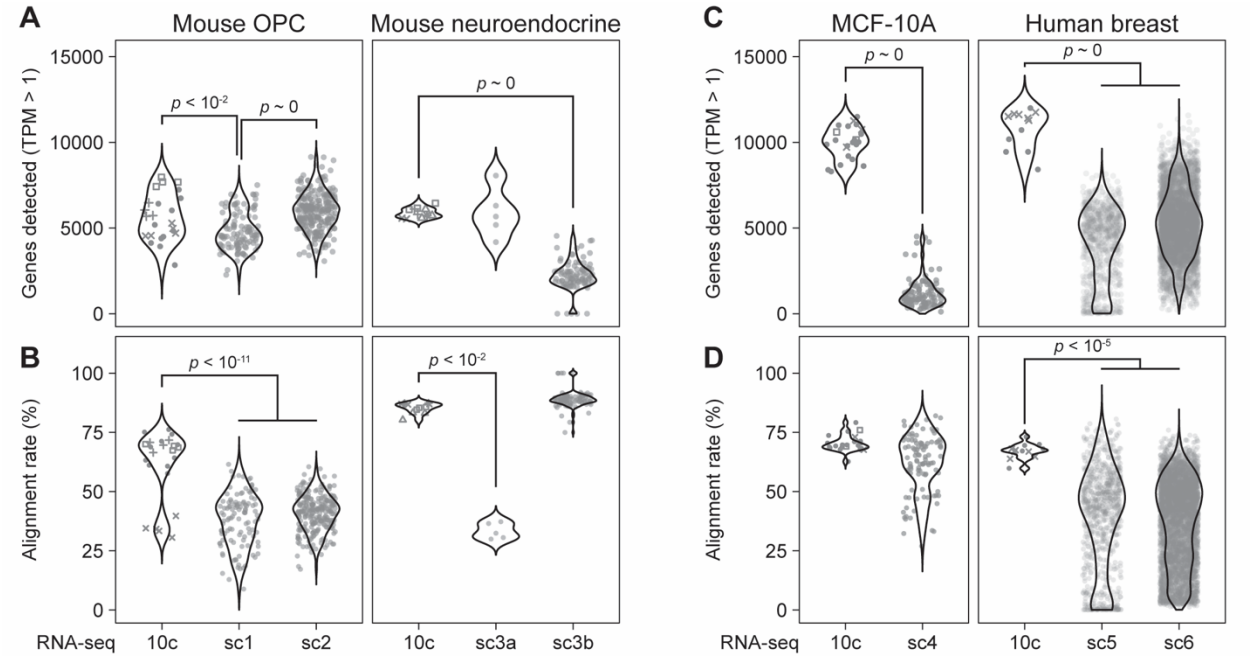

Figure 2.18 Increased gene detection and exonic alignment rates for 10cRNA-seq compared to scRNA-seq.
(A) Detection of murine Ensembl genes for mouse oligodendrocyte precursor cells (OPCs) and lung neuroendocrine-derived cells.
(B) Exonic alignment rate comparison for OPCs and lung neuroendocrine-derived cells.
(C) Detection of human Ensembl genes for MCF-10A cells and human breast cancer cells.
(D) Exonic alignment rate comparison for MCF-10A cells and human breast cancer cells.
Public scRNA-seq data were obtained from the indicated accession numbers: sc1=GSE75330, sc2=GSE60361, sc3a=GSE103354 (plate-based), sc3b=GSE103354 (droplet-based), sc4=GSE66357, sc5=GSE113197, sc6=PRJNA396019. 10cRNA-seq data were aggregated from independent 10-cell samples (circles) and 10-cell equivalents from pool-and-split controls. Pool-and-split controls from the same day are indicated with non-circular markers corresponding to the shared day. Pairwise differences between 10-cell and single-cell methods were assessed by permutation test.

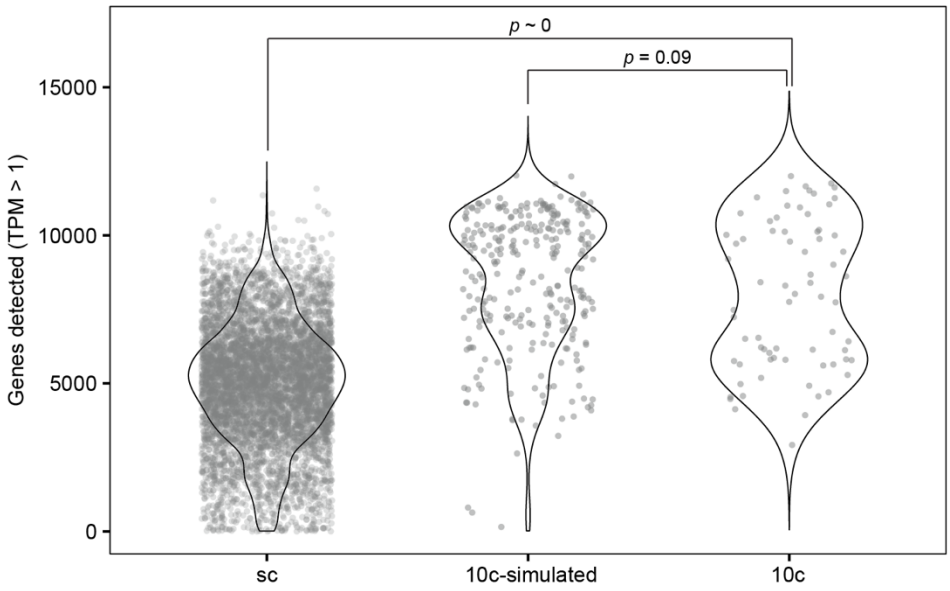

Figure 2.19 Gene-detection sensitivity increases predictably to that of 10cRNA-seq when scRNA-seq data are aggregated as 10-cell pools.
Single cells from various scRNA-seq datasets (described in **Figure 2.18**) were randomly sampled and grouped with their nine nearest neighbours, and dropouts were modelled using a binomial process for RNA-to-cDNA conversion (n = 48 random samples for each of six datasets; see Methods). Gene detection from the original scRNA-seq datasets (sc, reprinted from **Figure 2.18**), the simulated 10-cell pools (10c-simulated), and 10cRNA-seq (10c, reprinted from **Figure 2.18**) were compared by permutation test.

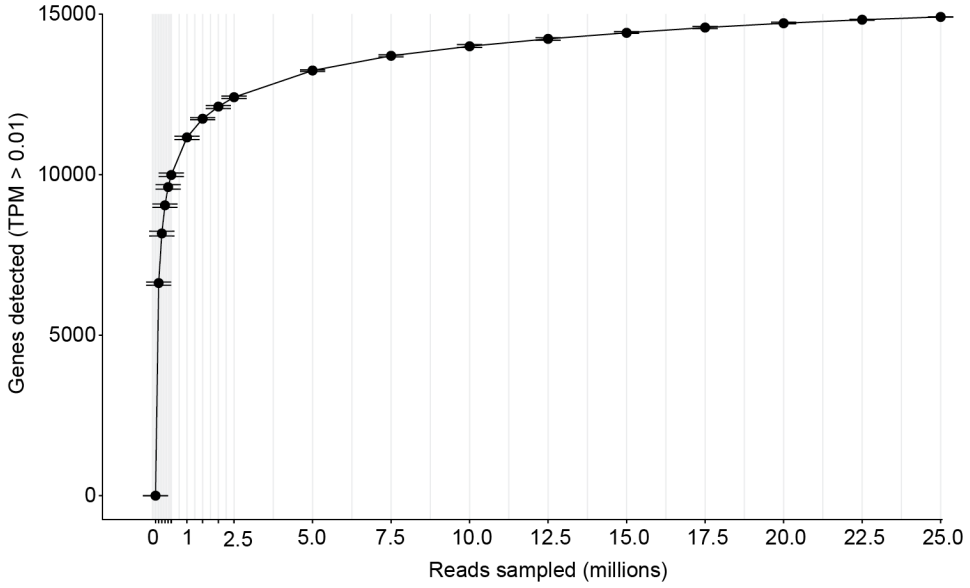

Figure 2.20 10cRNA-seq gene detection saturates above 5 million reads per sample. 10cRNA-seq reads from MCF10A-5E cells were aggregated, randomly sampled at the indicated depth, and aligned. Data are shown as the median number genes detected (TPM > 0.01) ± range of *n* = 10 random draws per depth.

2.4 Discussion

Single-cell transcriptomics has expanded or rewritten the catalogue of cell types in tissues, organs, and organisms (229,233–238). Yet, scRNA-seq does not obviate the need for complementary approaches, which accurately profile regulatory-state changes within a given cell lineage (70). The technical advances reported here demonstrate the immediate feasibility of 10cRNA-seq for mouse and human samples obtained in situ by LCM (Chapters 3 and 4). We combined straightforward extensions of ERCC spike-ins and tagmentation with new approaches for fluorescence-guided LCM and cDNA purification that may prove beneficial for other applications (**Figure 2.2**). Although small-sample RNA-seq is never fully dissociated from tissue acquisition or cell handling, our data illustrate a workflow that can be paused and restarted when LCM is used as an intermediate step.

Previous descriptions of fluorescence-guided LCM relied upon exogenous fluorophores added by lectins, antibodies, or viruses (193,209,210,239). Through careful optimization of cryoembedding and LCM, we identified conditions that preserved the most-common fluorescent proteins used to engineer the mouse germ line. Compatibility with genomically encoded labels creates new opportunities for combining 10cRNA-seq with lineage tracing to examine early regulatory-state changes in development and disease (240). Compared to fluorophore localization, RNA integrity was not as exquisitely sensitive to sample preparation and handling. Nevertheless, we recommend fresh cryoembedding of all samples in case other protein-guided approaches, such as immuno-LCM, might be pursued (241). The breast core biopsies profiled here were prospectively obtained and cryoembedded during an outpatient procedure. However, a nearly identical protocol has been deployed intra-operatively for surgical pathology, implying that fresh cryoembedding is not prohibitive for biobanked clinical samples (242).

A startling result from the revised protocol was the extent of poly(A) amplification observed in murine samples when reverse transcription was omitted. Nonspecific amplification

was not as prominent in human samples obtained by LCM, pointing to specific differences in genome composition and the susceptibility to priming with AL1. A plausible explanation lies in transposable elements—specifically, the distinct classes of short interspersed nuclear elements (SINEs) in rodents and humans (243). Human-specific Alu SINEs and rodent-specific B-type SINEs both contain stretches of 10–20 As that could partially anneal to the T homopolymer sequence on the 3' end of AL1 (244). However, to amplify during poly(A) PCR, an antisense SINE must be sufficiently nearby. The mouse genome is ~20% smaller than humans, and B-type SINEs are ~25% more numerous in mice compared to Alu SINEs in humans (243). The differences reduce the expected spacing of sense-antisense SINEs from ~6 kb in humans to ~4 kb in mice, consistent with a prior analysis of sense-antisense SINEs around transcription start sites (245). The shorter average spacing may be close enough for genomic fragments to compete with the ~500 bp cDNA amplicons generated during reverse transcription (**Figure 2.6**, **Figure 2.11A**). Such nonspecific products were prevented from coamplifying with cDNA by using biotinylated oligo$(dT)_{24}$ and streptavidin beads, akin to the bead capture and primer extension of droplet-based approaches (140,246). This strategy may prove useful in other non-murine settings, such as suspension cells, where genomic contamination will be more extensive than with LCM (157).

ERCC spike-ins provide a standard to compare 10cRNA-seq against single-cell methods for transcriptomic profiling. Using the metrics of Svennson et al., we estimate a 50% detection sensitivity of 45 copies per reaction (90% nonparametric CI: [15–485]) and a Pearson product-moment correlation coefficient of R = 0.86 (90% nonparametric CI: [0.71–0.91] from n = 72 samples). The R accuracy is somewhat lower than prevailing techniques, but that may be overly pessimistic because 10cRNA-seq uses such a dilute mix of spike-ins (4 million-fold dilution of the ERCC stock). Detection sensitivity is comparable to that reported for the most popular plate-based scRNA-seq methods, including SMART-seq2 and CEL-seq (136,247). The strength of 10cRNA-seq lies in the use of 10- cell pooling to improve the per-cell technical sensitivity

beyond the best microfluidic- and droplet- based approaches for scRNA-seq (150). LCM minimizes disruptive tissue handling and provides histologic cues for microdissecting pools of cells within the same lineage. Adopting a 10-cell approach may prove similarly beneficial for other microdissection-based approaches, such as GEO-seq and the recent pairing of SMART-seq2 with LCM (144,192).

When 10cRNA-seq was compared to scRNA-seq, we often observed significant improvements in exonic alignment. Methods for scRNA-seq typically yield exonic alignment rates below 50%, with the remainder of aligned reads splitting equally between intronic and intergenic sequences (136,231). 10cRNA-seq achieves exonic alignments of 70% or higher despite using oligo(dT)- primed reverse transcription with the same potential to prime internal A homopolymer sequence as with scRNA-seq (248,249). Interestingly, in one instance of similarly high exonic alignment (GSE66357, **Figure 2.18B**), the RNA-printing approach to scRNA-seq incorporated a DNase treatment absent from all other methods (231). This study also yielded a significantly reduced gene-detection sensitivity compared to 10cRNA-seq. Commingling genomic DNA may dilute exonic alignment percentages and inflate the number of genes detected due to chance sequencing of genomic DNA from exonic loci. Multiple scRNA-seq approaches incorporate unique molecular identifiers appended to oligo(dT) (150,250). The identifiers avoid redundantly counting the same product of reverse transcription, and they also retrospectively exclude sequenced reads that do not come from cDNA. The biotin cleanup approach we devised for mouse cells (**Figure 2.11**) achieves cDNA selection prospectively in situations where genomic contamination may be problematic.

Our work illustrates that 10-cell profiling can extend beyond microarrays and quantitative PCR to compete favorably with scRNA-seq. Although ill-suited for lineage mapping of highly mixed cell populations, 10cRNA-seq exploits the precision of LCM to target specific cell types in situ and define their regulatory heterogeneities. LCM is also advantageous for sequencing cells that are delicate or difficult to dissociate rapidly (70,144). In Chapters 3 and 4, we apply

10cRNA-seq to cancer biology to characterize the diversification of tumour cells in patient samples and animal models.

2.5 Methods

2.5.1 Cell and tissue sources

The MCF10A-5E breast epithelial cell samples were described previously (129). KP1 small-cell lung cancer cells were grown as spheroids in RPMI Medium 1640 with 10% FBS, 1% penicillin- streptomycin, and 1% glutamine (223). KP1 spheroids were pelleted and mixed in Neg-50 (Richard-Allan Scientific) before cryoembedding. Animal housing and experimental procedures were carried out in compliance with regulations and protocols approved by the IACUC at the University of Virginia. Cspg4-CreER;Trp53$^{F/F}$;Nf1$^{F/F}$;Rosa26-LSL-tdT mice were housed in accordance with IACUC Protocol #3955 at the University of Virginia (212). As per the approved protocol, animals were administered 200 mg/kg tamoxifen by oral gavage for five days, and brains were harvested at 12 days or 183 days after the last administration. A labelled glioma arising the olfactory bulb at 165 days after the last tamoxifen administration was also used. Human samples acquisition and experimental procedures were carried out in compliance with regulations and protocols approved by the IRB-HSR at the University of Virginia. In accordance with IRB Protocol #19272, breast cancer samples were collected as ultrasound-guided core needle biopsies during diagnostic visits from participants who provided informed consent. Each core biopsy was divided into multiple pieces before cryoembedding. Unless otherwise indicated, all samples were freshly cryoembedded in a dry ice-isopentane bath and stored at –80˚C wrapped in aluminium foil.

2.5.2 Cryosectioning

Samples were equilibrated to –24˚C in a cryostat before sectioning. 8 µm sections were cut and wicked onto Superfrost Plus slides. To preserve fluorescence localization of tdT and EGFP, slides were precooled on the cutting platform for 15–30 sec before wicking, and the

section was carefully placed atop the cooled slide with forceps equilibrated at –24˚C. Then, the slide was gently warmed from underneath by tapping with a finger until the section was minimally wicked onto the slide. All wicked slides were stored in the cryostat before transfer to –80˚C storage on dry ice. Frost build-up was minimized by storing cryosections in five-slide mailers.

2.5.3 Staining, dehydration, and laser-capture microdissection

For cryosections lacking fluorophores, slides were stained and dehydrated as described previously (129,157). Briefly, slides were fixed immediately in 75% ethanol for 30–60 sec, rehydrated quickly with water, stained with nuclear fast red (Vector Labs) containing 1 U/ml RNAsin-Plus (Promega) for 15 sec, and rinsed two more times with water before dehydrating with 70% ethanol for 30 sec, 95% ethanol for 30 sec, and 100% ethanol for 1 min and clearing with xylene for 2 min. tdT- and EGFP- labelled cryosections were not stained and instead began with the 70% ethanol dehydration step that also provided solvent fixation. After air drying, slides were microdissected immediately on an Arcturus XT LCM instrument (Applied Biosystems) using Capsure HS caps (Arcturus). The smallest spot size was used, and typical instrument settings of ~50 mW power and ~2 msec duration yielded ~25 µm spot diameters capturing 1–3 cells per laser shot.

2.5.4 RNA extraction and first-strand synthesis

RNA extraction and first-strand synthesis were similar to earlier protocols with some minor modifications (129,157). HS caps were eluted for 1 hr at 42˚C with 4 µl of digestion buffer containing 1.25x First- strand buffer (Invitrogen), 100 µM dNTPs (Roche), 0.08 OD/ml oligo$(dT)_{24}$ with or without 5'-biotin modification (IDT), and 250 µg/ml proteinase K (Sigma). Samples containing ERCC spike-ins included a four-million-fold dilution of ERCC spike-in mixture 1 (Ambion). Eluted samples were centrifuged into 0.5 ml PCR tubes at 560 rcf for 2 min, the digestion buffer was quenched with 1 µl of digestion stop buffer containing 2 U/µl SuperAse-in (Invitrogen) and 5 mM freshly prepared PMSF (Sigma). 4.5 µl of the quenched extract was

transferred to a 0.2 ml PCR tube, and reverse transcription was performed with 0.5 µl of SuperScript III (Invitrogen) for 30 min at 50˚C followed by heat inactivation at 70˚C for 15 min. Samples were placed on ice and centrifuged for 2 min at 18,000 rcf on a benchtop microcentrifuge.

2.5.5 Streptavidin bead cleanup of biotinylated first-strand products

For 5'-biotin-containing samples, streptavidin magnetic beads (Pierce) were prepared in a 0.2 ml PCR tube on a 96S Super Magnet Plate (Alpaqua). Beads (6 µl per sample) were magnetized, aspirated, and resuspended in binding buffer (5 µl per sample) containing 1x First-strand buffer (Invitrogen), 4 M NaCl, and 0.02% (vol/vol) Tween-20. 5 µl of resuspended beads were added after first-strand synthesis, and samples were incubated for 60 min at room temperature with mixing every 15 min. Beads were pelleted on the magnet plate, resuspended in 100 µl high-salt wash buffer (50 mM Tris [pH 8.3], 2 M NaCl, 75 mM KCl, 3 mM $MgCl_2$, 0.01% Tween-20). Beads were pelleted again on the magnet plate, and the pellet was washed once with 100 µl high-salt wash buffer. Next, beads were resuspended in 100 µl low-salt wash buffer (50 mM Tris [pH 8.3], 75 mM KCl, 3 mM $MgCl_2$) and transferred to a fresh 0.2 ml PCR tube. Beads were pelleted again on the magnet plate, and the pellet was washed once with 100 µl low-salt wash buffer. After the last wash, the beads were resuspended in 5 µl 1x First-strand buffer for RNAse H treatment and poly(A) tailing.

2.5.6 RNAse H treatment and poly(A) tailing

RNAse H digestion and poly(A) tailing were performed exactly as described previously (129,157). Briefly, template mRNA strands were hydrolysed for 15 min at 37˚C with 1 µl of RNAse H solution containing 2.5 U/ml RNAse H (USB Corporation) and 12.5 mM $MgCl_2$. After RNAse H treatment, cDNA templates were poly(A)-tailed with 3.5 µl of 2.6x tailing solution containing 80 U terminal transferase (Roche), 2.6x terminal transferase buffer (Invitrogen) and 1.9 mM dATP. The tailing reaction was incubated for 15 min at 37˚C and then heat-inactivated

at 65˚C for 10 min. Samples were placed on ice and spun for 2 min at 18,000 rcf on a benchtop centrifuge.

2.5.7 Poly(A) PCR

Poly(A) PCR was carried out with several modifications to the earlier procedure (129,157). To each tailed sample, 90 µl of poly(A) PCR buffer was added to a final concentration of 1x ThermoPol buffer (New England Biolabs), 2.5 mM $MgSO_4$, 1 mM dNTPs (Roche), 100 µg/ml BSA (Roche), 3.75 U Taq polymerase (NEB) and 1.5 U Phusion (NEB) and 2.5 µg AL1 primer (ATTGGATCCAGGCCGCTCTGGACAAAATATGAATTCTTTTTTTTTTTTTTTTTTTTTT). Each reaction was split into three thin-walled 0.2 ml PCR tubes and amplified according to the following thermal cycling scheme: four cycles of 1 min at 94˚C (denaturation), 2 min at 32˚C (annealing) and 2 min plus 10 sec per cycle at 72˚C (extension); 21 cycles of 1 min at 94˚C (denaturation), 2 min at 42˚C (annealing) and 2 min 40 sec plus 10 sec per cycle at 72˚C (extension). The tubes were cooled, placed on ice, and the reactions from three tubes for each sample were pooled and amplified according to the following thermal cycling scheme: five cycles of 1 min at 94˚C (denaturation), 2 min at 42˚C (annealing) and 6 min at 72˚C (extension). Amplified samples were stored at −20°C until further use.

2.5.8 Poly(A) PCR re-amplification

For sequencing, poly(A) PCR cDNA samples were reamplified as before in a 100 µl PCR reaction containing 1x High-Fidelity buffer (Roche), 3.5 mM $MgCl_2$, 200 µM dNTPs (Roche), 100 µg/ml BSA (Roche), 5 µg AL1 primer, and 1 µl of poly(A) PCR sample. Each reaction was amplified according to the following thermal cycling scheme: 1 min at 94˚C (denaturation), 2 min at 42˚C (annealing) and 3 min at 72˚C (extension). The appropriate number of PCR cycles was determined by a pilot reamplification containing 20 µl of the PCR reaction above plus 0.25x SYBR Green monitored on a CFX96 real-time PCR instrument (Bio-Rad). The number of amplification cycles for each sample was selected to ensure that the reamplification remained in

the exponential phase and there was sufficient cDNA for SPRI bead purification (typically 5–12 cycles).

2.5.9 qPCR

For detection of specific targets in poly(A) PCR samples, qPCR was performed on a CFX96 real-time PCR instrument (Bio-Rad) as previously described (251). 0.1 µl or 0.01 µl of each preamplification was used with the qPCR primers listed in Supplementary Table S2. For relative quantification between ERCC spike-ins, qPCR amplicons were purified by gel electrophoresis, extracted, ethanol precipitated, and quantified by spectrophotometry. Purified amplicons were used to create a six-log standard curve based on ERCC amplicon copy number. All spike-ins were normalized to ERCC130 copy numbers to obtain relative abundance.

2.5.10 SPRI bead purification

Re-amplified samples were purified twice with 70% (vol/vol) Ampure Agencourt XP SPRI beads. SPRI beads were equilibrated to room temperature for 30 min, and 70 µl beads were added to the 100 µl reamplification product. After a 15-min incubation at room temperature, samples were magnetized for 5 min. The supernatant was removed with a gel-loading pipette tip, leaving ~5 µl volume in the well. Beads were gently washed twice on the magnet with 200 µl freshly prepared 80% (vol/vol) ethanol and aspirated with a gel-loading pipette tip. Residual ethanol was removed after the second wash, and beads were air-dried at room temperature for 10 min before resuspension in 10 µl elution buffer (10 mM Tris-HCl [pH 8.5]). Samples were magnetized at room temperature for 1 min, and the eluted supernatant was transferred to a new 0.2 ml PCR tube. The 10 µl elution was purified a second time with 7 µl beads and the same incubation, ethanol wash, and elution conditions as the first purification.

2.5.11 RNA sequencing and analysis

Bead-purified cDNA libraries were quantified with the Qubit dsDNA BR Assay Kit (Thermo Fisher) using a seven-point standard curve and a CFX96 real-time PCR instrument (Bio-Rad) for detection. Samples were diluted to 0.2 ng/µl before tagmentation with the Nextera

XT DNA Library Preparation Kit (Illumina) according to the manufacturer's earlier recommendation to purify libraries with 180% (vol/vol) SPRI beads (**Figure 2.14**). For each run, samples were multiplexed at an equimolar ratio, and 1.3 pM of the multiplexed pool was sequenced on a NextSeq 500 instrument with NextSeq 500/550 Mid/High Output v1/v2 kits (Illumina) to obtain 75-bp paired-end reads at an average depth of 4.2 million reads per sample (**Figure 2.13**) or 7.5 million reads per sample (all others). Simulated read depths of 10cRNA-seq data from MCF10A-5E cells confirmed saturation of gene detection above ~5 million reads per sample (**Figure 2.20**). Adapters were trimmed using fastq-mcf in the EAutils package (version ea-utils.1.1.2-537) with the following options: -q 10 -t 0.01 -k 0 (quality threshold 10, 0.01% occurrence frequency, no nucleotide skew causing cycle removal). Quality checks were performed with FastQC (version 0.11.7) and multiqc (version 1.5). Datasets were aligned to either the human (GRCh38.84) or the mouse (GRCm38.82) transcriptome along with reference sequences for ERCC spike-ins using RSEM (version 1.3.0) with the following options: --bowtie2 --single-cell-prior --paired-end (Bowtie2 transcriptome aligner, single- cell prior to account for dropouts, paired end reads). RSEM read counts were converted to transcripts per million (TPM) by dividing each value by the total read count for each sample and multiplying by 10^6. Mitochondrial genes and ERCC spike-ins were not counted towards the total read count during TPM normalization. The number of genes with TPM > 1 for each sample was calculated relative to the number of unique Ensembl IDs for the organism excluding ERCC spike-ins.

2.5.12 Analysis of public scRNA-seq datasets

FASTQ files were downloaded from GSE75330, GSE60361, GSE103354 (plate-based), GSE66357, GSE113197, and PRJNA396019. FASTQ files were not available for the droplet-based dataset of GSE103354; therefore, BAM files were downloaded from SRR7621182 and converted to FASTQ format. Adapters were trimmed using fastq-mcf with the following options: -q 10 -t 0.01 -k 0 (quality threshold 10, 0.01% occurrence frequency, no nucleotide skew causing cycle removal). To compare with the other datasets, seqtk (version 1.3) was used to clip 15 bp

unique molecular identifiers from the beginning of sequences in GSE60361 and GSE75330. All RNA-seq datasets were aligned to either the human (GRCh38.84) or the mouse (GRCm38.82) transcriptome as well as reference sequences for ERCC spike-ins, using RSEM with the following options: --bowtie2 --single-cell-prior (Bowtie2 transcriptome aligner, single-cell prior to account for dropouts). GSE103354 (plate-based), GSE113197, and PRJNA396019 also used --paired-end (paired-end reads). TPM conversion and gene detection quantification were calculated as above. For post-hoc pooling (**Figure 2.19**), individual scRNA-seq profiles were selected at random (n = 48 per dataset) and grouped with the nine scRNA-seq profiles in the dataset that were nearest by Jaccard distance. To model dropouts, TPM values for each scRNA-seq profile were scaled to expected copies per cell assuming 250,000 mRNA copies per cell38 and transmitted to the 10-cell pool as a binomial random variable (N = expected copies per cell, p = RNA-to-cDNA conversion efficiency = 10% for **Figure 2.19**). Post-hoc pooling results were similar up to a conversion efficiency of ~30%.

2.5.13 Paired analysis of BeadChip microarrays and 10cRNA-seq

Microarray data (GSE120030) were batch processed with the lumi R package using a detection threshold of 0.05 and simple scaling normalization to obtain $\log_2$-normalized values that were converted to $\log_{10}$-normalized values (252). Gene names from the BeadChip files were merged to the extent possible with Ensembl IDs from the RSEM alignments by using HUGO Gene Nomenclature synonym tables to match current and retired gene names.

2.5.14 Monte Carlo simulations

Simulations of stochastic-profiling experiments were performed in MATLAB using StochProfGUI (157). Each parameter set was run 50 times to measure median p values and nonparametric confidence intervals. False positives were called when the median p value was less than 0.05 for a unimodal population (expression fraction = 0). False negatives were called when the median p value was greater than 0.05 for a bimodal population (expression fraction ≠ 0).

2.5.15 Data availability

All 10cRNA-seq data are available through the NCBI Gene Expression Omnibus (GSE120261; Reviewer token: evyzaokcxnwfhcn). Step-by-step protocols for 10cRNA-seq, including critical steps and troubleshooting, are available here as a Supplementary Note and will be maintained on the Janes Laboratory website (http://bme.virginia.edu/janes/protocols/).

3 Pan-cancer Drivers are Recurrent Transcriptional Regulatory Heterogeneities in Early-stage Luminal Breast Cancer

3.1 Foreword

As discussed in Chapter 1, the heterogeneous composition of solid tumors is known to impact disease progression and response to therapy. In Chapter 2, we developed 10cRNA-seq to access transcriptomes of cells *in situ* with increased sensitivity compared to scRNA-seq methods. In this chapter, we revised a statistical fluctuation analysis called stochastic profiling (Chapter 1) and combined it with 10cRNA-seq to identify signatures of different regulatory states of cancer cells. When applied to a cohort of late-onset, early-stage luminal breast cancers, the integrated approach identified thousands of candidate regulatory heterogeneities. Intersecting the candidates from different tumors yielded a relatively stable set of 710 recurrent heterogeneously expressed genes (RHEGs) that were significantly variable in >50% of patients. RHEGs were not confounded by dissociation artifacts, cell cycle oscillations, or driving mutations for breast cancer. Rather, we detected RHEG enrichments for epithelial-to-mesenchymal transition genes and, unexpectedly, the latest pan-cancer assembly of driver genes across cancer types other than breast. Heterogeneous transcriptional regulation conceivably provides a faster, reversible mechanism for malignant cells to sample the effects of potential oncogenes or tumor suppressors on cancer hallmarks.

Acquisition of patient samples profiled in this chapter was enabled by Dr. Jennifer Harvey and Kathy Repich at the UVA Breast Care Clinic. A manuscript of this work with me as first author is under review (253).

3.2 Introduction

Approximately 80% of human tumors are epithelial carcinomas (254). Epithelial cells and progenitors proliferate considerably during normal tissue development, maintenance, and repair (255,256). Proper epithelial organization is enforced by basement membrane extracellular matrix (ECM), which becomes compromised during the epithelial cell-state changes underlying carcinomagenesis (26,95,257). Advanced carcinomas show considerable cell-to-cell variation in chromosomal gains–losses (59), overall mutational burden (54), and hybrid epithelial-mesenchymal traits (98). However, the state trajectory of carcinoma cells within large, rapidly progressing tumors is not stereotyped (258), complicating general interpretations of this variability. It is not known when intra-carcinoma cell heterogeneity meaningfully emerges, nor whether there might be common themes early in tumorigenesis that go on to diverge at later stages.

Thoroughly deconstructing intratumor heterogeneity requires transcriptomic approaches that can separate lineages and distinguish regulatory states with high cellular resolution (70,124). Conventional methods for single-cell RNA sequencing (scRNA-seq) either a) profile dozens– hundreds of cells at the maximum depth afforded by the single-cell sample, or b) shallowly scan (tens of) thousands of cells for molecular phenotyping (150). Regardless of the output, all leading approaches dissociate tumors into single-cell suspensions, requiring up to an hour of tissue processing and yielding a range of carcinoma proportions depending on cancer type (**Table 3.3.1**). The impact of these preparative steps on the transcriptomes of live cells is recognized (143,259), but they are considered to be an unavoidable tradeoff of the scRNA-seq approach.

Considering the drawbacks, we (180) and others (144,192) have developed approaches orthogonal to scRNA-seq that combine high-sensitivity transcriptomics with laser-capture microdissection (LCM) (Chapter 2). Using LCM, histologically distinguishable cell types can be

isolated with single-cell resolution from cryosections of snap-frozen tissue or tumor. Unfortunately, all LCM- based sequencing methods have substantially decreased sensitivity and technical reproducibility with fewer than 10 cells of microdissected material (144,157,192). To gain single-cell information, we reasoned that our 10-cell RNA sequencing (10cRNA-seq) method could productively identify carcinoma cell-regulatory heterogeneities if it were implemented as a type of fluctuation analysis called stochastic profiling (129). Previously, stochastic profiling was applied to 10-cell averaged gene-expression profiles collected repeatedly within a clonal cell line and measured by qPCR or microarray, leading to predicted single-cell heterogeneities that were subsequently validated (129,158,201). The distinct statistical properties of RNA sequencing data required design of a customized analytical pipeline for stochastic profiling by 10cRNA-seq.

Here, we combine 10cRNA-seq with an abundance-dependent overdispersion statistic that enables stochastic profiling of tumor cells *in situ.* Without any sample dissociation, we deeply profile carcinoma cell-to-cell heterogeneity in a cohort of five closely matched, late-onset and early-stage luminal breast cancers (**Table 3.1** and **Table 3.2**), obtaining data on 21,255 genes from 1400 carcinoma cells. The LCM component of 10cRNA-seq proved critical to obtain carcinoma heterogeneity profiles from cases with extensive immune infiltration. 10-cell pooling minimized the contribution of periodic transcripts that covary with cell-cycle phases. Stochastic profiling inferred 710 candidate transcripts that were recurrently heterogeneous in ≥50% of tumors. Subsampling tumors in the cohort consistently yielded 500– 1500 candidates by the same criteria, suggesting bounds for reliable variability between similar tumors. The shared set of candidates was largely devoid of detachment-induced artifacts (143) and, surprisingly, breast-cancer driver genes (260). Recurrent heterogeneities were instead enriched in the collagen and matricellular constituents of a pan-cancer epithelial-to- mesenchymal transition (EMT) signature (261). The heterogeneities uncoupled canonical EMT marker genes that are tightly correlated at the population level. Most intriguing was an enrichment for dozens of non-cycling driver genes

for cancers other than breast (262). Our results raise the possibility that early-stage luminal breast cancers sample a much broader landscape of oncogenes and tumor suppressors through transcriptional heterogeneity than indicated by the genomic lesions characteristic of the subtype.

3.3 Results

3.3.1 Carcinoma-focused 10-cell profiling of early-stage luminal breast cancer

Women were selected for enrollment if they required ultrasound-guided biopsy for a suspected malignancy after screening mammography (BI-RADS 4C and higher). Just before diagnostic biopsy, we obtained written informed consent to collect an additional ultrasound-guided core sample, which was cryoembedded (180) within one minute of acquisition (**Figure 3.1A** and **Table 3.1**). After clinical diagnosis of hormone-positive, *HER2*-negative breast cancer, we selected five cases that were as closely matched as possible (**Table 3.2**). The median tumor biopsy was Stage 1, Grade 3, and aged 63 years—the late-onset group for breast cancer (263).

Despite gross similarities in tumor characteristics, we noted elevated lymphocyte infiltration in two cases (UVABC3 and UVABC5) that rendered them problematic to microdissect by nuclear histology alone (**Figure 3.1B** and **Table 3.2**). Therefore, we devised an immuno-LCM (241) procedure that combines an Alexa Fluor 488-conjugated, high-affinity monoclonal antibody against KRT8 with the red-orange nucleic acid stain YOPRO3 (see Materials and Methods). A one-minute incubation with the antibody-dye cocktail was sufficient to resolve KRT8-positive carcinoma cells from KRT8-negative stromal cells and the YOPRO3- negative autofluorescence of tissue collagen (**Figure 3.2A**). We could not find any evidence that the antibody or dye interfered with the critical early steps of 10cRNA-seq (**Figure 3.2B**) (180).

For each case, we collected 28 random pools of 10 carcinoma cells located throughout cryosections of the core biopsy. Pools were assembled as local 10-cell groups that reflect both

clonal and microenvironmental heterogeneity. We recorded the spatial position of all cells microdissected in the 10-cell samples to leave open the possibility of retroactively linking transcriptomic changes to histological or topological features of the tumor. Samples were deeply sequenced at 6.04 ± 0.75 million reads per 10-cell pool to ensure saturation of gene detection and provide maximum sensitivity for identifying non-carcinoma contaminants. Across all cases, we found that the luminal markers *KRT8* and *ESR1* predominated by transcripts per million (TPM), whereas markers for myoepithelial cells (*KRT14*, *KRT5*), T cells (*CD3D*), B cells (*CD19*), and macrophages (*FGCR1A*) were rarely detected (**Figure 3.1C**). Even though desmoplasia was marked for all but one case (**Table 3.2**), fibroblast contamination was problematic in only ~14% of 10-cell samples (using $\log_2$ *FAP* > 5 transcripts per million [TPM] ≈ 0.8 copies per cell (180) as a stringent threshold). These samples were excluded retroactively for downstream analyses related to EMT signatures (see below). Overall, the observations confirmed the fidelity of (immuno-)LCM for isolating spatially-resolved, carcinoma-specific transcriptomic profiles with minimal disruption of tumor architecture.

Table 3.1 Transcriptomic studies of intra-carcinoma cell heterogeneity from primary clinical cases.

Cancer type	Median stage [range]	Dissociation time	Carcinoma yield	Read depth per cell	Reference
Head and neck	T3N1 [T1N0–T4aN2c]	60 min	38%	1340K	(125)
Colon	T3N2 [T1n0–T4bN0]	30 min	28%	≥100K	(264)
Pancreas	IIb [Ib–IV]	45 min	7.8%	50K	(265)
	T2N1 [T1cN0–T4N2]	40 min	6.3%	50K	(122)
Breast	T2N0 [T1N0–T3N1]	n.a.	62%	5800K	(117)
	T2N0 [T0N0–T2N3]	60 min	62%	50K	(118)
Luminal breast	T1N0 [T1N0-T2N1]	0 min	~100%	604K	This study

n.a., not available.

Table 3.2 Early stage luminal tumors profiled in this study.

Sample	Age	Race	Stage	Grade	ER	PR	HER2	Mitoses (per hpf)	TIL foci (per hpf)	Desmoplastic reaction
UVABC1	76	Caucasian	1	3	+	+	—	21/10	2/5	Marked
UVABC2	52	Caucasian	2	3	+	+	—	32/10	2/5	Moderate
UVABC3	63	Caucasian	1	3	+	+	—	11/10	5/5	Marked
UVABC4	63	Caucasian	1	2	+	+	—	2/10	1/5	Marked
UVABC5	59	Caucasian	1	1	+	+	—	2/10	3/5	Marked

TIL = tumor infiltrating lymphocytes
hpf = high powered field (400X magnification)

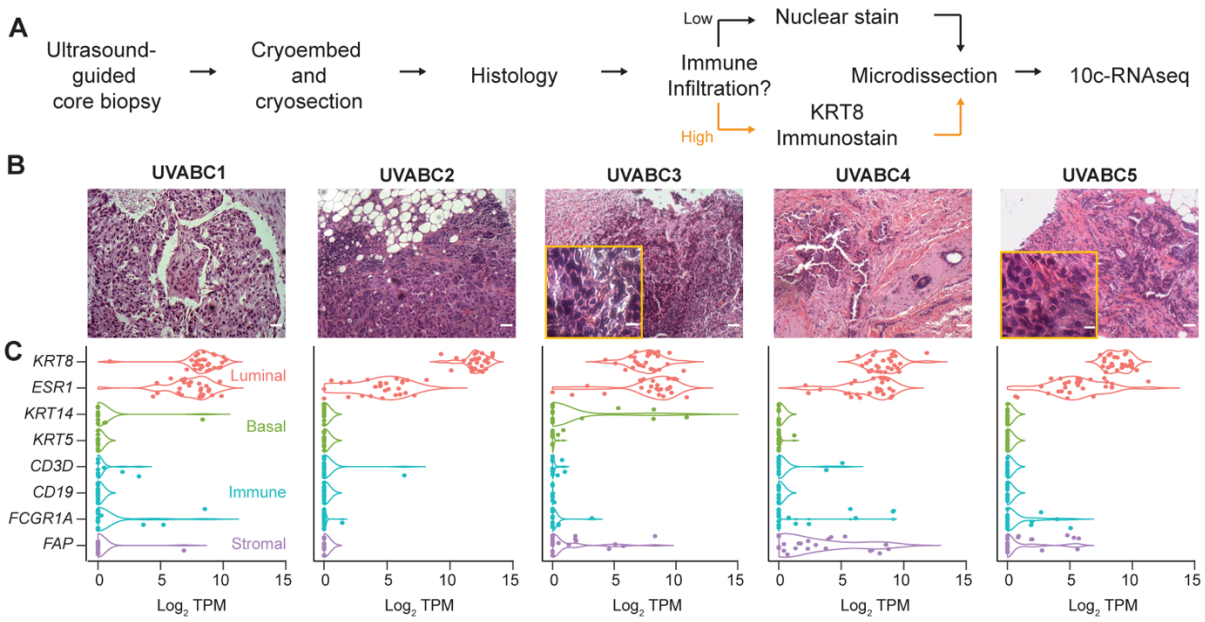

Figure 3.1 Focused transcriptional profiling of breast carcinoma cells without dissociation in early- stage tumor biopsies.
(A) Biopsy tissue was immediately cryoembedded and later cryosectioned for laser-capture microdissection. Tumor cells were visualized with either a rapid nuclear stain or KRT8 immunostain (**Figure 3.2**) before 10cRNA-seq (16).
(B) Tumor histology of the UVABC cohort visualized by hematoxylin-eosin staining. Two cases (marigold inset) showed increased tumor infiltrating lymphocytes requiring KRT8 immunostain-guided LCM.
(C) Selective capture of epithelial carcinoma cells assessed by marker transcripts for luminal (*KRT8*, *ESR1*), basal (*KRT14*, *KRT5*), immune (*CD3D*, *CD19*, *FCGR1A*), and stromal (*FAP*) cells. Scale bar is 80 μm (B) or 10 μm (B, inset)

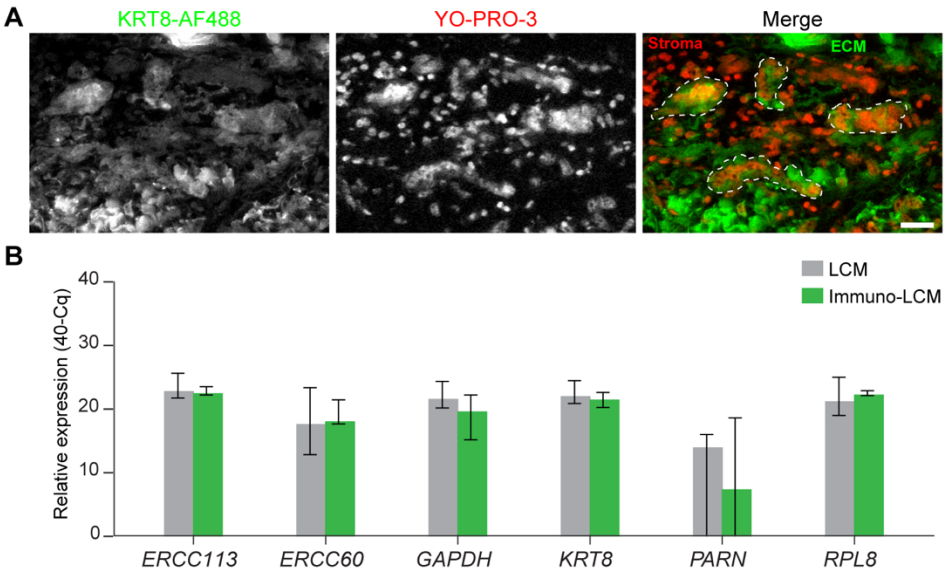

Figure 3.2 Immuno-LCM capture of 10-cell samples does not affect gene detection.
(A) Rapid immunostaining of UVABC3 with an Alexa Fluor 488 (AF488)-conjugated antibody recognizing KRT8 (left) combined with YO-PRO-3 to stain all nuclei (middle). Epithelial cells were identified by dual red-green staining (right, dashed lines) compared to stromal cells that are KRT8-negative and autofluorescent extracellular matrix that is free of nuclear staining. Scale bar is 25 µm.
(B) Relative abundance for the indicated transcripts as measured by quantitative PCR in UVABC4. Two exogenously spiked-in RNA transcripts (*ERCC113*, *ERCC60*) were quantified along with four endogenous genes. Data are shown as the median inverse quantification cycle (40–Cq) ± range from n = 4 replicates.

3.3.2 10cRNA-seq transcriptomes retain the inter-tumor and intra-tumor heterogeneity of luminal breast carcinoma cells profiled by scRNA-seq

As a first assessment, the 10cRNA-seq data were visualized by uniform manifold approximation and projection (UMAP) (266). Consistent with past descriptions of carcinoma heterogeneity by scRNA-seq (117,118,125), we found that 10cRNA-seq data clustered tightly by patient (**Figure 3.3A**). Batch effects were not evident within the patient clusters (**Figure 3.4A–E**). Furthermore, we did not observe any clustering in a separate UMAP visualization using only the data from ERCC spike-ins added to every sample at the time of RNA extraction (**Figure 3.4F**). The observed separation of 10cRNA-seq transcriptomes (**Figure 3.3A**) thus reflects bona fide inter-tumor differences between cases.

Next, we sought to classify individual 10cRNA-seq samples into intrinsic breast cancer subtypes (266). We adapted the microarray-based PAM50 classification of subtypes (12) to RNA-seq (268), but the 50-gene signature was not robust enough for 10cRNA-seq observations (**Figure 3.5**). As a substitute, we used expression-signature modules (269) associated with *ESR1*, *ERBB2*, and *AURKA* as proxies for the hormone, *HER2*, and proliferative status of each 10-cell sample (see Materials and Methods). Within patients, there was considerable variability in module scores (**Figure 3.3B**), corroborating an earlier scRNA-seq study of multiple breast-cancer subtypes (117). Although most 10-cell samples were classified as luminal A or luminal B subtype, all cases but UVABC4 contained observations scoring more strongly to other subtypes (**Figure 3.3C**). Two cases (UVABC1 and UVABC3) harbored instances of all four subtypes, analogous to scRNA-seq observations in glioblastoma (79). The clustering of mixed classifications implied that other patient-specific gene programs were dominant in the UMAP organization. Variations in subtype class were repeatedly documented in nearby samples microdissected from the same histologic section (**Figure 3.6**). Even for early-stage breast tumors, the 10cRNA-seq data suggested that local variations in regulatory state are pervasive.

Toward a more-direct comparison of 10-cell data with single-cell measurements of gene expression, we extracted 78 *KRT8*$^{+}$*EPCAM*$^{+}$ carcinoma cells from three cases of luminal breast cancer profiled by scRNA-seq (117) (see Materials and Methods). As previously reported by us and others (144,180,195,197), there were significantly more transcripts detected in local 10-cell pools compared to singly isolated cells (10,066 ± 1,416 genes vs. 5,957 ± 1,824 genes; $p < 10^{-15}$ by K- S test; **Figure 3.3D**). Overall, 3319 transcripts found in 10cRNA-seq pools were entirely undetected by scRNA-seq. Notwithstanding the differences in gene coverage, when scRNA-seq and 10cRNA-seq samples were projected on a shared UMAP, the separation between methods was comparable to that among patients (see Materials and Methods; **Figure 3.3E**). Together, the data argue that 10-cell pooling does not dilute out the cell-to-cell and tumor-to-tumor heterogeneities recognized by scRNA-seq.

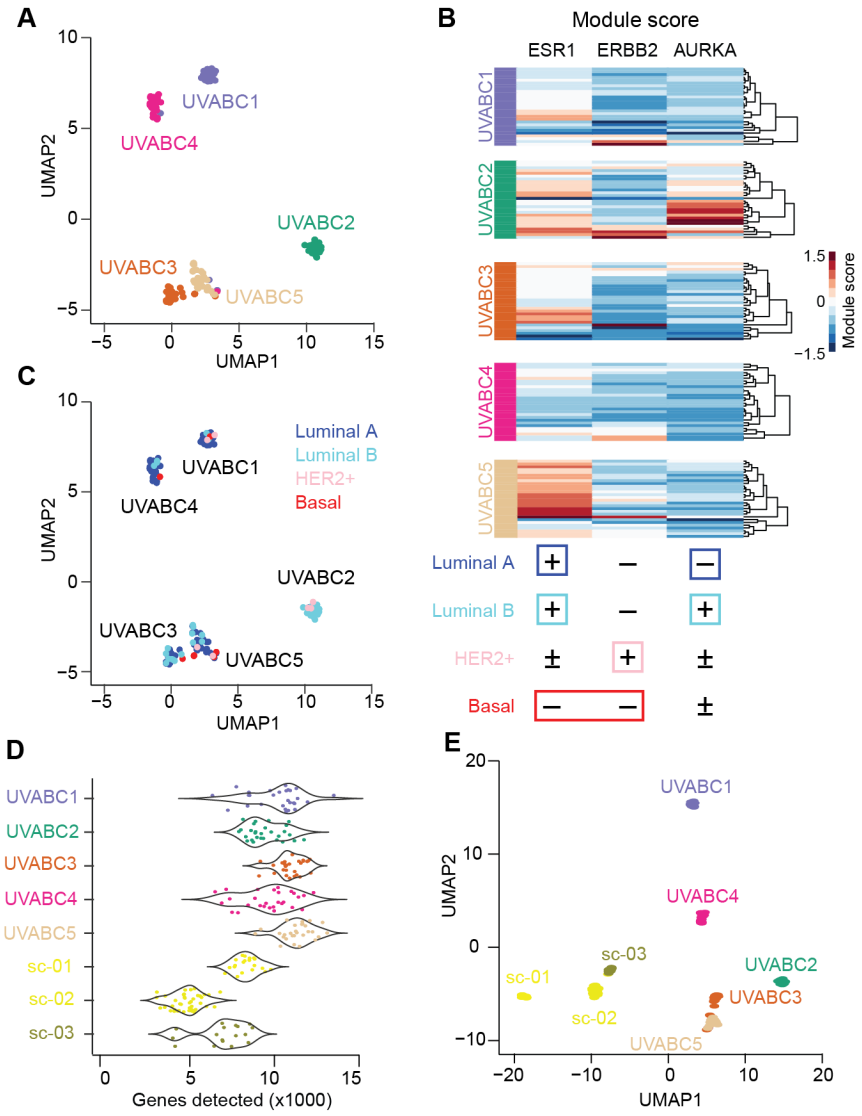

Figure 3.3 10-cell transcriptomes of luminal breast carcinomas are heterogeneous among and within tumors.
(A) UMAP embedding of 10cRNA-seq samples from the UVABC cohort colored by tumor.
(B) A three-signature classification system for identifying molecular subtypes of breast cancer in 10cRNA-seq data. Module scores were used to classify samples as indicated.
(C) Molecular subtype classifications of the 10cRNA-seq samples projected as in A.
(D) Genes detected by 10cRNA-seq in the UVABC cohort compared to three luminal tumors profiled by scRNA-seq (sc-01, sc-02, sc-03) (117). $p < 10^{-15}$ by K-S test.
(E) UMAP embedding of tumors profiled by 10cRNA-seq and scRNA-seq.

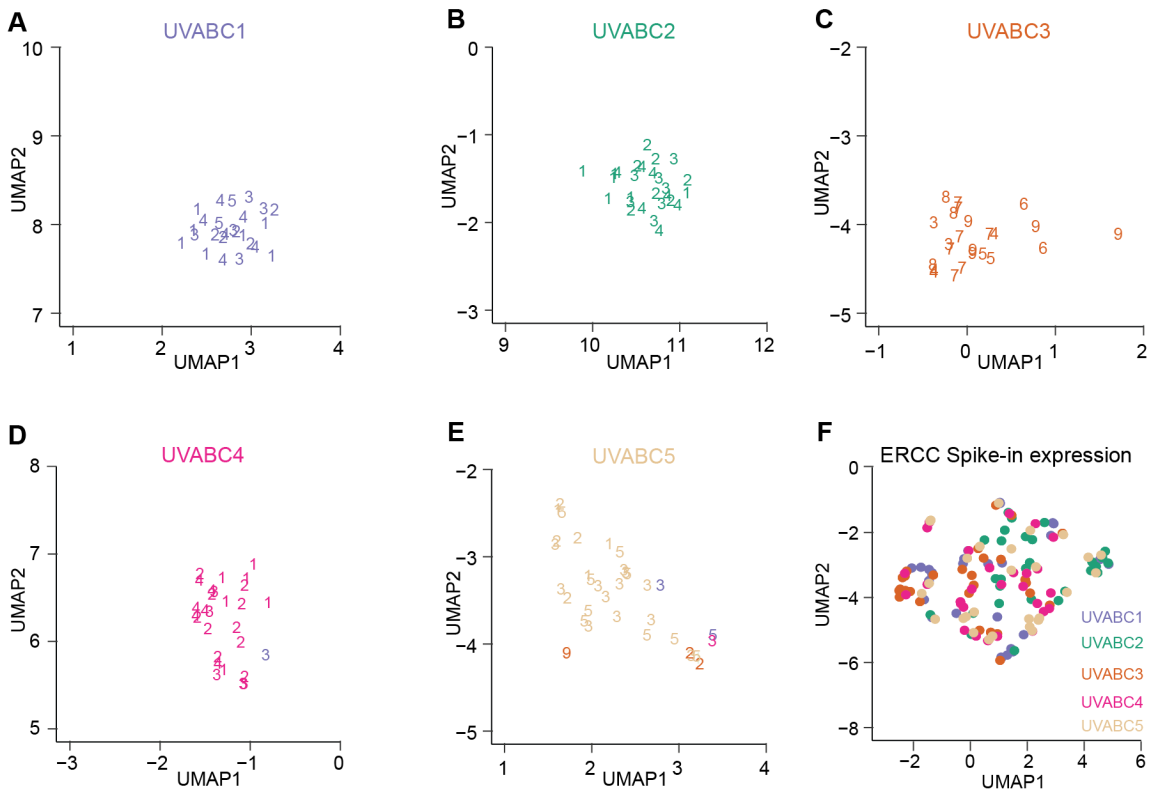

Figure 3.4 Clustering of 10cRNA-seq data by tumor does not arise from batch effects.
(A–E) Enlarged UMAP embedding for UVABC tumors from **Figure 3.3A**, with different batches of sample collections annotated by number.
(F) UMAP embedding of the UVABC cohort based only on the ERCC spike-in transcripts of 10cRNA-seq samples.

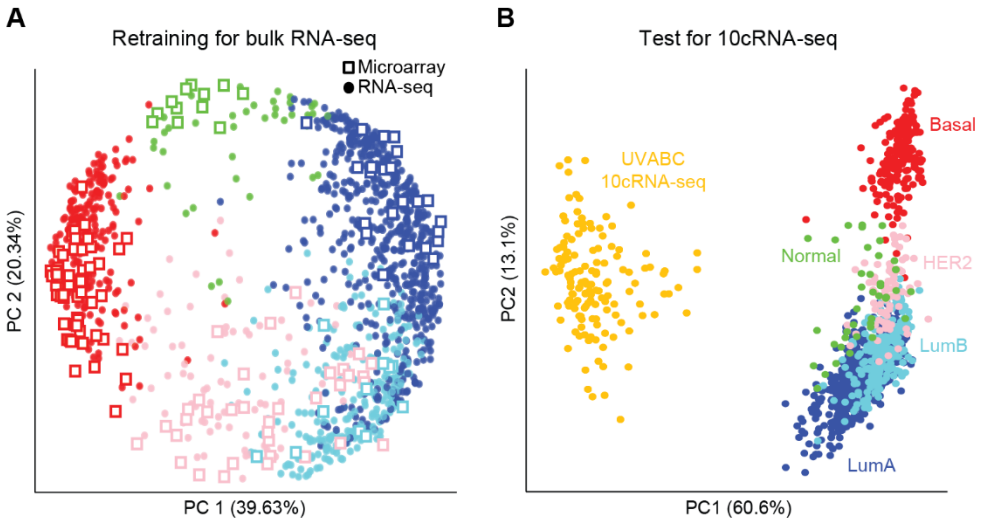

Figure 3.5 Microarray-based PAM50 classification is not adaptable to 10cRNA- seq.
(A) Principal component plot adapting the microarray-based PAM50 predictor (12) to bulk RNA-seq data of breast tumors from The Cancer Genome Atlas (TCGA) (268) (see Materials and Methods). Similar projections of the different data types indicate successful data fusion: 77% of RNA-seq samples were correctly subtyped, with an average confidence score of 0.99.
(B) Principal component plot of UVABC 10cRNA-seq observations together with TCGA tumors subtyped with the adapted PAM50 classification of A. 10cRNA-seq projections are distinct from all classified subtypes of TCGA tumors, even though dispersion along PC2 indicates underlying subtype differences.

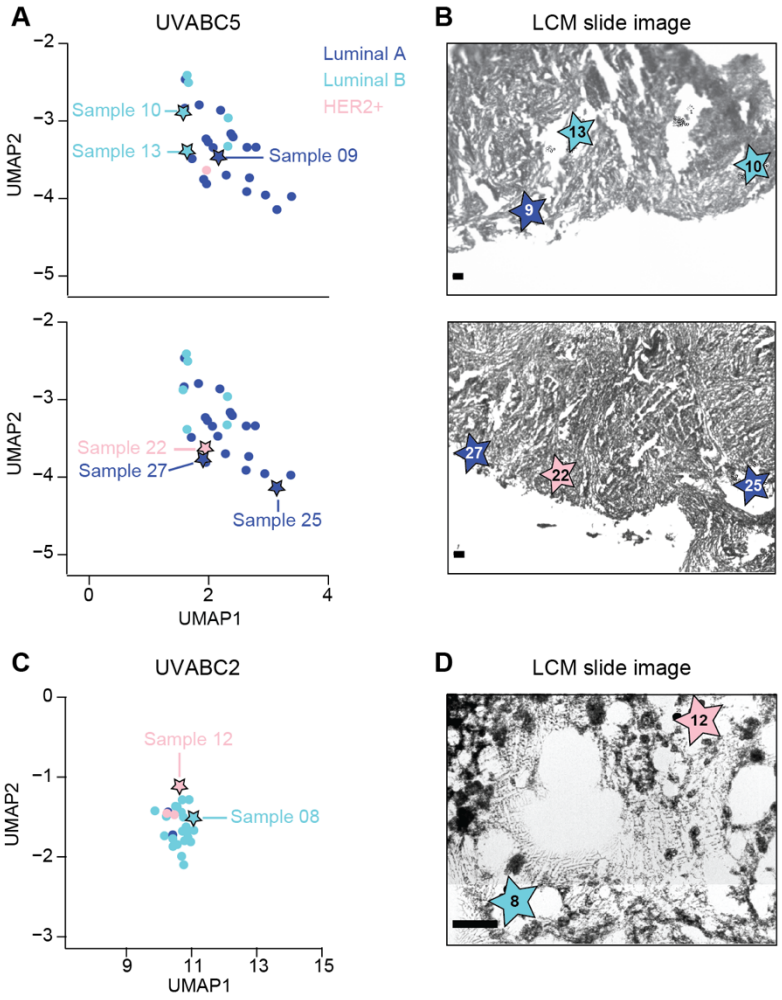

Figure 3.6 Different molecular subtypes are assigned to tumor cells microdissected from the same cryosection.
(A) Enlarged UMAP embedding for UVABC5 from **Figure 3.3C**, highlighting six 10cRNA-seq samples (stars) obtained together in two separate cryosections.
(B) Low-magnification grayscale LCM slide images indicating the regions microdissected in the UVABC5 cryosections.
(C) Enlarged UMAP embedding for UVABC2 from **Figure 3.3C**, highlighting two 10cRNA-seq samples (stars) obtained together in one cryosection.
(D) High-magnification grayscale LCM slide image indicating the regions microdissected in the UVABC2 cryosection. Scale bar is 50 μm (B and D).

3.3.3 Stochastic profiling by 10cRNA-seq identifies candidate regulatory heterogeneities

To go beyond qualitative descriptions of molecular subtype and inter-tumor differences, we sought to adapt the theory of stochastic profiling (129,157) to 10cRNA-seq (180). Fluctuation analysis by RNA-seq brings additional challenges and opportunities compared to microarray-based transcriptomics. The data are not biased by the position or quality of hybridization probes [as discussed in (180)], but they are sensitive to read depth, and low-abundance transcripts are susceptible to noise from counting statistics. Specifically, the discrete and left-censored character of rare transcripts partially obscures sample-to-sample fluctuations (**Figure 3.7A**) and deviates from lognormally distributed models used in earlier analyses (129,157,158). We surmounted these hurdles by extracting an abundance-dependent dispersion module from the SCDE package (153,154) and redeploying it as a separate inference tool for stochastic profiling. The module relates the squared coefficient of variation (CV) of each gene in a study (here, a patient) to the abundance magnitude of that gene, building an expectation model of variance at a given abundance (**Figure 3.7B**). The variance of each transcript is then normalized by the expected variance for that transcript's abundance, yielding an overdispersion score for the transcript. For high-abundance genes, overdispersed transcripts show multiple modes or heavier tails than expected (**Figure 3.7C, D**). Low-abundance genes with overdispersion are skewed by multiple instances of moderate-to-high TPM (**Figure 3.7E, F**). The dispersion module incorporates discrete negative-binomial and Poisson processes to model aligned reads and dropouts. The overdispersion score thus provides a principled metric for stochastic-profiling analysis of 10cRNA-seq data.

In conventional scRNA-seq, each cell is considered as an N-of-1 observation that convolves biological variability and technical noise. For our study, technical noise could be quantified more rigorously by pool-and-split sequencing of 10-cell equivalents from hundreds of carcinoma cells microdissected in the same vicinity as the samples. Accordingly, we sequenced

20 pool-and-split controls in parallel with the 28 10-cell samples, analyzing the controls separately to construct a null distribution for transcript overdispersion in each tumor. The upper 95th percentile in the null model defined an overdispersion cutoff for the 10-cell samples—transcripts above this threshold in the 10-cell samples (but not in the null) were considered candidate regulatory heterogeneities (**Figure 3.7G**, **H**, and **Figure 3.8**).

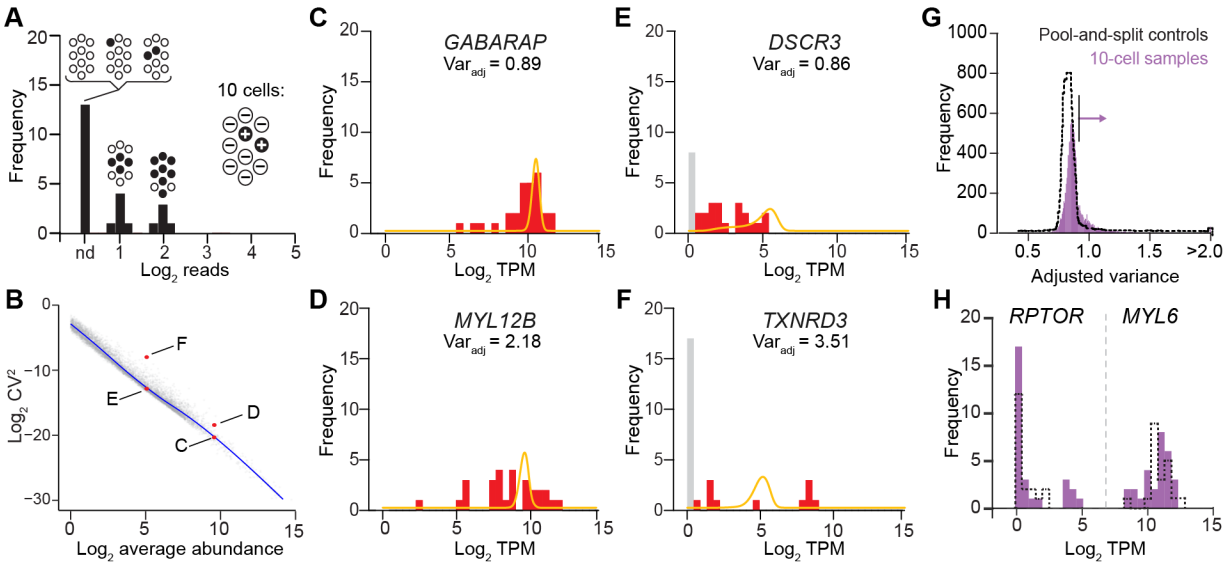

Figure 3.7 Stochastic profiling by 10cRNA-seq through abundance-dependent overdispersion statistics.

(A), Illustration of theoretical left-censored transcript, which is heterogeneously expressed at one copy per cell (filled, +) and measured with 50% efficiency. Low-frequency mixtures will be identically not detected (nd).

(B), Dispersion-abundance plot illustrating the expected inverse relationship. Example transcripts with similar abundance but different dispersion (red) are annotated by the **Figure 3.7** subpanel in which they appear. The blue trace indicates a smoothing cubic spline fit of the summarized 10cRNA-seq fluctuations per transcript (gray).

(C) and **(D),** 10cRNA-seq sampling fluctuations for high-abundance transcripts with expected dispersion (*GABARAP*, **C**) and overdispersion (*MYL12B*, **D**).

(E) and **(F),** 10cRNA-seq sampling fluctuations for low-abundance transcripts with expected dispersion (*DSCR3*, **E**) and overdispersion (*TXNRD3*, **F**).

(G), Distribution of adjusted variance scores for each gene measured transcriptomically in separate 10-cell samples (purple) compared to pool-and-split controls estimating technical variation (black dashed). The arrow indicates the upper 5th percentile of adjusted variance for the pool-and-split controls, which is used as the cutoff for 10-cell samples.

(H), Examples of low-abundance (*RPTOR*) and high-abundance (*MYL6*) transcripts deemed to be significantly overdisperse given their relative abundance in 10-cell samples (purple) and their technical variation in pool-and-split controls (black dashed). For **C–F**, continuous traces (marigold) indicate the idealized dispersion expected given the observed transcript abundance, and the adjusted variance (Var_{adj}) is reported. Gray bars represent drop-out events, which are modeled by a separate posterior (47,48). For **B–H**, data from UVABC4 were used as representative examples.

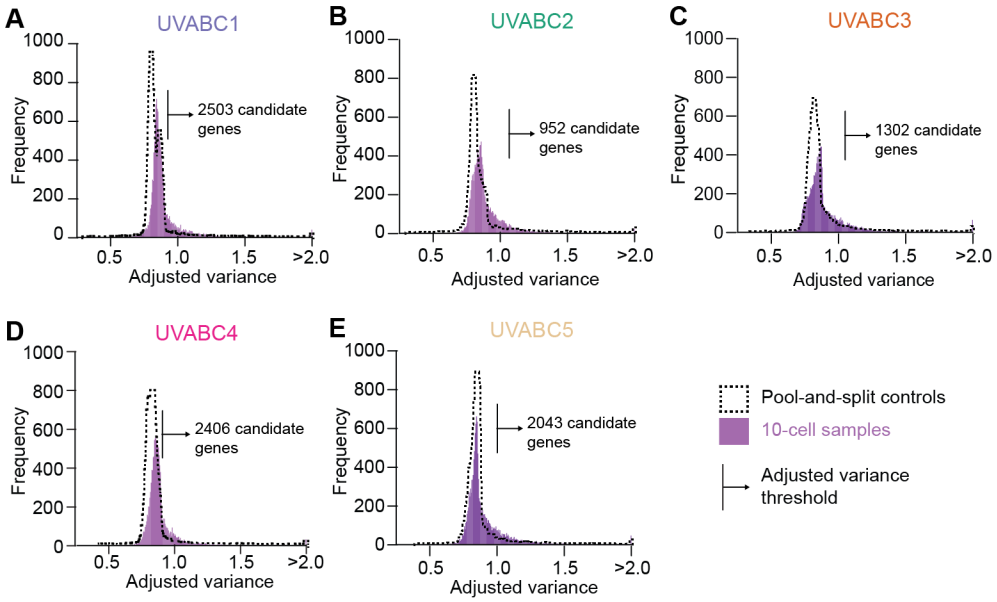

Figure 3.8 Abundance-dependent overdispersion statistics of individual cases in the UVABC cohort.

(A–E), Distribution of adjusted variance scores for each gene measured transcriptomically in separate 10-cell samples (purple) compared to pool-and-split controls estimating technical variation (black dashed). The arrow indicates the upper 5th percentile of adjusted variance for the pool-and-split controls, which is used as the cutoff for 10-cell samples.

3.3.4 Stochastic profiling identifies recurrent transcriptional regulatory heterogeneities

By abundance-dependent dispersion, stochastic profiling identified 9206 candidate heterogeneities in the UVABC cohort, 161 transcripts of which were undetected by scRNA-seq (117). Only a few percent of candidate genes were found in loci of inferred copy-number variation (**Figure 3.9**), excluding major contributions from heritable differences among subclones in a tumor. 3627 candidates were exclusive to one of five patients, laying bare the extraordinary challenge of interpreting malignant cell-state heterogeneity beyond well- known markers of differentiation (124). Encouragingly, when the candidates were intersected, we noted a significant enrichment of transcripts that appeared repeatedly in three or more breast-cancer cases (**Figure 3.10A**). Such transcripts were classified as recurrent heterogeneously expressed genes (RHEGs), for which there were 710 in total. Generalizing the RHEG definition to candidates appearing in >50% of the cases considered, we examined the stability of RHEG numbers by subsampling the UVABC cases (see Materials and Methods). As the quantity of patients increased, RHEGs stabilized in the range of 500–1500 for this highly circumscribed cohort (**Figure 3.10B** and **Table 3.2**). RHEGs provide a conceptual framework for prioritizing cell-state regulatory heterogeneities identified in vivo (see Chapter 4) (270,271).

To evaluate RHEGs as an organizing principle, we revisited the UMAP visualization of UVABC cases from the standpoint of regulatory heterogeneity (**Figure 3.3A**). Because abundance-dependent dispersion evaluates fluctuations local to each tumor (**Figure 3.7**), it was first necessary to standardize the 10cRNA-seq transcriptomes separately and regenerate the UMAP (see Materials and Methods). Tumor-specific standardization intermingled the 10cRNA-seq observations considerably, but two clusters remained enriched in samples from UVABC1 and UVABC2 (**Figure 3.10C**). When the same approach was applied using RHEGs exclusively, we observed a projection that was different from when the whole 10cRNA-seq transcriptome was used (**Figure 3.10D**). For the same UMAP parameters, samples were more distributed than

clustered, with the UVABC1-enriched cluster disappearing and a new UVABC4-enriched cluster appearing. This analysis suggested that RHEGs could be used as a lens to refocus transcriptome-wide heterogeneity on the most-robust variations.

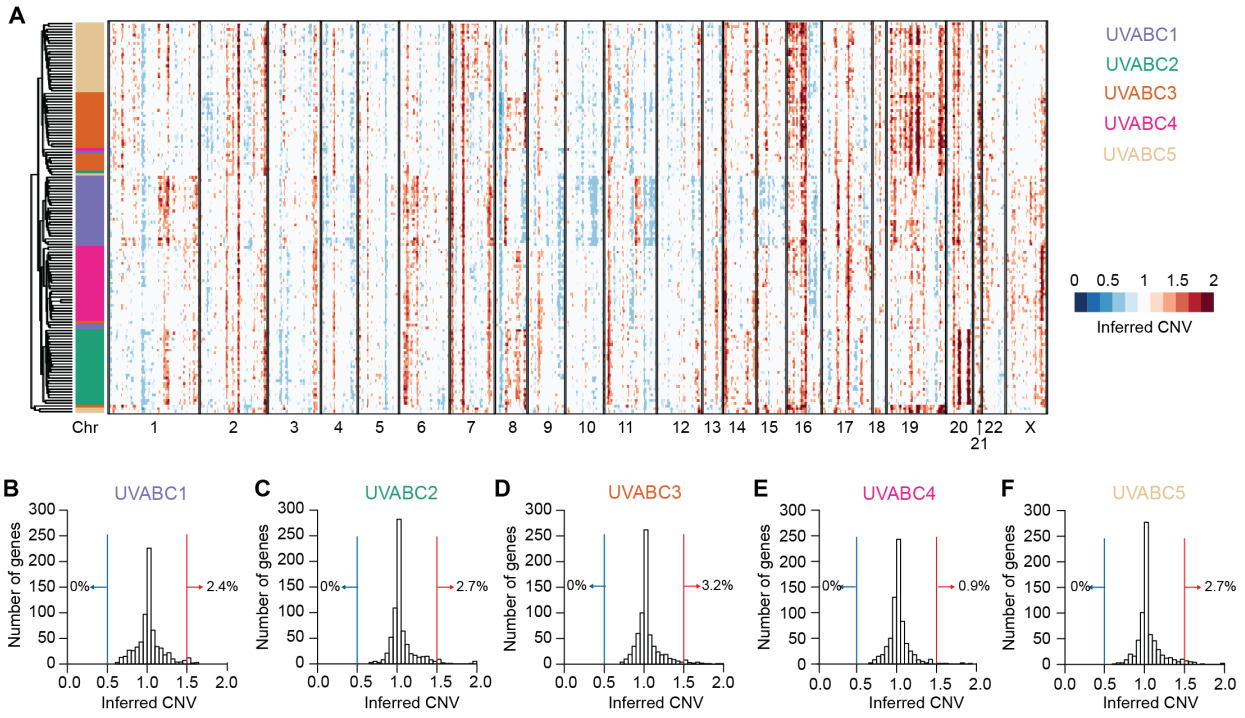

Figure 3.9 Most candidate heterogeneities do not reside in loci with inferred copy-number variations (CNVs).
(A) Chromosomal gains and losses predicted from 10cRNA-seq data by inferCNV (30). RNA-seq data from normal human luminal breast tissue obtained through GTEx (29) was used as the reference transcriptome. Gains in 1q (UVABC1, UVABC2) and 8q (UVABC2, UVABC4) and losses in 8p (all) and 16q (UVABC1, UVABC4) are characteristic of luminal A breast tumors (64). **(B–F)** Distribution of inferred CNVs corresponding to the candidate heterogeneities of individual cases in the UVABC cohort. Percentage of transcripts with inferCNV scores suggesting gain (red) or loss (blue) is shown.

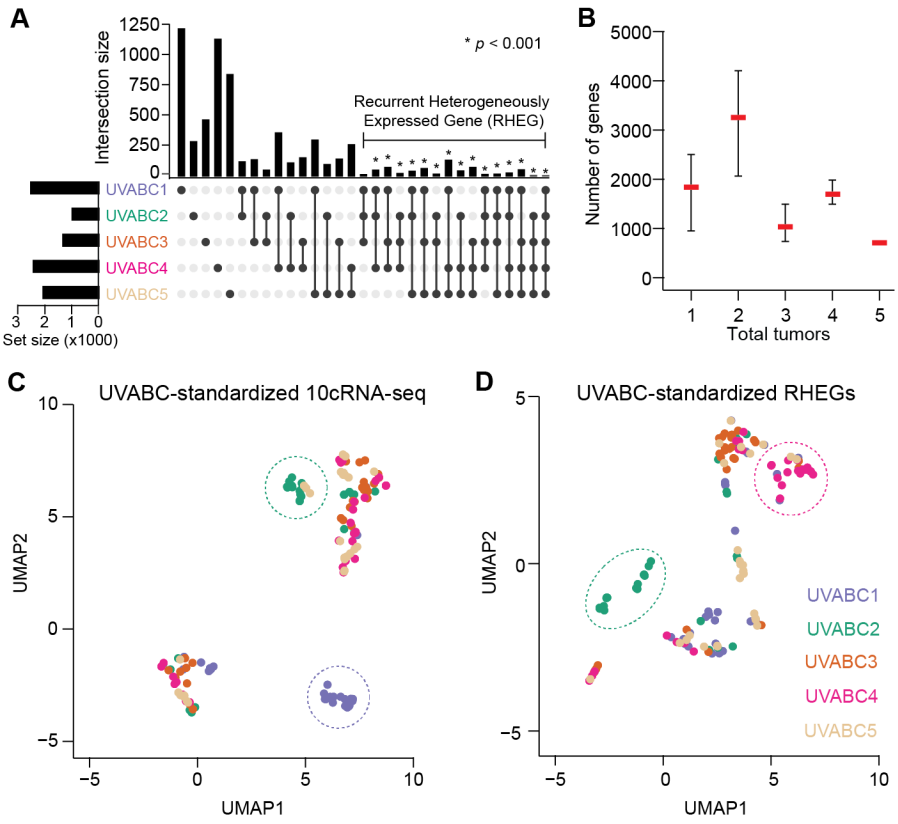

Figure 3.10 Intersecting candidate regulatory heterogeneities across tumors yields a set of recurrent heterogeneously expressed genes (RHEGs).
(A) Gene overlaps between and among cases in the UVABC cohort. Enriched categories were assessed statistically by Monte-Carlo simulation (see Materials and Methods).
(B), RHEG size as a function of the number of UVABC tumors included. RHEGs were defined as transcripts that appear in >50% of the tumors included. Data are shown as the median ± range of ${}_5C_n$ combinations of *n* tumors in the UVABC cohort.
(C) and (D), UMAP embedding of patient-standardized 10cRNA-seq transcriptomes (**C**) or RHEGs (**D**). Patient- enriched clusters are highlighted with dashed ovals.

3.3.5 RHEGs are not dominated by cell-cycle covariates

In scRNA-seq, the most-overarching contributor to heterogeneity is the phase of the cell cycle (125,246,272). However, it was not obvious whether such single-cell variations would also overrepresent in RHEGs derived from 10-cell pools. Using a panel of 863 transcripts associated with replicating cells (272), we identified 62 among the 710 RHEGs, a significant overlap ($p < 10^{-6}$ by hypergeometric test) but one comprising <10% of the list overall. Most of the overlapping genes were expressed acutely during one cell-cycle transition (e.g., G1/S, G2/M), which is akin to the two-state expression models foundational for the theory of stochastic profiling (129,157,158). When the cell-cycle search was restricted to 361 oscillating transcripts (273), the RHEG intersection reduced to 24 genes ($p < 0.01$ by hypergeometric test). Moreover, when we compared the periodicity of the overlapping genes to the most-symmetrically cycling transcripts, we found that RHEGs were significantly more biased toward up- or downregulation (see Materials and Methods; **Figure 3.11A, B**). These results are consistent with Monte-Carlo simulations of stochastic profiling that model a three-state distribution corresponding to G1, S, and G2/M populations (**Figure 3.11C–F**). Stochastic-profiling theory thus bolstered our experimental results indicating that cycling transcripts contribute <5% to the RHEGs identified.

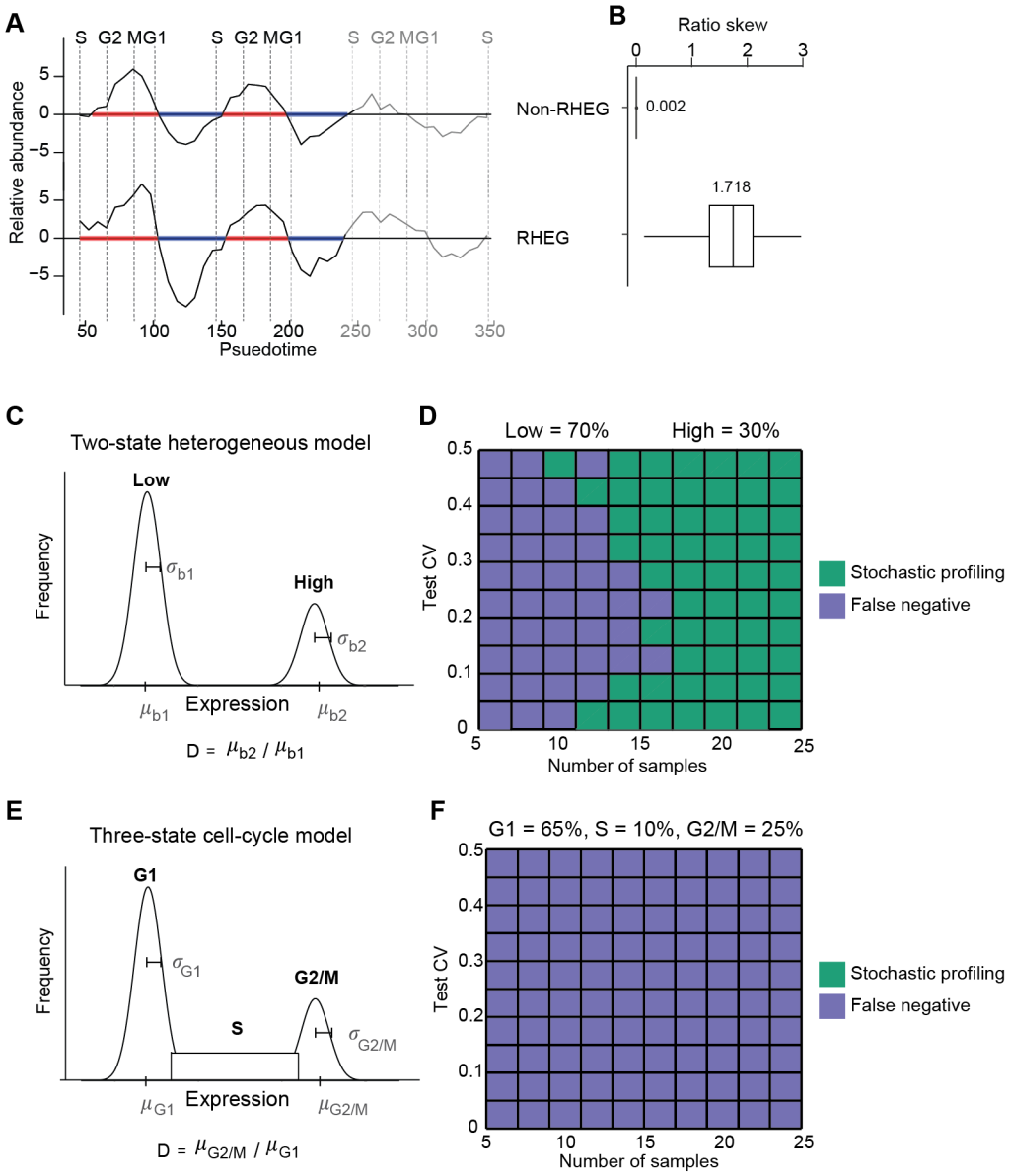
A
S G2 MG1 S G2 MG1 S G2 MG1 S
Relative abundance
5
0
−5
5
0
−5
50 100 150 200 250 300 350
Psuedotime
B
Ratio skew
0 1 2 3
Non-RHEG
0.002
1.718
RHEG
C
Two-state heterogeneous model
Low
High
Frequency
σ_{b1}
σ_{b2}
μ_{b1}
Expression
μ_{b2}
D = μ_{b2} / μ_{b1}
D
Low = 70%
High = 30%
Test CV
0.5
0.4
0.3
0.2
0.1
0
5 10 15 20 25
Number of samples
Stochastic profiling
False negative
E
Three-state cell-cycle model
G1
G2/M
S
Frequency
σ_{G1}
$\sigma_{G2/M}$
μ_{G1}
Expression
$\mu_{G2/M}$
D = $\mu_{G2/M}$ / μ_{G1}
F
G1 = 65%, S = 10%, G2/M = 25%
Test CV
0.5
0.4
0.3
0.2
0.1
0
5 10 15 20 25
Number of samples
Stochastic profiling
False negative

Figure 3.11 Periodically cycling transcripts are disfavored by stochastic profiling.
(A) Asymmetric cell-cycle oscillations in the RHEG *DCTPP1* compared to the non-RHEG *HJURP*. Microarray data from synchronized HeLa cells (77) and pseudotime estimates were obtained from Cyclebase 3.0 (273). Time intervals above (red) and time (blue) the midpoint of each transcript are shown.
(B) Asymmetry of cycling RHEGs quantified by ratio skew and compared to non- RHEG cycling transcripts (see Materials and Methods). Data are shown as boxplots from n = 1000 bootstrapping runs.
(C) Abstraction of a two-state regulatory heterogeneity. Parameters of the probability distribution are described elsewhere (129,157,158).
(D) Monte-Carlo simulations (157) of stochastic profiling in the two-state case. False-negative regimes are marked when a two-state heterogeneity is not detected in the 10-cell pool.
(E) Abstraction of a three-state cell-cycle model. A uniform S-phase interval is added in between the first and second regulatory states modeling G1 and G2/M phases, and the fractional proportions are updated accordingly (see Materials and Methods).
(F) Monte-Carlo simulation of stochastic profiling in the three-state case. The false- negative regime marks three-state heterogeneities that are not detected in the 10-cell pool. For D and F, the following simulation parameters were used: D = 3, $\sigma_{b1}=\sigma_{b2}=\sigma_{G1}=\sigma_{G2}=0.2$.

3.3.6 RHEGs are largely devoid of detachment artifacts and influence from breast-cancer driver genes

We also considered other trivial explanations for genes in the RHEG set. Although tumors were not dissociated, it was possible that detachment-like regulatory variation was induced locally and rapidly from the tissue damage of the biopsy procedure. Using a 138-gene signature for dissociation-induced transcripts in muscle satellite cells (143), we intersected with the RHEG set and found that only two were shared (**Figure 3.12A**). This marginal under-enrichment ($1 - p < 0.05$) indicated that our clinical-procurement and sample-handling procedures had avoided detachment-like damage responses in the breast carcinoma cells.

Next, we looked at known breast cancer drivers with the premise that mutations may arise subclonally in a breast carcinoma (274) and disrupt abundance of the encoded transcript (275). Among 29 robust driver genes for breast cancer (260), only one was shared with the RHEG set: *CTCF* (**Figure 3.12B**, left). As an insulator protein, CTCF abundance changes could cause secondary transcriptional alterations; however, we did not observe any enrichment for conserved CTCF-sensitive genes (276) in the RHEG set (**Figure 3.12B**, right). Although not classified as a RHEG, we also investigated transcriptional targets for the most-prevalent transcription factor mutated in luminal breast cancer, GATA3 (260). Again, we found no RHEG enrichment among 1213 transcripts altered by mutant GATA3 in luminal breast cancer cells (277) (**Figure 3.12C**). The lack of association collectively supported that RHEGs were more than a reflection of known sources of cell-to-cell heterogeneity in cancer.

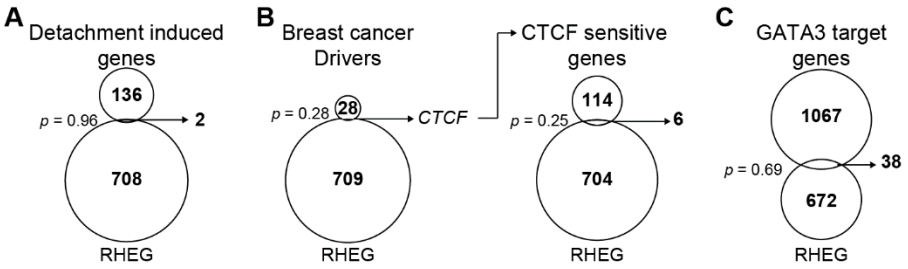

Figure 3.12 RHEGs have little in common with detachment signatures or mutational drivers of breast cancer.
(A–C), Venn diagrams intersecting the UVABC RHEG set with a cell-detachment signature (143) **(A),** a set of robust drivers for breast cancer (260) (**B,** left), a list of transcripts altered by CTCF knockout in a luminal breast cancer cell line (276) (**B**, right), and a list of GATA3 target genes (277) **(C).**
Statistical significance of overlaps was assessed by the hypergeometric test.

3.3.7 RHEGs are enriched for EMT signatures and correlate with canonical EMT markers

scRNA-seq of dissociated tumors has identified (partial-)EMT states in some carcinomas (118,125) but not others (264). EMT-like transcriptional profiles also arise normally in the cell type of origin for serous ovarian cancer (278). For breast cancer, changes along the EMT spectrum are mostly described in hormone-negative cell lines, but more-recent work reports EMT-like activation patterns in 65–85% of primary luminal breast cancers (279). The literature thus supported a focused search for EMT states among RHEGs.

We intersected a pan-cancer EMT signature (280) with the RHEG set and found significant overlap in multiple collagens, matricellular proteins, and other transcripts in the signature (**Figure 3.13A**). The data suggest that ECM dysregulation in these tumors is jointly mediated by the carcinoma cells together with cancer-associated fibroblasts. RHEG enrichment was also found with an independently derived EMT signature (261) (**Figure 3.14A**), reinforcing the result. Formally, none of the canonical EMT regulators [*ZEB1* (M), *ZEB2* (M)] or markers [*CDH1* (epithelial, E), *VIM* (mesenchymal, M), *FN1* (M)] were RHEGs, even though all were detected reliably enough for stochastic-profiling analysis. We clustered these canonical EMT transcripts with the EMT RHEGs after stringently removing samples with any evidence of fibroblast contamination (14% of samples with *FAP* > 5 TPM ≈ 8 copies in one cell of the 10-cell pool). There was clear separation of E- and M-associated transcripts among 10-cell pools along with several notable subclusters by gene and by sample (**Figure 3.13B**). The organization by patient was unexpected; for example, UVABC2 showed the most evidence for the E state, even though it was one of the most-advanced stage tumors of the cohort (**Table 3.2**). Reciprocally, 10-cell profiles of the high-grade UVABC3 tumor were no more scattershot in EMT transcripts than UVABC4 (grade 2) or UVABC5 (grade 1) (**Figure 3.13B**). Among samples with M characteristics, *ZEB2* appeared to track with those samples abundant for some transcripts (*FN1*, *COL6A1*, *SPARC, VIM*) but not others (*COL5A2*, TAGLN). M-state fragmentation was

also observed in UVABC1, which was predominated by samples positive for *VIM*, *SPARC*, and *SERPING1* but negative for CTSK, TAGLN, and mesenchymal collagens (**Figure 3.13B** and **Figure 3.14B**). The RHEG set thereby provided a transcriptomic resource for looking more deeply at EMT regulatory patterns in early-stage luminal breast cancers.

3.3.8 RHEGs are enriched for pan-cancer driver genes and suggest transcription factor-target relationships in single cells

The dearth of breast-cancer driver genes among RHEGs (**Figure 3.12B**) prompted us to look at cancer drivers more broadly. We merged 299 robust drivers for any cancer type (260) with the latest pan-cancer analysis reporting 803 drivers from 2,658 tumors (281) and intersected with the RHEG set. There were multiple instances where RHEGs resided in the same complex, pathway, or gene subfamily as a cancer driver (**Table 3.3**). We included these proximal RHEGs and altogether found 46 genes as “RHEG drivers” shared between the two datasets (p = 0.001 by hypergeometric test; **Figure 3.13C**). Even with the expanded driver set, we found no enrichment for mutated breast-cancer driver genes (p = 0.6 by hypergeometric test). In the UVABC cohort, RHEG drivers may be leveraged noncanonically through transcriptional regulation rather than mutation.

Last, we clustered the RHEG drivers to ask whether there were interpretable covariations spanning multiple patients (**Figure 3.13D**). Associations among 10-cell samples were tightest for UVABC2 and UVABC4, in line with their separation on the earlier UMAP (**Figure 3.10D**). Repeatedly, co-clustering RHEG drivers suggested direct modes of action between transcription factors and target genes (**Figure 3.13D**, arrows). For example, knockdown of *NFATC4* blocks induction of the neighboring RHEG driver, *TNFSF10* (282), and there is literature that the reprogramming factor *KLF4* is required for full induction of *CDKN1A* (283). Although no functional studies are available for *CDKN2D*, another co-clustering cyclin-dependent kinase inhibitor (**Figure 3.13D**, arrows), the *CDKN2D* locus is occupied by KLF4

(284) and may warrant further study. Likewise, the *MAD1L1* locus is reportedly among the top 250 binding events in the genome for *TP73* (285)—a RHEG driver absent from all luminal breast cancer cells profiled by scRNA-seq (117). Some of the debate involving *MAD1L1* as a TP53 target gene [reviewed in (286)] could be explained by compensation from TP73 (287). RHEG drivers are variably expressed within luminal breast cancers, and our data suggest that some are variably active.

Table 3.3 Multiple RHEGs proximal to established cancer driver genes

RHEG	Driver gene	Cancer type(s)	Proximal relationship
EFNA4	EPHA4	LUAD	EPHA4 signals through EFNA4
GDF15	RET	PANCAN, THCA, SKCM	RET is a coreceptor for GDF15
NQO1	NFE2L2	PANCAN, LUSC, LIHC	NFE2L2 activity is marked by *NQO1* abundance
WNT4	WNT5A	PRAD	WNT5A and WNT4 are both non-canonical Wnt ligands
RRAS	KRAS	PANCAN, COAD, LUAD, PAAD, UCEC, ESCA	KRAS and RRAS are both in the Ras family
TP73	TP53	PANCAN, BLCA, BRCA, GBM, COAD, ESCA, HNSC, KIRC, LIHC, LUAD, LUSC, DLBC, OV, PAAD, PRAD, SKCM, STAD, UCEC, SARC	TP53 and TP73 are in the same family of transcription factors
IKBKG	IKBKB	DLBC	IKBKB and IKBKG are in the same IKK complex
CDKN2D	CDKN2A	PANCAN, HNSC, ESCA, LUSC, PAAD, SKCM	CDKN2A and CDKN2D are related CDK inhibitors
MLST8	MTOR	KIRC	MTOR and MLST8 are in the same MTORC complex

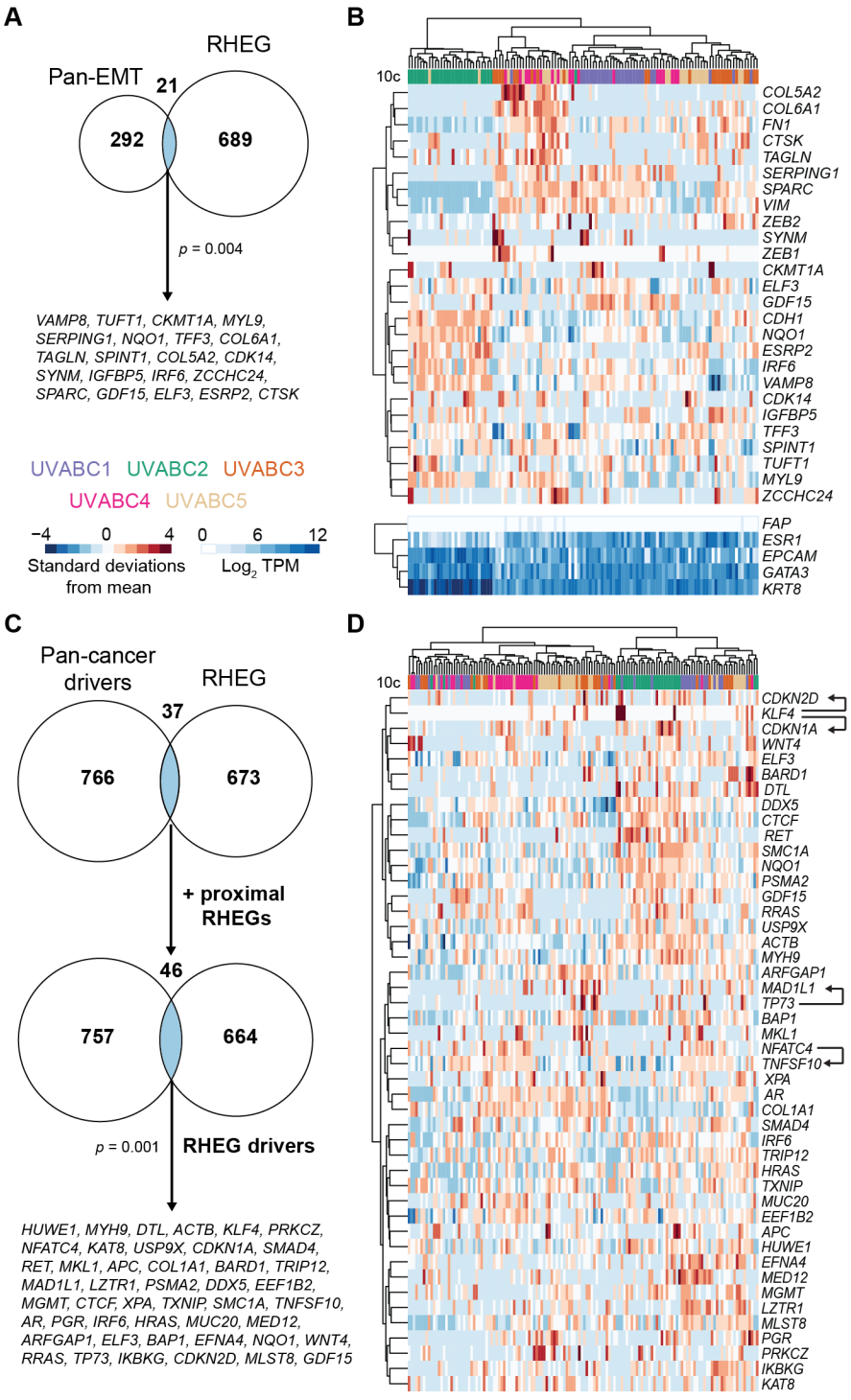
A
RHEG
Pan-EMT
21
292
689
p = 0.004
VAMP8, TUFT1, CKMT1A, MYL9, SERPING1, NQO1, TFF3, COL6A1, TAGLN, SPINT1, COL5A2, CDK14, SYNM, IGFBP5, IRF6, ZCCHC24, SPARC, GDF15, ELF3, ESRP2, CTSK
UVABC1 UVABC2 UVABC3
UVABC4 UVABC5
−4 0 4
Standard deviations from mean
0 6 12
Log2 TPM
B
10c
COL5A2
COL6A1
FN1
CTSK
TAGLN
SERPING1
SPARC
VIM
ZEB2
SYNM
ZEB1
CKMT1A
ELF3
GDF15
CDH1
NQO1
ESRP2
IRF6
VAMP8
CDK14
IGFBP5
TFF3
SPINT1
TUFT1
MYL9
ZCCHC24
FAP
ESR1
EPCAM
GATA3
KRT8
C
Pan-cancer drivers
RHEG
37
766
673
+ proximal RHEGs
46
757
664
p = 0.001
RHEG drivers
HUWE1, MYH9, DTL, ACTB, KLF4, PRKCZ, NFATC4, KAT8, USP9X, CDKN1A, SMAD4, RET, MKL1, APC, COL1A1, BARD1, TRIP12, MAD1L1, LZTR1, PSMA2, DDX5, EEF1B2, MGMT, CTCF, XPA, TXNIP, SMC1A, TNFSF10, AR, PGR, IRF6, HRAS, MUC20, MED12, ARFGAP1, ELF3, BAP1, EFNA4, NQO1, WNT4, RRAS, TP73, IKBKG, CDKN2D, MLST8, GDF15
D
10c
CDKN2D
KLF4
CDKN1A
WNT4
ELF3
BARD1
DTL
DDX5
CTCF
RET
SMC1A
NQO1
PSMA2
GDF15
RRAS
USP9X
ACTB
MYH9
ARFGAP1
MAD1L1
TP73
BAP1
MKL1
NFATC4
TNFSF10
XPA
AR
COL1A1
SMAD4
IRF6
TRIP12
HRAS
TXNIP
MUC20
EEF1B2
APC
HUWE1
EFNA4
MED12
MGMT
LZTR1
MLST8
PGR
PRKCZ
IKBKG
KAT8

Figure 3.13 RHEGs contain epithelial-to-mesenchymal transition (EMT) markers and driver genes for cancers other than breast.
(A) Venn diagram intersecting the UVABC RHEG set with a pan- cancer EMT signature (280). Shared genes are listed.
(B) Hierarchical clustering of the shared genes in A along with epithelial (*CDH1*) and mesenchymal (*VIM*, *FN1*, *ZEB1*, *ZEB2*) markers that were reliably detected by 10cRNA-seq. Stromal contamination was excluded by the relative abundance of the fibroblast marker *FAP* compared to the luminal markers *ESR1*, *EPCAM*, *GATA3*, and *KRT8*.
(C) Venn diagram intersecting the UVABC RHEG set with a pan-cancer set of driver genes (260,281). The intersection was updated to include proximal RHEGs as described in **Table 3.3.** Shared genes ("RHEG drivers") are listed.
(D) Hierarchical clustering of RHEG drivers. Arrows between co-clustering drivers indicate possibly direct transcription factor–target gene relationships as described in the text. For A and C, statistical significance of overlaps was assessed by the hypergeometric test.

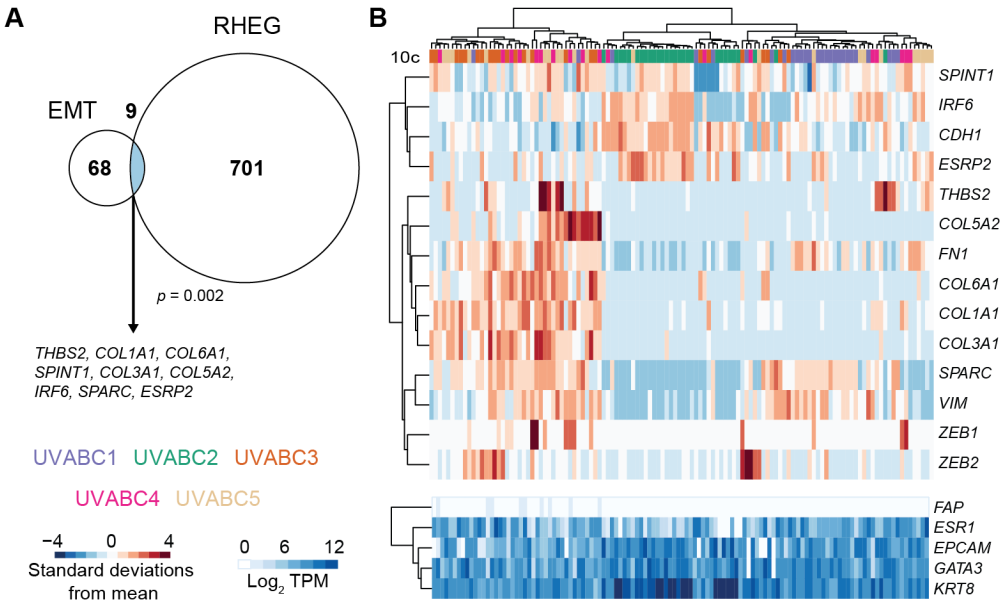

Figure 3.14. RHEGs are enriched for additional epithelial-to-mesenchymal transition (EMT) signatures.
(A) Venn diagram intersecting the UVABC RHEG set with an independently derived EMT signature (261) from that of **Figure 3.13A**. Shared genes are listed.
(B) Hierarchical clustering of the shared genes in A along with epithelial (*CDH1*) and mesenchymal (*VIM*, *FN1*, *ZEB1*, *ZEB2*) markers that were reliably detected by 10cRNA-seq. Stromal contamination was excluded by the relative abundance of the fibroblast marker *FAP* compared to the luminal markers *ESR1*, *EPCAM*, *GATA3*, and *KRT8*.

3.4 Discussion

This work combines 10cRNA-seq (180) and stochastic profiling (129) for disruption-free isolation of cancer cell-regulatory heterogeneities in a clinically practicable way. We targeted LCM isolation to breast-carcinoma cells here by using nuclear cytology or epithelial-targeted antibodies; the approach is also compatible with genetically encoded fluorophores (Chapter 4). For 10cRNA-seq, artifactual cell stress (143) is avoided by LCM (Chapter 4), and dominant cycling transcripts are mitigated by 10-cell averaging. But in many respects, 10cRNA-seq of malignant cells shares similarities with scRNA-seq: cases are very different from one another, and samples vary substantially within cases. What differs is overall gene coverage per sample (10cRNA-seq > scRNA-seq), as well as the analytical approach needed to discern single-cell differences. Abundance-dependent overdispersion can identify candidates from 10cRNA-seq, much like the nonparametric distribution tests first deployed for microarray data (157). As with microarrays, we anticipate future developments toward parameterizing the underlying single-cell distributions, which combine to yield 10cRNA-seq observations (158). The 10cRNA-seq-based subtype classifications predicted local differences not obvious from histology, and tools for spatial analysis of biomolecules are rapidly advancing (288,289). It will be especially intriguing when spatial transcriptomics (290) reaches the resolution and sensitivity of 10 cells.

RHEGs open the possibility of making more specific claims about intratumor heterogeneity beyond cell stress, cell cycle, and cell type (124). Partial EMTs in carcinomas have been documented by scRNA-seq (117,125), which we verify here in earlier-stage tumors without any pre-dissociation (**Table 3.1**). While there are many ways to elicit EMT-like states, a leading explanation for the UVABC cohort is tissue stiffness (99) given their marked desmoplasia. Notably, one RHEG is the hemidesmosomal integrin *ITGB4*, which acts as a critical sensor for matrix stiffness in breast epithelia (26). *ITGB4* was undetected in every luminal breast cancer cell analyzed by conventional scRNA-seq (117).

Secreted ligands and receptors (291) are not overly prevalent among RHEG drivers, but a pair with some coordination is the p53 target gene *GDF15* (292) and its cognate receptor *RET* (293) (**Figure 3.13D**). In another carcinoma type in Chapter 4, we suggest that such receptor-ligand pairs could engage as locally varying autocrine–paracrine circuits within a tumor and shape the immune microenvironment. For instance, among the 136 candidates shared by the two cases requiring immuno-LCM because of extensive lymphocytic infiltration, we identified an inhibitory ligand for NK cells [*ADGRG1* (294)], a major histocompatibility class II receptor [*HLA-DRA* (295)], a macrophage stimulatory ligand [*MST1* (296)], and the palmitoyltransferase for PD-L1 [*ZDHHC9* (297)]. Natural variation in such carcinoma transcripts could one day be mapped to associating changes in the type and extent of immune-cell recruitment (298).

Single-cell cancer biology must trade off coverage, throughput, and handling artifacts to retain the conceptual allure of measuring one cell. The approaches described and implemented here for late-onset, early-stage breast cancer are also a compromise, but one that triangulates differently by using cell pools to reduce handling and improve coverage. We see great potential in using stochastic profiling by 10cRNA-seq to deconstruct the very earliest stages of tumor initiation and premalignancy in engineered systems (271) and in precancerous *in situ* lesions of the breast where the need for treatment is actively debated (Chapter 5) (299).

3.5 Materials and methods

3.5.1 Tissue procurement and processing

Human sample acquisition and experimental procedures were carried out in compliance with regulations and protocols approved by the IRB-HSR at the University of Virginia in accordance with the U.S. Common Rule. In accordance with IRB Protocol #19272, breast cancer samples were collected as ultrasound-guided core needle biopsies during diagnostic visits from participants who provided informed consent. Each core biopsy was cut into two

pieces, freshly cryoembedded in NEG-50 medium (Richard-Allan Scientific) in a dry ice-isopentane bath, and stored at −80°C wrapped in aluminum foil. Cryosectioning and slide storage was performed exactly as described previously (180).

3.5.2 Rapid histology–immunofluorescence and laser-capture microdissection

For samples with low immune infiltration (UVABC1, UVABC2, and UVABC4) slides were stained and dehydrated as described previously (129). For samples with high immune infiltration (UVABC3 and UVABC5), slides were fixed immediately in 75% ethanol for 30 seconds, rehydrated quickly with PBS, and stained with a mixture of Alexa 488-conjugated KRT8 antibody (Abcam ab192467, 1:20 dilution), YO-PRO-3 (Fisher/Invitrogen Y3607, 1:1000 dilution) and 1 U/ml RNAsin-Plus (Promega) in PBS for one minute. Slides were rinsed twice with PBS before dehydrating with 70% ethanol for 15 seconds, 95% ethanol for 15 seconds, and 100% ethanol for one minute and finally clearing with xylene for two minutes.

Slides were microdissected immediately on an Arcturus XT LCM instrument (Applied Biosystems) using Capsure HS caps (Arcturus). Cells were either visualized by brightfield microscopy (UVABC1, UVABC2, and UVABC4) or with a dual FITC/TRITC filter (UVABC3, UVABC5). The smallest spot size and typical instrument settings (~50 mW power and ~2 msec duration) yielded ~25 μm spot diameters capturing 1–3 breast carcinoma cells per laser shot.

For each biopsy, adjacent clusters of 10 cells were captured as 10-cell samples throughout multiple cryosections to access different regions of the tumor. In addition, a matched number of cells was captured nearby the 10-cell samples on the same LCM cap, extracted as a pool, and diluted into 10-cell equivalents that serve as measurement controls (pool-and-split controls). For each tumor, 10cRNA-seq datasets include 10-cell samples and matched pool-and-split controls captured across multiple days.

3.5.3 RNA extraction and library preparation

RNA extraction and amplification of microdissected samples was performed as described previously (180). Briefly, RNA was eluted from the LCM caps by digesting with

proteinase K, and oligo-dT primed cDNA was synthesized. Residual RNA was degraded by RNAse H (NEB) digestion, and cDNA was poly(A) tailed with terminal transferase (Roche). Poly(A)-cDNA was amplified using an AL1 primer (ATTGGATCCAGGCCGCTCTGGACAAAATATGAATTCTTTTTTTTTTTTTTTTTTTTTTT) and a blend of Taq polymerase (NEB) and Phusion (NEB) for 30 cycles.

Poly (A) PCR-amplified samples were first assessed by quantitative PCR for exogenous ERCC spike in standards and endogenous genes (*GAPDH* and *RPL8* as loading controls and the epithelial marker *KRT8*) as previously described (180). New primers for this study were *KRT8* (Fwd: GCCGTGGTTGTGAAGAAGAT, Rev: CCCCAGGTAGTAAACTCCCC) and *RPL8* (Fwd: CCCAGCTCAACATTGGCAAT, Rev: ACGGGTCTTCTTGGTCTCAG). Samples were retained if the geometric mean of quantification cycles for the *GAPDH–RPL8* loading controls was within 3x the interquartile range of the median calculated across all 10-cell samples of the biopsy. Samples beyond that range were excluded because of over- or under-capture during LCM. For samples with increased immune infiltration, we additionally excluded samples with a detectable quantification cycle for the T cell marker *CD3D* (Fwd: TGCTTTGCTGGACATGAGACT, Rev: CAGGTTCACTTGTTCCGAGC).

Libraries for sequencing were re-amplified, purified, and tagmented as described previously (180). Briefly, each poly(A) PCR cDNA sample was re-amplified for a number of cycles where the amplification remained in the exponential phase (typically 10 to 20). Re-amplified cDNA was then twice purified with Ampure Agencourt XP SPRI beads. After bead purification, samples were quantified on a CFX96 real-time PCR instrument (Bio-Rad) using a Qubit BR Assay Kit (Thermo Fisher). Samples were diluted to 0.2 ng/µl and tagmented with the Nextera XT DNA Library Preparation Kit (Illumina).

3.5.4 RNA sequencing

Libraries from 10-cell samples were multiplexed at an equimolar ratio, and 1.3 pM of the multiplexed pool was sequenced on a NextSeq 500 instrument with NextSeq 500/550 Mid/High

Output v1/v2/v2.5 kits (Illumina) to obtain 75-bp paired-end reads. From the sequencing reads, adapters were trimmed using fastq-mcf in the EAutils package (version ea-utils.1.1.2-779), and with the following options: -q 10 -t 0.01 -k 0 (quality threshold 10, 0.01% occurrence frequency, no nucleotide skew causing cycle removal). Quality checks were performed using FastQC (version 0.11.8) and MultiQC (version 1.7). Data were aligned to the human transcriptome (GRCh38.84) along with reference sequences for ERCC spike-ins using RSEM (version 1.3.0) and Bowtie 2 (version 2.3.4.3). RSEM options for the 10cRNA-seq data also included the following: --single-cell-prior --paired-end. RSEM read counts were converted to transcripts per million (TPM) by dividing each value by the total read count for each sample and multiplying by 10^6. Mitochondrial genes and ERCC spike-ins were not counted towards the total read count during TPM normalization.

3.5.5 Molecular subtype assignments

Microarray-based PAM50 centroids and associated code were obtained from the UNC Microarray Database (https://genome.unc.edu/pubsup/breastGEO/PAM50.zip). To adapt the signature for RNA-seq data, RSEM aligned TPM data for TCGA breast tumors was obtained from the UCSC genome browser (268). Using a balanced number of estrogen receptor (ER)-negative and ER-positive tumors, median RNA-seq values for PAM50 genes were calculated and subtracted from the entire cohort for standardization. Standardized values were used to predict PAM50 subtypes by using downloaded centroids and code from the UNC Microarray Database (12). Successful model training was visualized by a principal component plot showing both training (microarray) and test (TCGA RNA-seq) data clustering by molecular subtype. The same median correction method was attempted for 10cRNA-seq data from UVABC tumors, but model calibration was unsuccessful due to a lack of ER-negative samples and some large differences in overall abundance of PAM50 genes between bulk and 10-cell data. As a substitute, 10cRNA-seq samples were scored for transcriptional modules associated with ESR1 (464 genes), ERBB2 (27 genes), and AURKA (229 genes) using the “subtype.cluster” function

within the package “genefu” (version 2.16.0) in R. On the basis of these module scores, samples were subtyped as Luminal A (ESR1+, ERBB2–, AURKA–), Luminal B (ESR1+, ERBB2–, AURKA+), HER2 (ERBB2+), and Basal (ESR1–, ERBB2–).

3.5.6 UMAP projections

All UMAP projections were generated using the R package “umapr” (version 0.0.0.9001) with the following parameters: neighbors = 4, distance metric = "correlation", minimum distance = 0. For UVABC embedding, RSEM counts for UVABC tumors were converted to TPM values and projected onto a UMAP using all endogenous transcripts. The same UMAP projection was re-colored by subtype classifications and batch number. Batch effects were excluded by visualizing TPM estimates of exogenous ERCC spike-in expression for all UVABC tumors on a separate UMAP. scRNA-seq data of breast tumors (117) were obtained as RSEM-aligned TPM values from the Gene Expression Omnibus (GSE75688). Epithelial carcinoma cells were separated from infiltrating cell types in the scRNA-seq data by selecting cells that expressed *KRT8* and *EPCAM* at TPM > 1. 10cRNA-seq UVABC samples and filtered scRNA-seq data (78 cells) were merged by transcript and projected on the same UMAP.

To account for tumor-specific differences in overall transcript abundance, we centered the expression of every transcript by the 25th quartile of its expression in 10-cell samples from each tumor. Quartile-centered samples for each tumor were then merged by transcript and projected onto a UMAP. From the merged quartile-centered data, expression of 710 RHEGs was extracted and projected onto a separate UMAP.

3.5.7 Overdispersion-based stochastic profiling

10cRNA-seq analysis consisted of an identical algorithm applied separately to 28 10-cell samples and 20 pool-and-split controls from each UVABC tumor. RSEM values were rounded to integer counts, and transcripts with zero counts throughout were removed. Abundance-dependent expected expression and error models were generated separately for 10-cell samples and pool-and-split controls using the “knn.error.models” function in the package “scde”

(version 1.99.4) with *k* nearest neighbors set to 1⁄4 of the sample size for both sets. Only transcripts that had a minimum transcript count of 5 (min.count.threshold) in at least 5 samples (min.non.failed) were considered for model generation. From the abundance-adjusted error models, adjusted-variance estimates of overdispersion were calculated using the "pagoda.varnorm" function of the same package using the generated error models as input. The variance was further adjusted to account for read-depth using the "pagoda.subtract.aspect" function. Transcripts with adjusted variances in 10-cell samples that exceeded the 95th percentile of adjusted variances in pool-and-split controls were considered candidate heterogeneously expressed transcripts. Finally, transcripts with adjusted variances in the top 5th percentile of pool-and-split controls reflecting high measurement variability were filtered out of candidate lists. Overdispersion-based stochastic profiling was applied identically to each of the five UVABC cases to obtain candidate gene lists.

3.5.8 Cohort subsampling for RHEG estimation

The five UVABC tumors were exhaustively downsampled as groups of $n = 1$ (five total possibilities), 2 (10 total), 3 (10 total), 4 (five total), or 5 (one total) and intersected with the operational RHEG definition of transcript heterogeneities that recur in ≥50% of the cases considered: one for $n = 1$ or 2, two for $n = 3$ or 4,and three for $n = 5$.

3.5.9 CNV inference

To predict CNVs, we used InferCNV, which corrects the input expression data (here, 10cRNA-seq) for average gene expression based on normal reference cells and applies a moving average with a sliding window of 101 genes within each chromosome. Smoothed expression values are once again corrected against the normal reference and estimated copy number alterations are reported. Normal breast tissue gene-expression data was obtained from GTEx (300) as a reference dataset for copy-number variation. The reference GTEx data, UVABC 10cRNA-seq data, and a reference genome position file (GRCh38.86) were input to the "CreateInfercnvObject" function in the package "InferCNV" (version 1.0.3) in R (301). The

InferCNV object was then analyzed with the “infercnv::run” function with dynamic threshold denoising to infer copy-number variations as previously described (79).

3.5.10 Periodicity of cell-cycle RHEGs vs. non-RHEGs

We obtained 361 oscillating transcripts from Cyclebase 3.0 (273), reconciled aliases with official gene names, and intersected with the RHEGs from 10cRNA-seq, obtaining 24 shared transcripts. Next, using microarray data from synchronized HeLa cells (77), we identified probesets for 21 of the 24 shared transcripts along with those of the top 10 cycling transcripts according to Cyclebase 3.0 (273). The centered probeset data and HeLa pseudotime estimates are available through Cyclebase 3.0 for three complete cell cycles, but only the first two cycles show strong synchronization (273). We calculated the time interval above the mean value and compared it as a ratio to each of the adjacent time intervals below the mean value. For the ratio, the larger pseudotime interval (above or below the mean) was placed in the numerator. We calculated the skewness of the ratio distributions (indicating time-interval asymmetry above- vs.-below the mean) for the 21 RHEGs with identifiable probesets vs. the top 10 cycling transcripts and estimated uncertainty by bootstrapping.

3.5.11 Monte Carlo simulations of three-state stochastic profiling

Monte Carlo simulations of stochastic profiling under the assumption of two regulatory states were performed in MATLAB as described with available software (157). The two-state model assumes a binomial distribution for the cellular dichotomy and log-normal distribution of measured transcripts. To build a three-state model reflecting cell cycle-regulated variation, we used a multinomial distribution to reflect cell-cycle fractions, two lognormal regulatory states for G1 and G2/M phases, and a uniform “S-phase” interval between the two other regulatory states. Two- or three-state distributions were compared against a null distribution of lognormal variation using the Kolmogorov-Smirnov test. The simulations for a parameter set were run 50 times to measure the median p values and the associated nonparametric confidence intervals.

Stochastic sampling was deemed effective when the median p value for $F \neq 0$ (multiple states) was less than 0.05 and the median p value for $F = 0$ (one state) was greater than 0.05.

3.5.12 Gene signature overlaps with RHEGs

All overlaps were viewed using the "venn" function in the R package gplots (version 3.0.1.1), and intersections were obtained using the "intersect" function. Significance of overlap was calculated using the hypergeometric test in R through the "phyper" function and a total gene count of 20,000. For assessing detachment-induced artefacts in the RHEG gene set, we obtained a murine detachment-induced gene signature (143) and converted mouse genes to human orthologs with the Ensembl biomart in R. Any remaining mouse gene names were capitalized in accordance with human gene symbol conventions. The human ortholog mapping was verified against the human transcriptome reference gene list used for 10cRNA-seq data. Breast driver genes were obtained from a larger gene list of cancer drivers (260) filtered for tissue of origin. We similarly assessed overlap with transcripts altered by CTCF knockout in luminal breast cancer cells (276), GATA3 target genes (277), EMT signature gene sets (261,280), and an aggregated pan-cancer driver gene set (260,281). For the last comparison, nine RHEGs were considered proximal to driver genes (**Table 3.3**) and treated as equivalent between the two sets. All mismatched gene aliases were corrected before overlap testing.

3.5.13 Statistics

Sample sizes for stochastic profiling were determined by Monte Carlo simulation (157). Differences in genes detected per sample between 10cRNA-seq and scRNA-seq were assessed by Kolmogorov-Smirnov test using the "ks.test" function in R. Significance of overlaps between candidate genes identified in different UVABC tumors were assessed by Monte Carlo simulations that drew the total number transcripts per tumor randomly from a common pool of 14,824 genes (total transcripts eligible for overdispersion analysis in all tumors). Observed overlaps were compared with 1000 Monte Carlo simulations to estimate a p value, which was adjusted for multiple comparisons by using the Šidák correction. All overlaps between the

RHEG set and other gene sets were assessed for significance by hypergeometric test using the "phyper" function in R and a background of 20,000 genes. Kolmogorov-Smirnov tests for Monte Carlo simulations for stochastic profiling were assessed using the "kstest" function in MATLAB. Hierarchical clustering was performed using "pheatmap" in R using Euclidean distance and "ward.D2" linkage.

3.5.14 Data availability

10cRNA-seq data from this study is available through the NCBI Gene Expression Omnibus (GSE147356, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147356 Reviewer token: wzcfouoijfodzij).

4 Fragmentation of Small-cell Lung Cancer Regulatory States in Heterotypic Microenvironments

4.1 Foreword

Small-cell lung cancers (SCLC) derive from pulmonary neuroendocrine cells, which have stem-like properties to reprogram into other cell types upon lung injury. SCLC cells display phenotypic plasticity, which can be difficult to uncouple from heritable changes that evolve in primary tumors or select in metastases to distant organs. Conventional profiling approaches are problematic for SCLC if the required sample dissociation activates injury-like signaling and reprogramming.

In the previous chapter, we defined cell-state heterogeneities through 10cRNA-seq coupled with stochastic-profiling fluctuation analysis. In this chapter, we applied these approaches to a SCLC mouse model initiated by neuroendocrine deletion of tumor suppressor genes *p53* and *Rb*. To dissect microenvironmental influences, we compared SCLC cells in spheroid cultures and in murine liver colonies seeded intravenously. Fluctuating transcripts in vitro were partly shared with other epithelial-spheroid models, and candidate heterogeneities increased considerably when cells were delivered to the liver. Liver colonization of mice drove the fractional appearance of alveolar type II-like markers and poised cells for paracrine stimulation from immune cells and hepatocytes. Candidate heterogeneities recurrent in the liver also stratified primary human tumors into discrete groups not readily explained by molecular subtype. We conclude that heterotypic interactions in the liver and lung are an accelerant for intratumor heterogeneity in small-cell lung cancer.

The work presented in this chapter includes significant experimental contributions from Dylan Schaff, an undergraduate student in the Janes Lab that I mentored. A manuscript of this work with Mr. Schaff and I as co-first authors is under review (270).

4.2 Introduction

The categories, origins, and organization of tumor cell-to-cell heterogeneity are open questions of fundamental importance to cancer biology (124). Within normal tissues, single cells differ by lineage type and regulatory state (233). These distinctions blur, however, when cells lose their proper context because of tissue damage (170,302), transformation (303–305), or metastatic colonization (98,306). The details of such adaptive heterogeneity are expected to depend heavily on the originating cell type, the state of the cell when perturbed, and the local microenvironment where the cell resides.

Within the lung, the pulmonary neuroendocrine cell (PNEC) is a rare-but-important cell type that acts as an airway sensor for damaging stimuli (307). PNECs self-organize into 20–30-cell clusters at airway branch points through dynamic rearrangement of cell-cell contacts and reversible state changes suggesting epithelial-to-mesenchymal transition (EMT) (308). The latest evidence supports that certain PNECs have a reservoir of plasticity to convert into other lung cell types during tissue damage (170). A state of chronic wounding characterizes many tumors– metastases (309), and PNECs are the main cell type of origin for small-cell lung cancer (SCLC) (167–169), a deadly form of lung carcinoma.

Regulatory mechanisms of SCLC plasticity are beginning to be dissected through systems-biology approaches (177,178) and genetically-engineered mouse models (GEMMs) (173). Human SCLC requires loss of *RB1* and *TP53* (6,310)—two tumor suppressors that also developmentally restrict pluripotency (311,312). GEMMs with *Rb1*–*Trp53* deleted by intratracheal delivery of Cre-expressing adenovirus (AdCMV-Cre) give rise to murine SCLCs similar to the classic ASCL1-high subtype of human SCLC (175,313). Deletion of additional tumor suppressors can synergize with *Rb1*–*Trp53* loss (314–316). For example, progression is accelerated by compound deletion of the Rb-family member *p130* (223). These GEMMs

(167,168) and others (169) were instrumental in defining PNECs as the cell type of origin for SCLC.

Interestingly, phenotypes of the resulting murine tumors depend on the maturation state of PNECs targeted for *Rb1–Trp53* deletion. Restricting adenoviral Cre to PNECs positive for Calca causes far fewer SCLCs to develop compared to when Cre expression is driven by a strong cytomegalovirus (CMV) promoter (317). Both tumor models are metastatic, but only the CMV-driven GEMM upregulates the transcription factor Nfib, which promotes widespread chromatin opening (318) and cell-lineage changes in both primary and metastatic sites (317). Many murine SCLC-derived cell lines are admixtures of cells with neuroendocrine and “non-NE” mesenchymal features (319). Other non-NE SCLC subpopulations are maintained by Notch signaling (173), which may also become activated in normal PNECs during injury-induced reprogramming (170). There might be other triggers of cell-fate heterogeneity to uncover if SCLC regulatory states could be examined at single-cell resolution without injury-like dissociation of cellular context.

In this work, we examined the in situ transcriptomic regulatory heterogeneities of an established murine SCLC culture derived from an $Rb1^{F/F}; Trp53^{F/F}$ animal administered AdCMV-Cre [KP1 cells (179); **Figure 4.2A**]. Using GFP-labeled cells, fluorescence-guided laser capture microdissection (LCM), and 10-cell RNA sequencing (10cRNA-seq) (180), we considered three biological contexts: 1) tumor spheroids cultured in vitro and liver colonies in mice 2) lacking or 3) retaining an intact immune system (**Figure 4.2B**). KP1 tumorspheres exhibited cell-to-cell regulatory heterogeneities in cell biology, aging, and metabolism that were shared with spheroid cultures of breast epithelia (129,180). Liver colonization gave rise to pronounced cell-state changes suggesting that paracrine signaling from the lung was partially resurrected in the liver. Liver colonies in immunocompetent animals showed an exacerbated breadth of cell fates, with observed alveolar type II (ATII) markers intermingling with many non-NE stromal markers documented in SCLCs (319) and PNECs (170). Intersecting the three datasets yielded core

recurrent heterogeneously expressed genes (RHEGs) and an in vivo RHEG set that was shared by all liver colonies but absent in tumorspheres. Core RHEGs from KP1 cells were broadly shared in bulk human SCLC transcriptomes, yet covariations among cases were not discriminating. By contrast, in vivo RHEGs showed weaker overall correlations but clustered human data into discrete groups that were separate from any 10-cell KP1 transcriptomes. The in vivo RHEGs defined here may reflect a set of injury-like SCLC adaptations that are possible during tumor growth and metastasis at different organ sites.

4.3 Results

4.3.1 Study design and rationale

We sought to define how SCLC regulatory heterogeneity was compiled in different microenvironments. To avoid confounding variation in GEMM tumors that arise autochthonously, we used KP1 cells, a polyclonal *Trp53*$^{\Delta/\Delta}$*Rb1*$^{\Delta/\Delta}$ line derived from a tumor initiated by intratracheal administration of AdCMV-Cre. We sequenced the bulk transcriptome of KP1 cells and found that they were very similar to three other *Trp53*$^{\Delta/\Delta}$*Rb1*$^{\Delta/\Delta}$ lines prepared in similar genetic backgrounds (GSE147358). By contrast, autochthonous SCLC tumors from related GEMMs (317) were different and also more variable among primary tumors, as expected (**Figure 4.1**). Before starting, we genetically labeled KP1 cells with EGFP for unambiguous isolation of cells administered in vivo (**Figure 4.2A**).

SCLCs frequently metastasize to the liver (320). We mimicked the terminal steps of metastatic colonization and outgrowth by tail-vein injection of EGFP-labeled KP1 cells, which readily establish lesions in the livers of athymic nude mice (**Figure 4.2B**). Although KP1 cells have a mixed genetic background, we discovered that subcutaneous xenografts and liver colonies were 100% successful in first-generation crosses (F_1) of C57/B6 and 129S inbred

strains. Inoculating C57/B6 x 129S F_1 hybrid animals thus afforded a third setting in which liver colonization and expansion could occur in the presence of a cell-mediated immune response.

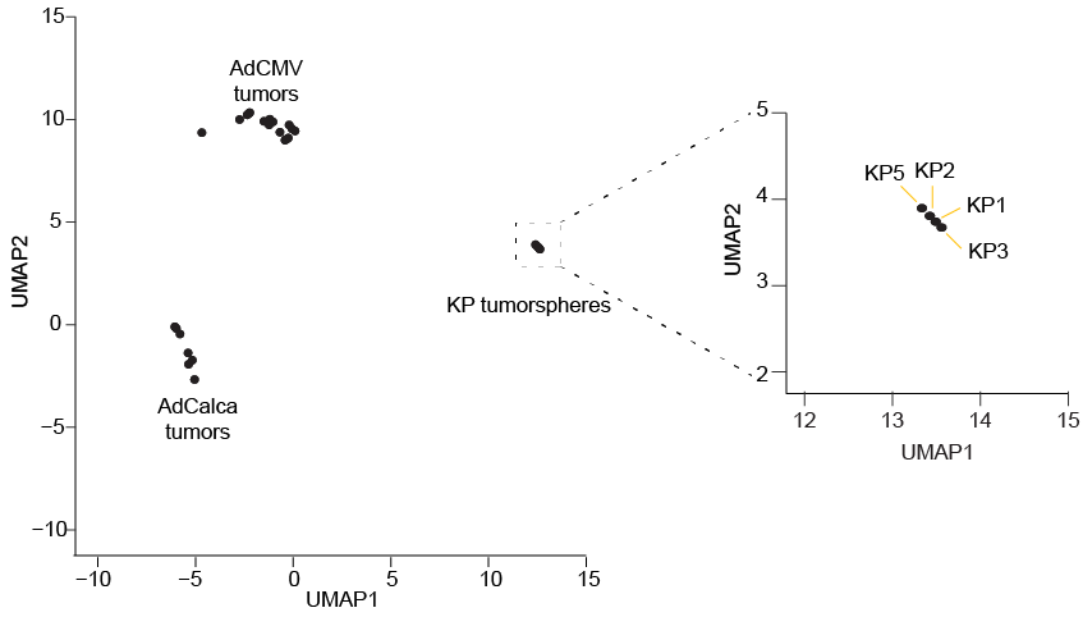

Figure 4.1 KP1 cells are representative of SCLC lines derived from *Rb1$^{F/F}$Trp53$^{F/F}$* mice administered CMV-driven adenoviral Cre (AdCMV-Cre).
Autochthonous AdCMV-Cre-initiated and AdCalca-Cre-initiated tumors from *Rb1$^{F/F}$Trp53$^{F/F}$Rbl2$^{F/F}$* mice (317) are shown for comparison. Genotypes for the KP SCLC lines are: KP1, *Rb1$^{\Delta/\Delta}$Trp53$^{\Delta/\Delta}$* (32); KP2, *Rb1$^{\Delta/\Delta}$Trp53$^{\Delta/\Delta}$Ptch1$^{+/Lacz}$* (179); KP3, *Rb1$^{\Delta/\Delta}$Trp53$^{\Delta/\Delta}$Axin2$^{+/Lacz}$* (321); KP5, *Rb1$^{\Delta/\Delta}$Trp53$^{\Delta/\Delta}$Tg$^{BAT\text{-}lacZ}$* (322).

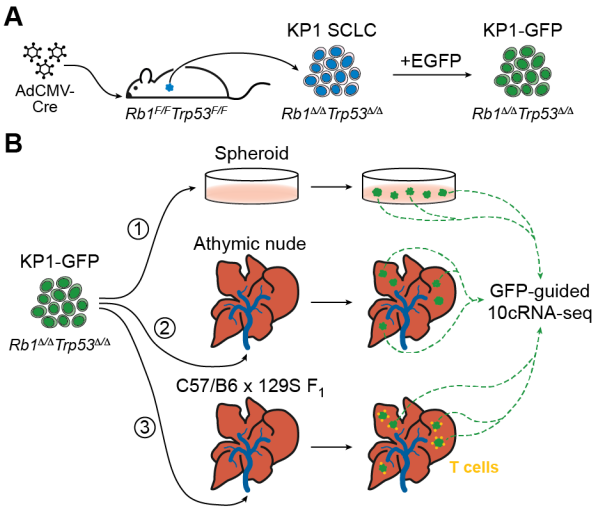

Figure 4.2 Stochastic profiling of transcriptional regulatory heterogeneity in three isogenic SCLC contexts.

(A) Derivation of KP1 small-cell lung cancer (SCLC) cells by intratracheal administration of adenovirus delivering cytomegalovirus promoter-driven Cre recombinase (AdCMV-Cre) to *Rb1*$^{F/F}$*Trp53*$^{F/F}$ animals. The KP1 SCLC line was engineered to express ectopic enhanced green fluorescent protein (EGFP) for fluorescence-guided microdissection.

(**B)** The KP1-GFP derivative line was 1) cultured as three-dimensional spheroids in vitro or colonized to the liver of 2) athymic nude mice or 3) C57/B6 x 129S F_1 hybrid mice harboring an intact immune system. GFP-positive cells from multiple spheroids and liver colonies were randomly captured and measured by 10-cell RNA sequencing (10cRNA-seq) for stochastic profiling.

4.3.2 KP1 tumorspheres share adaptive transcriptional regulatory heterogeneities with breast- epithelial spheroids

Cultured KP1 cells grow as spheroidal aggregates that can be readily dissociated enzymatically, but juxtacrine cell-cell interactions may contribute to the overall heterogeneity of the population (173). Therefore, we processed KP1 spheroids exactly as if they were tissue, cryoembedding within seconds and sectioning–staining as described (180). Cells were microdissected from the outermost periphery of each spheroid to ensure that all cells profiled had equal availability of nutrients. We gathered 10-cell pools across multiple spheroids to average out subclonal differences within the line and highlight pervasive heterogeneities that characterize spheroid culture. Using 10cRNA-seq (180), we measured the transcriptomes of 28 separate 10-cell groups of KP1 cells along with 20 pool-and-split controls as 10-cell equivalents obtained by LCM. The data were analyzed for candidate regulatory heterogeneities by stochastic profiling (129) implemented with an overdispersion metric optimized for RNA-seq as described in Chapter 3 (153,154,253). The analysis yielded 405 candidate genes that were much more variable in the 10-cell samples than expected given their average abundance and technical reproducibility (**Figure 4.3A**).

Samples were collected across multiple days to assess whether batch effects dominated the fluctuation analysis. We clustered gene candidates hierarchically and asked whether the fluctuation signatures clustered according to when the 10-cell samples were collected (**Figure 4.3B**). Each grouping was comprised of 10-cell profiles from all batches, supporting that the analytical strategy was robust amidst day-to-day variations in LCM, RNA extraction, and sample amplification. Standard gene set enrichment analysis indicated hallmarks for cell-cycle transitions, Myc–mTORC1 signaling, and metabolism, consistent with the variable growth of spheres in the culture.

Previously, our group used 10cRNA-seq to revisit an earlier analysis of transcriptional regulatory heterogeneity in 3D cultured MCF10A-5E breast-epithelial spheroids (129,180). With

an analytical pipeline for stochastic profiling by 10cRNA-seq now in hand (Chapter 3), we quantified the gene-by-gene overdispersion and identified 1129 candidate heterogeneities (**Figure 4.4A** and **B**). The list included multiple transcripts that were independently validated to be heterogeneous by RNA fluorescence in situ hybridization (129,158), including one transcript (*SOX4*) that we validated here (**Figure 4.4B and C**). The analysis provided a second context for regulatory heterogeneity that exists during spheroidal growth.

Murine SCLC cells and human breast epithelial cells are undoubtedly very different, but normal PNECs derive from an epithelial lineage (308) and often adopt a columnar morphology similar to that seen in the breast. The KP1 study detected significantly fewer genes as regulated heterogeneously compared to MCF10A-5E ($p < 10^{-15}$ by binomial test), corroborating the differences in spheroid culture format. MCF10A-5E cells were 3D cultured in reconstituted basement membrane, which traps secreted factors locally around the spheroids (202), whereas KP1 spheroids develop freely in suspension. Despite differences in the overall number of candidates, we found significant overlap in shared genes after mapping mouse and human orthologs (see Materials and Methods; **Figure 4.3C**). Intersecting the two gene groups only marginally enriched for cell-cycling transcripts (273) (six of 57 genes, $p = 0.04$ by hypergeometric test), suggesting other biological processes in addition to proliferation. The intersection raised the possibility that cell growth–competition within epithelial spheroids elicits a set of RHEGs, which generalize beyond a specific culture format.

We next asked whether there might be any common heterogeneities in regulation between the two contexts after correcting for transcript abundance. When the standardized fluctuations of KP1 and MCF10A-5E spheroids were coclustered by gene ortholog, there were multiple close pairings consistent with shared biology or biological category (**Figure 4.3D**). For instance, the importin *KPNA2* covaried with its exportin, *CSE1L* (**Figure 4.3E**) (323). We also observed cross-species correlations in genes functioning at the interface of the plasma membrane and endoplasmic reticulum: *ESYT1* and *SPTAN1* (**Figure 4.3F**) (324). Although near

the detection limit for both cell types, we noted cofluctuations in *SIRT3* and *CTC1*, two factors implicated in cellular longevity (**Figure 4.3G**) (325,326). Together, these gene pairings provide a basis for hypotheses about single-cell regulatory pathways that become co-activated when epithelia proliferate outside of their normal polarized context.

Elsewhere among the spheroid RHEGs, we found instances of mutually exclusive transcript heterogeneities, such as with *HUWE1* and *TRIP12* (**Figure 4.3H**). These E3 ubiquitin ligases have been reported to operate independently in triggering ubiquitin fusion degradation (327), an unusual proteasomal pathway not studied in cancer. Separately, we recognized a preponderance of metabolic enzymes related to lipids and clustered the fatty acid elongase *ELOVL1*, the β-oxidation dehydrogenase *ACADVL*, and the α-oxidation hydroxylase *PHYH* (**Figure 4.3I**). Even with 10-cell pooling, we rarely observed these enzymes abundantly expressed in the same sample, suggesting independent states of lipid synthesis and degradation that could be mined deeply in the future for covariates. Expanding candidate lists around positive and negative covariates has proved powerful in mechanistic follow-on work (202,328).

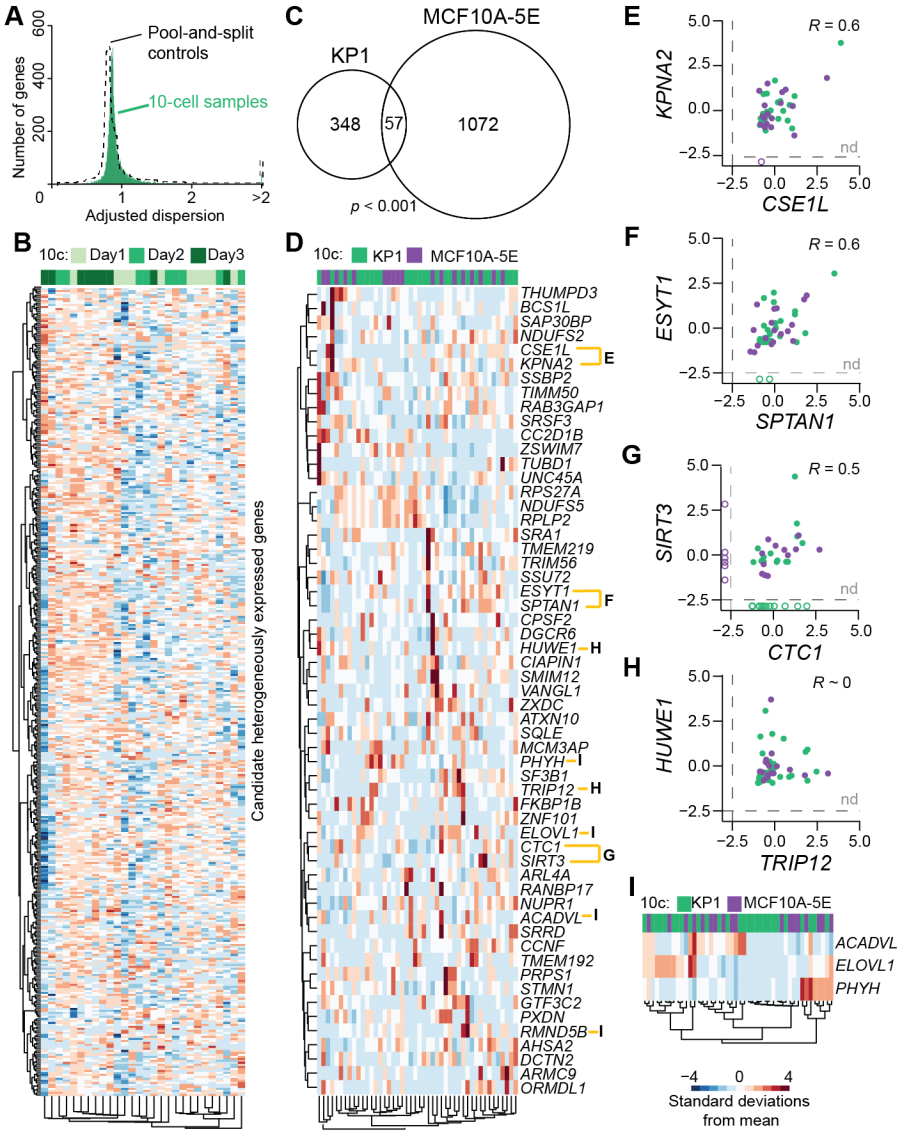
A
Number of genes
Pool-and-split controls
10-cell samples
Adjusted dispersion
C
MCF10A-5E
KP1
348
57
1072
p < 0.001
E
KPNA2
CSE1L
R = 0.6
nd
F
ESYT1
SPTAN1
R = 0.6
nd
G
SIRT3
CTC1
R = 0.5
nd
H
HUWE1
TRIP12
R ~ 0
nd
B
10c: Day1 Day2 Day3
Candidate heterogeneously expressed genes
D
10c: KP1 MCF10A-5E
THUMPD3
BCS1L
SAP30BP
NDUFS2
CSE1L
KPNA2
E
SSBP2
TIMM50
RAB3GAP1
SRSF3
CC2D1B
ZSWIM7
TUBD1
UNC45A
RPS27A
NDUFS5
RPLP2
SRA1
TMEM219
TRIM56
SSU72
ESYT1
SPTAN1
F
CPSF2
DGCR6
HUWE1
H
CIAPIN1
SMIM12
VANGL1
ZXDC
ATXN10
SQLE
MCM3AP
PHYH
I
SF3B1
TRIP12
H
FKBP1B
ZNF101
ELOVL1
I
CTC1
SIRT3
G
ARL4A
RANBP17
NUPR1
ACADVL
I
SRRD
CCNF
TMEM192
PRPS1
STMN1
GTF3C2
PXDN
RMND5B
I
AHSA2
DCTN2
ARMC9
ORMDL1
I
10c: KP1 MCF10A-5E
ACADVL
ELOVL1
PHYH
Standard deviations from mean

Figure 4.3 Shared transcriptional regulatory heterogeneities between KP1 spheroids and MCF10A-5E breast-epithelial spheroids.
(A) Overdispersion plot showing 10-cell sample distribution for KP1 spheroids (green) overlaid on pool-and-split controls of 10-cell equivalents (dashed).
(**B)**Clustergram of the 405 transcripts identified as candidate heterogeneities within KP1 spheroids. Sample acquisition days are annotated. Data were log transformed before standardization.
(C) Venn diagram of orthologous candidates between KP1 spheroids and MCF10A-5E spheroids analyzed in **Figure 4.4**. Significance of the intersection was assessed by hypergeometric test with 12,612 total detectable transcripts in KP1 cells and 12,927 total detectable transcripts in MCF10A-5E cells.
(D) Clustergram of the spheroid RHEGs annotated by human ortholog.
(E–H) Pairwise Pearson correlations between the indicated gene pairs in **C** among samples where both genes were detected (filled). nd, not detected.
(**I)** Clustergram of three transcripts encoding enzymes for lipid metabolism.
In **D–I**, 10-cell samples of KP1 cells (green) and MCF10A-5E cells (purple) were standardized separately by z-score before clustering or correlation.

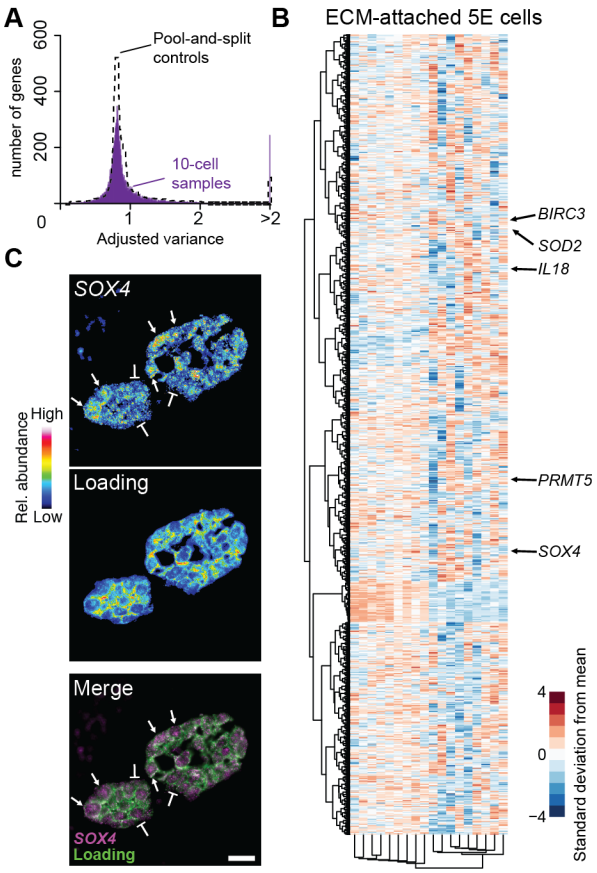

Figure 4.4 Stochastic profiling of transcriptional regulatory heterogeneities in MCF10A-5E spheroids by 10cRNA-seq.

(A) Overdispersion plot showing 10-cell sample distribution for MCF10A-5E spheroids (purple) overlaid on pool-and-split controls of 10-cell equivalents of KP1 spheroids (dashed).

(B) Clustergram of the 1129 transcripts identified as candidate heterogeneities within MCF10A-5E spheroids. Data were log transformed before standardization. Indicated transcripts were independently validated as heterogeneous by RNA FISH (129,158).

(C) RNA FISH validation of *SOX4*, a gene candidate identified by 10cRNA-seq stochastic profiling. Pseudocolor image for *SOX4* is shown above a loading control hybridization comprised of *GAPDH*, *HINT1*, and *PRDX6 (201)*. Scale bar is 20 μm.

4.3.3 SCLC reprogramming and paracrine signaling are initiated by colonization of KP1 cells to the liver

To begin examining how heterotypic interactions augment SCLC regulatory heterogeneity, we dissociated KP1 spheroids and colonized the liver of athymic nude mice (**Figure 4.2B**). Upon entering the liver circulation, cancer cells extravasate from sinusoids and proliferate amidst hepatocytes. We ensured that SCLC-hepatocyte communication was reflected in the 10cRNA-seq data by sampling the margins of separate GFP+ KP1 liver colonies, analogous to the spheroid margins sampled in vitro (**Figure 4.5A**). Focusing on the KP1–hepatocyte interface implied that some level of cell contamination would be introduced by collateral pickup during the LCM step. We rigorously controlled for hepatocyte contamination through a two-step negative- selection procedure. Samples were excluded if hepatocyte markers were abundant by qPCR, and transcripts were removed post-analysis if they covaried with the residual hepatocyte content in the sequenced sample (**Figure 4.5B**). Additionally, we oversampled the in vivo samples, collecting 33 10-cell pools that were subsampled 100 times as random groups of 28 for the dispersion analysis (see Materials and Methods; **Figure 4.5B** and **Figure 4.6**). The pipeline collectively identified 898 robust candidates fluctuating independently of residual liver markers and appearing in ≥75% of subsampling runs (**Figure 4.5C**).

Enriched gene sets were virtually identical to spheroid cultures, except for the addition of STAT5 and interferon γ hallmarks likely resulting from innate immune responses. Beyond the significant increase in candidates ($p < 10^{-15}$ by binomial test), we noted that sample-to-sample fluctuations were qualitatively more dramatic on the margin of liver colonies when compared to KP1 spheroids (**Figure 4.3B and Figure 4.5C**). Multiple, smaller subsets of candidates were especially interesting. For example, among the robust candidates were the alveolar type II (ATII) markers *Cd74* (174) and *Lyz2* (329), as well as the Cd74 ligand, *Mif*. Surprisingly, when the sample-by-sample fluctuations of these three genes were clustered, we did not detect any significant co-occurrence that would have suggested full transdifferentiation to an ATII

phenotype (see Materials and Methods; **Figure 4.5D**). The results agree with scRNA-seq data obtained in deprogrammed PNECs (170), where *Cd74* and *Lyz2* markers are anti-correlated among cells with a non-NE phenotype (**Figure 4.7**). The patterns detected by stochastic profiling suggest that a subset of KP1 cells reprogram into partial ATII-like states, only one of which senses Mif produced locally (**Figure 4.5E**).

Other groups of transcripts required inputs from non-KP1-derived cell types in the liver to rationalize. Two robust candidates were the NF-κB subunit *Rela* and an NF-κB target gene, *Sod2* (330), which co-occurred strongly ($p < 0.1$) when considering that NF-κB is mostly regulated posttranslationally (**Figure 4.5F**). Within the candidate heterogeneities, we also identified the NF-κB-inducing receptor *Ltbr* (331), which varied separately from *Rela*–*Sod2*. However, the Ltbr ligand (*Ltb*) was effectively absent in KP1 cells (less than 1.5 TPM in bulk samples and pool-and-split controls from liver colonies). We searched Tabula Muris (332) and found that *Ltb* is abundantly expressed in hepatic natural killer (NK) cells, the most-prevalent lymphocyte population in the liver (333). Given that NK cell activity is retained or enhanced in athymic nude mice (334), their paracrine communication with KP1 cells is a plausible mechanism for heterogeneous NF-κB pathway activation in the liver.

We also found evidence for variable regulation in signal transducers of interleukin 1-family cytokines. The inhibitory adaptor *Tollip* (335), the mitogen-activated protein kinase (MAPK) kinase kinase *Map3k7* and its activator *Tab1* (336), the downstream stress-activated MAPKs *Mapk8* (or *Jnk1*) and *Mapk14* (or *p38α*), and the MAPK phosphatase *Dusp8* were all robust transcript heterogeneities in KP1 liver colonies (**Figure 4.5C**). Clustering the 10-cell fluctuations of these genes indicated that elevated *Mapk14* levels co-occurred with reduced abundance of *Mapk8* and *Dusp8* ($1 - p < 0.1$; **Figure 4.5G**). Signaling along these parallel MAPK effector pathways may be weighted differently among SCLC cells in the liver colony. Although no relevant receptors were detectably overdispersed in KP1 cells, we consistently detected *Il1rl1*, which is the receptor for Il33 of the interleukin 1 family. PNECs normally receive

Il33 stimulation as an alarmin from ATII cells during lung injury or infection (337), but Il33 is also highly expressed in hepatocytes and liver sinusoids (332,338). The widespread single-cell adaptations downstream of Il33 support the hypothesis that SCLC cells redeploy native damage-response pathways in the liver microenvironment.

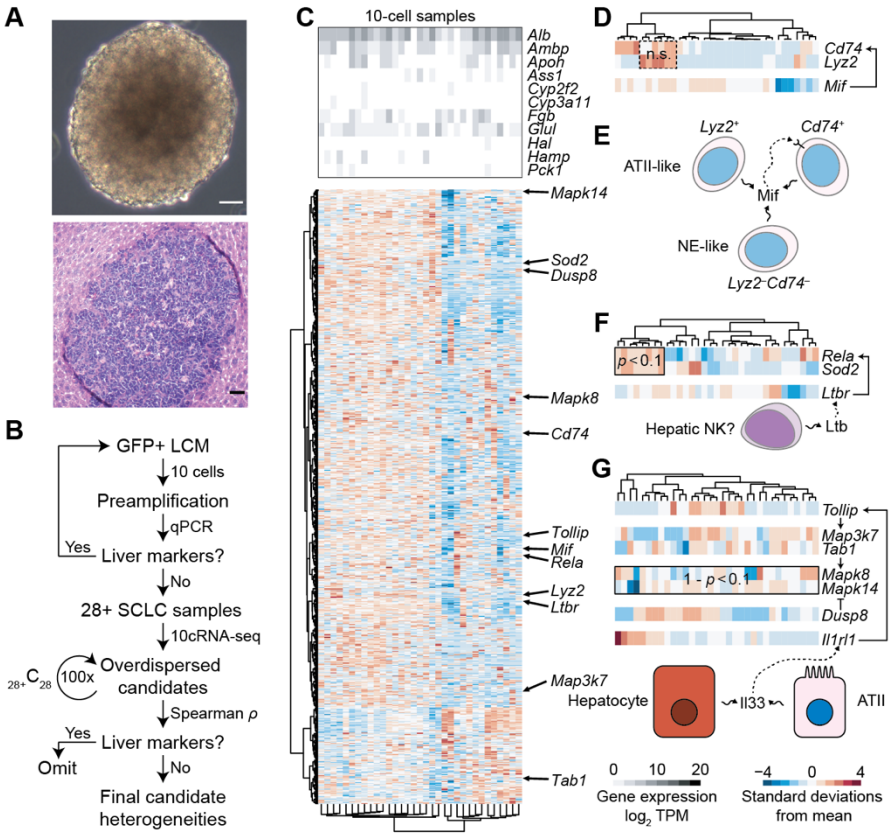

Figure 4.5 KP1 liver colonization in athymic nude mice causes partial reprogramming and engages heterotypic paracrine-signaling networks.

(A) Phase-contrast image of cultured KP1 spheroids (upper) compared to a brightfield hematoxylin-eosin stain of a KP1 liver colony in an athymic nude mouse (lower). Scale bar is 80 μm.

(B) Flowchart illustrating the experimental and analytical strategy controlling for liver contamination in 10cRNA-seq data and in candidate heterogeneities identified by stochastic profiling. Subsampling results from the 100 dispersion analyses of 28 10-cell samples are shown in **Figure 4.6**.

(C) Relative abundance of liver markers (upper) and log-standardized 10-cell sampling fluctuations of robust candidate heterogeneities identified by stochastic profiling (lower).

(D) Log-standardized sampling fluctuations of the alveolar type II (ATII) markers *Cd74* and *Lyz2* together with the Cd74 ligand, *Mif*.

(E) Schematic illustrating the hypothesized relationship between neuroendocrine (NE)- and ATII-like states and Mif signaling.

(F) Log-standardized sampling fluctuations of the *Rela* transcription factor, *Sod2* (a Rela target gene), and *Ltbr* (a Rela-inducing receptor). The ligand for Ltbr is produced by liver-resident NK cells (332).

(G) Log-standardized sampling fluctuations of the indicated signaling transcripts and their pathway relationships downstream of the Il33 receptor, Il1rl1, which is present in KP1 cells but not heterogeneously regulated. Il33 is produced by hepatocytes in the liver and ATII cells in the lung.

For **D, F,** and **G,** enriched or unenriched coexpression was evaluated by hypergeometric test of 10-cell observations above their respective logmeans. n.s., $p > 0.1$ and $1–p > 0.1$.

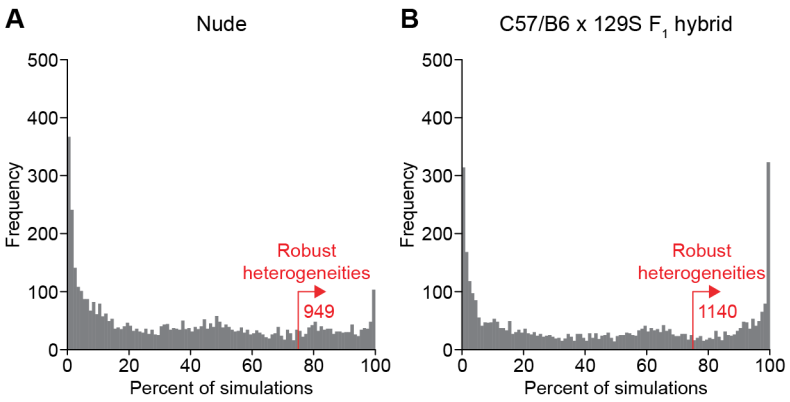

Figure 4.6 Subsampling identifies robust transcriptional regulatory heterogeneities within KP1 liver colonies.
(A) Subsampled dispersion analysis of 33 10-cell observations of KP1 cells colonized to the liver of athymic nude mice.
(B) Subsampled dispersion analysis of 31 10-cell observations of KP1 cells colonized to the liver of immunocompetent C57/B6 x 129S F_1 hybrid mice. Datasets were randomly downsampled to 28 10-cell observations and analyzed for overdispersion as described in Chapter 3, and candidates appearing in >75% of subsampling runs were considered robust heterogeneities.

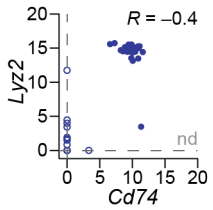

Figure 4.7 *Cd74* and *Lyz2* are anti-correlated in single PNEC-derived non-NE cells.
scRNA-seq reads of *Lyz2* and *Cd74* are shown for non-NE cells (170). Significance of the Pearson correlation (*R*) was tested after Fisher Z transformation (one-sided $p < 0.05$).

4.3.4 Immunocompetency exacerbates stromal non-NE phenotypes in SCLC liver colonies

We built upon the results in athymic nude mice by repeating the liver colonization experiments in C57/B6 x 129S F_1 hybrid mice. Compared to KP1 colonies in athymic mice, the C57/B6 x 129S F_1 hybrid colonies had a higher proportion of Cd3+ T cells and a reduced proportion of F4/80+ macrophages along the colony margin (**Figure 4.8A–D**), indicating different microenvironments. Stochastic profiling of the KP1 colony margins was performed in C57/B6 x 129S F_1 hybrid livers exactly as for athymic nude animals (**Figure 4.5B**). Abundance of liver markers in the C57/B6 x 129S F_1 hybrid samples were as low and uncorrelated as in the nude samples (**Figure 4.9**). From 31 10-cell transcriptomic profiles, we robustly identified 1025 regulatory heterogeneities within KP1 cells colonized to an immunocompetent liver (**Figure 4.8E**).

Gene set enrichment of the C57/B6 x 129S F_1 hybrid candidates reconstituted most of the hallmarks identified previously along with a moderate signature for hypoxia. In search of shared themes, we compared the KP1 candidate genes from the three biological contexts and found that all two- and three-way intersections were significant ($p < 0.001$ by Monte-Carlo simulation; **Figure 4.8F**). This suggested that biological meaning might be embedded in the heterogeneity trends between groups. In lieu of hard overdispersion thresholds (as in **Figure 4.3A**), we next analyzed the adjusted variance as a continuous measure of predicted heterogeneity. Beginning with the 2007 transcripts predicted to be heterogeneously regulated in at least one context (**Figure 4.8F**), we searched for genes with significant overdispersion increases in the immunocompetent setting (see Materials and Methods). We identified 202 transcripts meeting these criteria, which included multiple neuroendocrine markers (*Rtn2*, *Pcsk1*), *Cd74*, and a new group of stromal transcripts (*Bgn*, *Sparc*, *Mgp*, *Cep19*; **Figure 4.8G**) (170). Mesenchymal transitions of SCLC cells can be driven by activated Kras (319), and we noticed that the dispersion of wildtype *Hras* increased alongside the stromal transcripts.

However, when 10-cell fluctuations were clustered, we found that the co-occurrence of *Cd74–Bgn–Sparc* associated with a lack of elevated *Hras* abundance (**Figure 4.8H**), excluding a straightforward EMT-like state change. The stromal markers *Mgp* and *Cep19* were also uncoupled from *Cd74–Bgn–Sparc*. We conclude that immunocompetency drives a further diversification of SCLC toward stromal phenotypes in the setting of liver colonization.

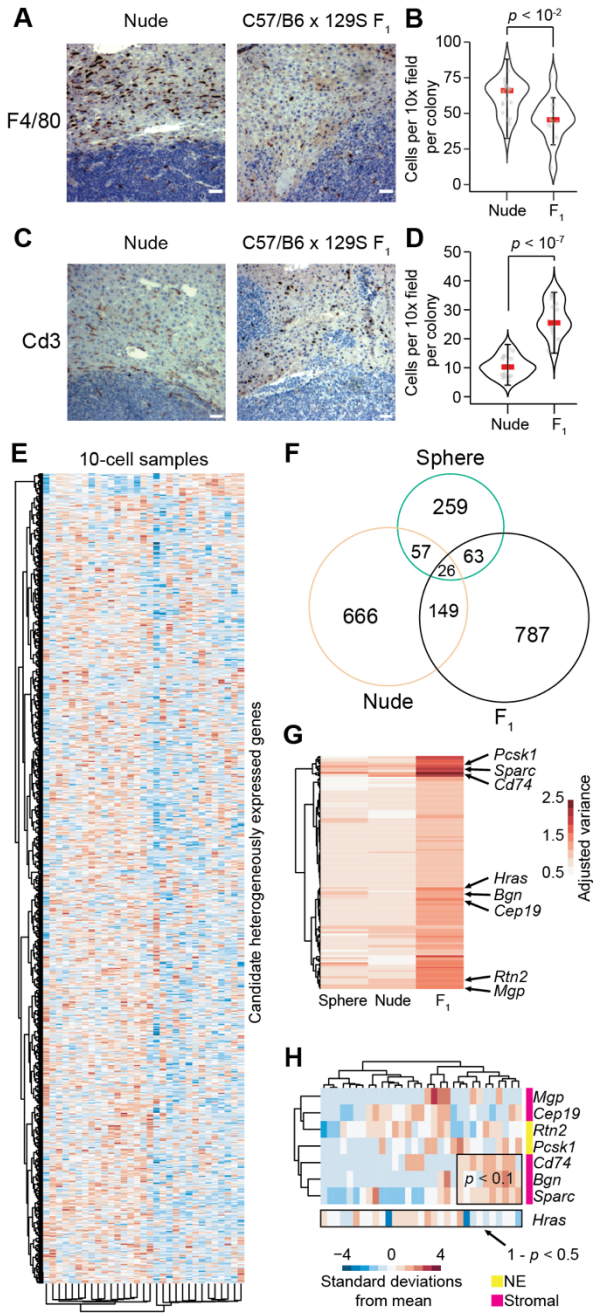
A
Nude
C57/B6 x 129S F1
F4/80
B
p < 10-2
Cells per 10x field per colony
100
75
50
25
0
Nude
F1
C
Nude
C57/B6 x 129S F1
Cd3
D
p < 10-7
Cells per 10x field per colony
50
40
30
20
10
0
Nude
F1
E
10-cell samples
Candidate heterogeneously expressed genes
F
Sphere
259
57
63
26
666
149
787
Nude
F1
G
Pcsk1
Sparc
Cd74
Hras
Bgn
Cep19
Rtn2
Mgp
2.5
1.5
0.5
Adjusted variance
Sphere
Nude
F1
H
Mgp
Cep19
Rtn2
Pcsk1
Cd74
Bgn
Sparc
Hras
p < 0.1
1 - p < 0.5
−4
0
4
Standard deviations from mean
NE
Stromal

Figure 4.8 Stromal markers emerge heterogeneously when KP1 cells colonize immunocompetent liver.
(A) Immunohistochemistry of athymic nude (left) and C57/B6 x 129S F_1 hybrid (right) livers stained for the macrophage marker F4/80.
(B) Quantification of F4/80+ cells per 10x field is shown as the median of n = 23 nude colonies and n = 14 C57/B6 x 129S F_1 hybrid colonies.
(C) Immunohistochemistry of athymic nude (left) and C57/B6 x 129S F_1 hybrid (right) livers stained for the T cell marker, Cd3.
(D) Quantification of Cd3+ cells per 10x field is shown (right) as the median of n = 24 nude colonies and n = 23 C57/B6 x 129S F_1 hybrid colonies.
(E) Log- standardized 10-cell sampling fluctuations of robust candidate heterogeneities identified by stochastic profiling.
(F) Venn diagram comparing the heterogeneous transcripts identified in **Figure 4.3B**, **Figure 4.5C**, and subpanel **E**. All two- and three-way intersections were significant ($p < 0.001$ by Monte-Carlo simulation).
(G) Regulatory heterogeneities with abrupt increases in abundance-adjusted variance in C57/B6 x 129S F_1 hybrid liver colonies. Stromal and neuroendocrine (NE) markers are highlighted.
(H) Log-standardized sampling fluctuations for the markers highlighted in **G.**
Enriched or unenriched coexpression was evaluated by hypergeometric test of 10-cell observations above their respective log means. For **B** and **D**, differences were assessed by rank-sum test. Scale bar in **A** and **C** is 80 μm.

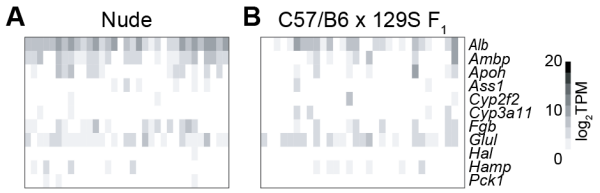

Figure 4.9 Liver contamination in KP1 samples from C57/B6 x 129S F_1 hybrid liver colonies is low and uncorrelated as in athymic nude liver colonies.
(A) Relative abundance of liver markers in 10-cell samples from nude liver colonies, reprinted from **Figure 4.5C** for comparison.
(B) Relative abundance of liver markers in 10-cell samples from C57/B6 x 129S F_1 hybrid liver colonies, column clustered as in **Figure 4.8E**.

4.3.5 Marker gene aberrations are partly retained in autochthonous SCLC tumors and metastases

The non-NE markers identified by stochastic profiling prompted a more-systematic evaluation of marker-gene status in 10-cell and bulk samples. For comparison, we used RNA-seq data from autochthonous tumors and metastases of *Rb1*$^{F/F}$*Trp53*$^{F/F}$*Rbl2*$^{F/F}$ mice administered AdCMV-Cre or adenoviral Cre driven the Calca promoter (AdCalca-Cre) (317). Curiously, for the ATII markers *Cd74* and *Lyz2*, the autochthonous samples indicated that abundance was higher in the primary tumor and reduced in the metastasis (**Figure 4.10A and B**, black squares vs. brown filled triangles). Similar results were obtained with the stromal markers, *Bgn* and *Sparc*, although the tumor-metastasis differences were less dramatic (**Figure 4.10C and D**). These observations are reconcilable with the 10-cell data if the spheroid observations are not taken as a proxy for the primary tumor. Rather, in vitro cultures reflect the SCLC states achievable from purely homotypic cell-cell interactions. Paracrine inputs from non-NE cells of the lung could just as feasibly drive SCLC reprogramming as non-NE cells of the liver.

Interestingly, abundance of the stromal marker *Mgp* was quite different between the two autochthonous GEMMs (**Figure 4.10E**). KP1 cells were isolated from an animal infected with AdCMV-Cre (179). The sporadic increases in *Mgp* abundance observed upon liver colonization were consistent with the other stromal markers found to be high in AdCMV-Cre tumors and metastases. By contrast, *Mgp* abundance in AdCalca-Cre-derived samples was uniformly low. AdCalca-Cre has been speculated to target a more-differentiated subset of PNECs compared to AdCMV-Cre (317). In support of this claim, we found that the KP1 gains in *Mgp* expression in vivo coincided with loss of endogenous *Calca* itself (**Figure 4.10E and F**). *Mgp* is an inhibitory morphogen for lung development (339) and its inducibility may mark the PNEC progenitor pool targeted by AdCMV-Cre.

In addition to *Calca*, other neuroendocrine markers (*Ascl1*, *Pcsk1*) declined substantially when KP1 cells were engrafted to the liver (**Figure 4.10G and H**). Yet, deprogramming appeared incomplete, as multiple neuroendocrine markers (*Uchl1*, *Resp18*) remained largely unchanged and at comparable abundance to autochthonous models (**Figure 4.10I and J**). Among the markers identified by a gradient of 10-cell dispersion (**Figure 4.8G**), several showed no discernible change in median abundance (**Figure 4.10K and L**) and thus would be impossible to identify in bulk samples. One of the transcripts correlating strongly with non-NE markers in PNECs (*Ldhb*) (170) recurred as a candidate heterogeneity in all three KP1 settings (**Figure 4.10M**). Lastly, we identified a characteristic non-NE marker (*Igfbp7*) (319) where both median abundance and dispersion increased specifically in immunocompetent livers (**Figure 4.10N**). Such miscoordination of markers could occur if SCLC cells fragmented their regulatory states upon encountering progressively more-diverse cellular microenvironments.

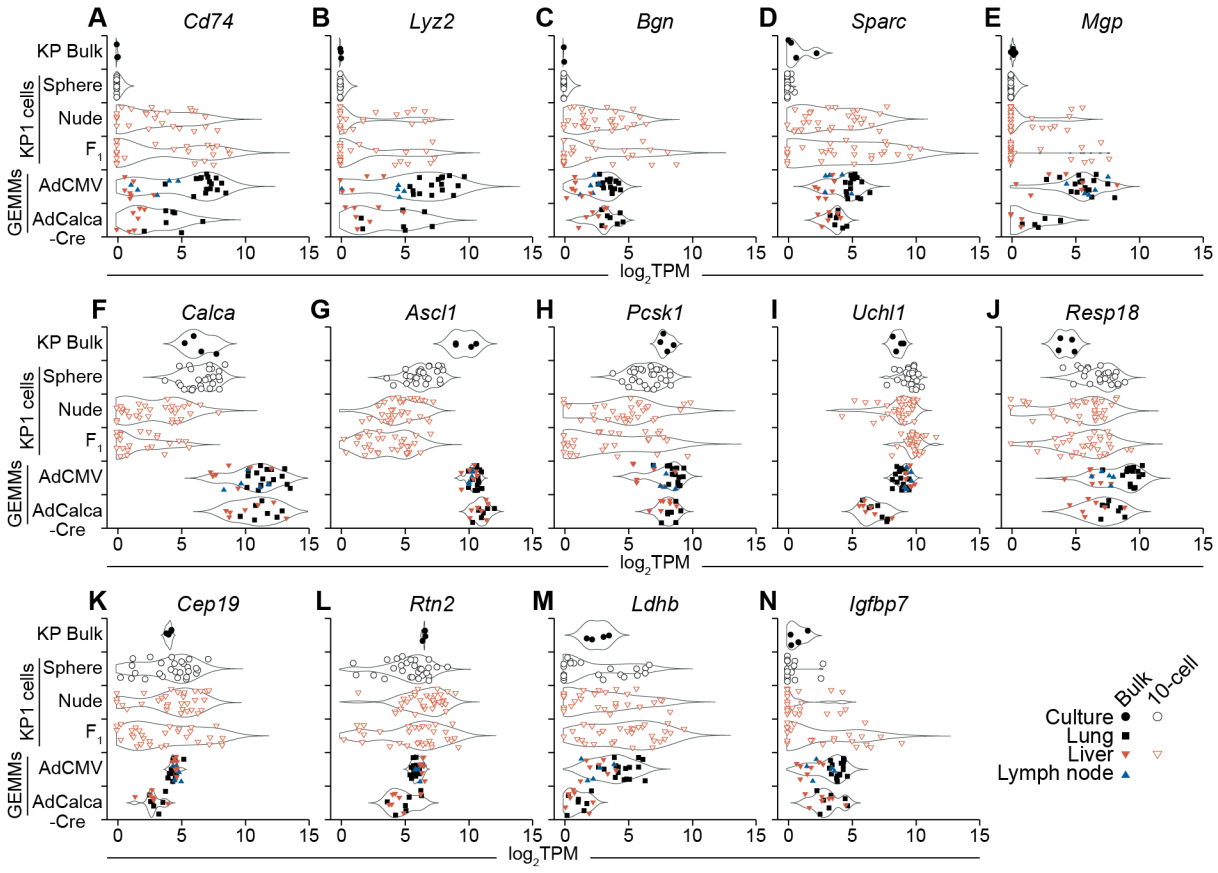

Figure 4.10 Comparison of 10cRNA-seq observations to bulk RNA-seq from primary SCLC tumors and metastases of SCLC GEMMs.
(A and **B)** Sporadic expression of the ATII markers *Cd74* (**A**) and *Lyz2* (**B**) upon liver colonization of KP1 cells compared to SCLC GEMM tumors and metastases.
(C–E) Abundance changes in the stromal markers *Bgn* (**C**), *Sparc* (**D**), and *Mgp* (**E**) in vivo.
(F–J) Reduced in vivo abundance of the neuroendocrine markers *Calca* (**F**), *Ascl1* (**G**), and *Pcsk1* (**H**), but not *Uchl1* (**I**) or *Resp18* (**J**).
(K and **L)** Context-dependent dispersion changes without abundance changes for the stromal marker *Cep19* (**K**) and the neuroendocrine marker *Rtn2* (**L**).
(M) The non-neuroendocrine marker *Ldhb* is a RHEG in KP1 cells.
(N) Heterogeneous regulation of the non-neuroendocrine marker *Igfbp7* in KP1 cells colonized to the liver of immunocompetent animals.

4.3.6 Mature Notch2 protein abundance is rapidly altered during KP1 cell dissociation

KP1 cells expressed multiple non-NE transcripts in the liver of C57/B6 x 129S F_1 hybrid mice (**Figure 4.8G and H**), and single-cell transcriptomics has associated non-NE changes with activation of the Notch pathway (170,173). Unexpectedly, despite measurable expression of *Notch2* by 10cRNA-seq (3.4 ± 9.4 TPM), we almost never detected the Notch target gene *Hes1* in vivo (0 ± 0.1 TPM). For normal PNECs, dedifferentiation to a non-NE state occurs during tissue damage, which may be mimicked by the cell-dissociation steps required for conventional single-cell expression profiling (170). Notch-pathway activation of cell lines also reportedly occurs during routine passaging (340), prompting us to ask whether such artifacts could arise in KP1 cells. Notch1 is nearly absent in the line (less than 0.5 TPM for *Notch1* vs. 29 TPM for *Notch2* in bulk; GSE147358), and reliable activation-specific antibodies for Notch2 are not available. Therefore, we used an antibody recognizing an intracellular epitope of full-length Notch2 and its processed transmembrane (NTM) subunit, which is the precursor for pathway activation (341). Within five minutes of KP1 dissociation using either trypsin or accutase, we noted considerable decreases in total Notch2 protein (full-length + NTM; **Figure 4.11A–C**). Furthermore, trypsin significantly increased the ratio of NTM-processed to full-length Notch2 ($p < 0.01$ by ANOVA; **Figure 4.11D**), suggesting that trypsinized cells may be more primed to activate Notch signaling. Our results support earlier speculation (170) that Notch activation in PNEC-like cells may be an artifact of the sample processing that precedes scRNA-seq but is avoided by 10cRNA-seq (180).

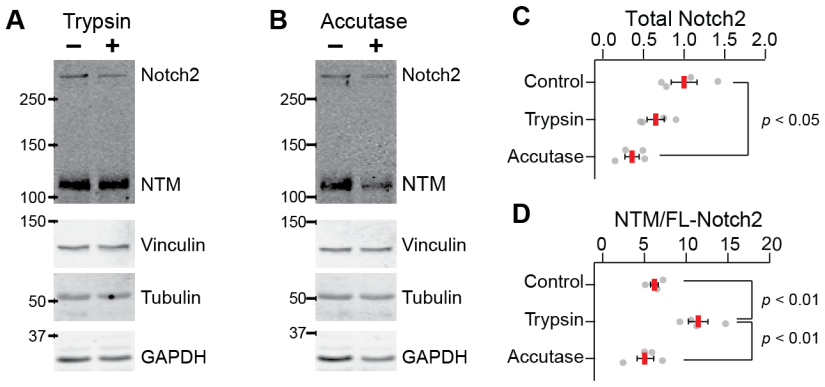

Figure 4.11 Cell-dissociation enzymes rapidly disrupt intracellular precursors of Notch2 signaling in KP1 cells.
(A and **B)** Immunoblots of full-length Notch2 and the processed Notch transmembrane (NTM) subunit in KP1 cells after treatment of 0.05% trypsin or 1x accutase for five minutes. Vinculin, tubulin, and GAPDH were used as loading controls.
(C) Relative abundance of total Notch2 (full-length + NTM) for the indicated conditions. Data are normalized to control KP1 cells lysed without dissociation.
(D) Ratiometric abundance of NTM / full-length (FL) Notch2 for the indicated conditions.
For **C** and **D**, data are shown as the mean ± s.e.m. from $n = 4$ independent biological samples. Differences in means were assessed by ANOVA with Tukey HSD post-hoc test.

4.3.7 Human SCLCs are merged or stratified by different classes of KP1 RHEGs

We returned to two statistically significant overlaps from the three studies in KP1 cells (**Figure 4.8F**). The three-way intersection of 26 transcripts defined a core group of RHEGs, which we viewed as a set of cell-autonomous heterogeneities intrinsic to KP1 cells and perhaps SCLCs more generally. We tested this concept by identifying the human orthologs of the KP1 core RHEG set and clustering our data alongside bulk RNA-seq profiles from 79 cases of SCLC in humans (342). The standardized fluctuations of the core RHEGs in human samples were largely indistinguishable from the KP1 observations, with most sample co-clusters containing mouse and human data (**Figure 4.12A**). Moreover, when the pairwise correlations of core RHEGs were organized hierarchically, it was difficult to discern any strongly linked groups of observations (**Figure 4.12B**). This would be expected if core RHEGs were broadly but independently “active” (induced heterogeneously). Accordingly, we found very little evidence of coordination outside a small row cluster of genes involved in biological processes that were largely unrelated—cell cycle-dependent ubiquitination (*CCNF*), carbonyl stress (*HAGH*), splicing (*SNRNP200*), calcium homeostasis (*CHERP*), and DNA methylation (*MBD1*) (**Figure 4.12A**). Although the existence of core RHEGs in mammalian SCLCs awaits direct testing in human samples, the analysis here provides a GEMM-informed set of targets worth examining further.

The second overlap of interest was the two-way intersection of 149 genes that emerged as candidate heterogeneities in both settings of liver colonization (**Figure 4.8F**). We defined these in vivo RHEGs as reflecting the SCLC regulatory heterogeneity triggered by heterotypic cell-cell interactions in the microenvironment. In contrast to core RHEGs, we expected different activation patterns of in vivo RHEGs in the liver versus the lung, and even among different SCLC subtypes or primary-tumor sites in the lung. We extracted human orthologs of the in vivo RHEGs and clustered the KP1 observations together with the human SCLCs (**Figure 4.12C**). There was far less intermixing between human and KP1 samples, consistent with the different

heterotypic interactions anticipated between primary and metastatic sites. The pairwise correlation structure of in vivo RHEGs was also qualitatively distinct, with multiple groups of covariates comprised entirely of human SCLCs (**Figure 4.12D**). Importantly, these clusters each contained mixtures of various SCLC subtypes based on the relative abundance of key transcription factors (175) (**Figure 4.13**). The KP1 observations in the liver reflect only one of four SCLC subtypes and do not precisely capture human variation in the lung. However, the in vivo RHEG set derived from those observations may stratify clinical cases by differences in tumor ecosystem.

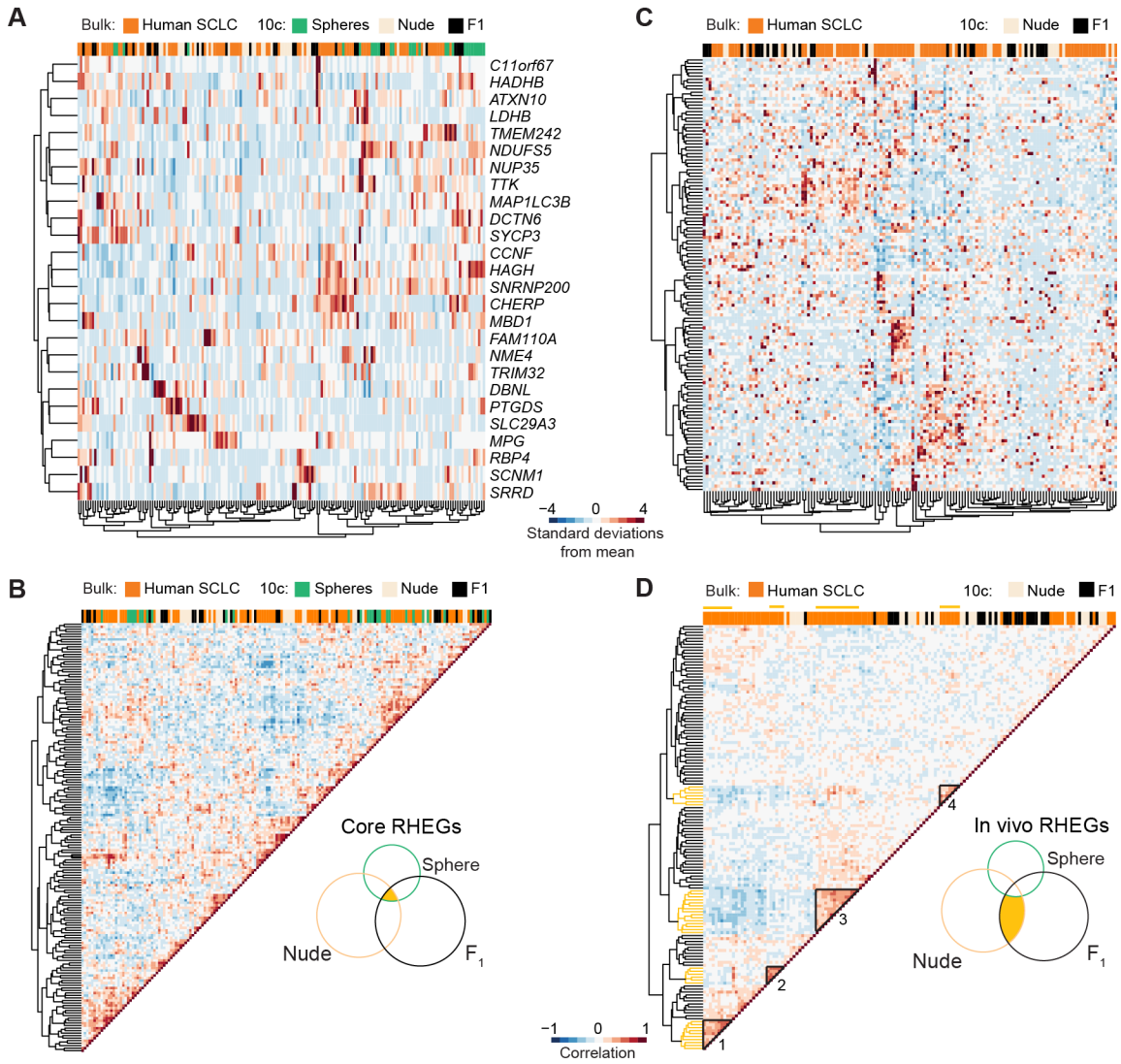

Figure 4.12 Orthologous RHEG fluctuations in primary human SCLCs.
(A) Core RHEGs and their orthologs intermix KP1 10-cell observations and bulk RNA-seq data from human cases of SCLC (342).
(B) Pearson correlation matrix for core RHEGs clustered hierarchically. The Venn diagram intersection for core RHEGs is highlighted from **Figure 4.8F**.
(C) In vivo RHEGs and their orthologs do not merge KP1 and human observations but identify subgroups of clinical SCLCs.
(D) Pearson correlation matrix for in vivo RHEGs clustered hierarchically. Groups of covarying human SCLC cases are indicated in black triangles and yellow margins and numbered as in **Figure 4.13**. The Venn diagram intersection for core RHEGs is highlighted from **Figure 4.8F**. Murine data and human data were standardized separately by z-score before clustering or correlation.

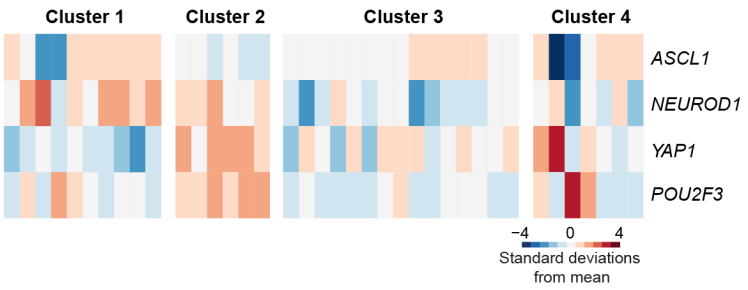

Figure 4.13 In vivo RHEG clusters of human SCLC are not entirely explained by known SCLC subtypes.
The SCLC-A, SCLC-N, SCLC-Y, and SCLC-P subtypes are based on the above transcription factor abundances (175). Cluster numbers are as in **Figure 4.12D**.

4.4 Discussion

PNECs are a particularly versatile cell type (170), and it is perhaps unsurprising that derivative SCLC cells show the deranged plasticity reported here. It is less obvious whether dispersed SCLC states are engaged hierarchically or chaotically—our work with a representative GEMM-derived SCLC line argues for the former. Cell-autonomous regulatory heterogeneities expand qualitatively in vivo through heterotypic cell-cell interactions absent from in vitro culture. The documented cell-state changes upon liver colonization could simply reflect the injury-like state of tumors and metastases (309). Alternatively, the reprogramming events could provide trophic support to the cellular ecosystem (173,305). The candidate heterogeneities identified by stochastic profiling and 10cRNA-seq create a resource to guide future functional studies that perturb specific emergent heterogeneities in vivo.

The KP1 results with Notch2 reinforce that SCLC cells are very sensitive to juxtacrine inputs (173). SCLC tumorsphere growth in vitro elicits its own cell-to-cell heterogeneities, which have some commonalities with spheroids of MCF10A-5E basal-like breast cells, a distant epithelial cell type. Intrinsic to spheroid culture are subclonal reorganization and competition, two processes important for primary tumor initiation and the end stages of metastatic colonization. Cell crowding and sequestration alter lipid metabolism (343,344), which could explain the catabolic and anabolic lipid enzymes identified within the spheroid RHEG set. The notion of spheroid RHEGs may generalize to clonogenic soft-agar assays of anchorage-independent growth, which remain widely used as surrogates for tumorigenicity (345).

The candidate regulatory heterogeneities identified in KP1 liver colonies reflect several of the deprogramming and reprogramming events recently described in PNECs (170). In addition, they suggest routes of paracrine communication that are equally realistic for the lung as for the liver. From this perspective, the stratification of primary human SCLCs by in vivo RHEGs is intriguing. SCLCs usually initiate in the bronchi, but there are differences in cell

composition at different depths of the lung (346) as well as lobular biases in the primary sites typical for SCLC (347). Different SCLC subtypes (175) might arise in similar microenvironments, yielding the mixed-subtype clusters identified here. The stromal heterogeneities induced by the immunocompetent setting may also relate to fibrotic lung diseases, where PNECs hyperplasia is known to occur (348). The genome of SCLCs is known to be highly mutated (6), but our study indicates that cell-fate variability arises on a much faster time scale in vivo.

4.5 Materials and methods

4.5.1 Cell and tissue sources

KP1 cells (179) were cultured as self-aggregating spheroids in RPMI medium 1640 (Gibco) with 10% FBS, 1% penicillin-streptomycin, and 1% glutamine. There was no cell-line authentication, and cells were not tested for mycoplasma contamination. KP1-GFP cells were prepared by transducing cells overnight with saturating lentivirus and 8 µg/ml polybrene as previously described (201). GFP-encoding lentivirus was prepared with pLX302 EGFP-V5 cloned by LR recombination of pLX302 (Addgene #25896) and pDONR221_EGFP (Addgene #25899). Stable transductants were selected with 2 µg/ml puromycin until control plates had cleared. Cultured KP1-GFP spheroids were kept to within 10 passages and cryoembedded as described previously (180).

To seed liver colonies, KP1-GFP spheroids were dissociated with 0.05% Trypsin/EDTA (Life Technologies), counted using a hemocytometer, and $2x10^5$ cells were injected via the tail vein of athymic nude (Envigo) or C57/B6 x 129S F_1 hybrid strain of mice (Jackson laboratory). Animals were not randomized. Liver colonies were resected after ~30 days and immediately cryoembedded in NEG-50, frozen in a dry ice-isopentane bath, and stored at –80˚C (180). KP1 spheroids were cryosectioned at –24˚C and liver colonies were cryosectioned at –20˚C, both at 8 µm thickness as previously described (180). All mice were maintained according to practices

prescribed by the National Institutes of Health in accordance with the IACUC protocol #9367. All animal procedures were approved by the Animal Care and Use Committee at the University of Virginia, accredited by the Association for the Assessment and Accreditation of Laboratory Animal Care (AAALAC).

4.5.2 Fluorescence-guided LCM

KP1-GFP sections were fixed and dehydrated with ethanol and xylene as described previously for fluorescent cryosections (180). Freshly fixed samples were immediately microdissected on an Arcturus XT LCM instrument (Applied Biosystems) using Capsure HS caps (Arcturus). The smallest spot size on the instrument captured 3–5 SCLC cells per laser shot.

4.5.3 RNA extraction and amplification

RNA extraction and amplification of microdissected samples was performed as described previously to minimize contaminating genomic amplification (180). Briefly, biotinylated cDNA was synthesized from RNA eluted from captured cells and purified with streptavidin magnetic beads (Pierce) on a 96S Super Magnet Plate (Alpaqua). Residual RNA was degraded with RNAse H (NEB), and cDNA was poly(A) tailed with terminal transferase (Roche). Poly(A)-cDNA was amplified using AL1 primer (ATTGGATCCAGGCCGCTCTGGACAAAATATGAATTCTTTTTTTTTTTTTTTTTTTTTT) and a blend of Taq polymerase (NEB) and Phusion (NEB) for 25 cycles. RNA from bulk KP cell lines was isolated by RNEasy kit (Qiagen).

4.5.4 10-cell sample selection by quantitative PCR (qPCR)

Detection of transcripts by qPCR was performed on a CFX96 real-time PCR instrument (Bio-Rad) as described previously (251). 0.1 µl of preamplification material was used in the qPCR reaction. For each sample, we quantified the expression of *Gapdh* and *Rpl30* as loading controls. Samples were retained if geometric mean quantification cycle of *Gapdh–Rpl30* was within 3.5x interquartile range of the median; samples outside that range were excluded

because of over- or under-capture during LCM. For liver colonies, we also excluded samples with detectable quantification cycles of three high-abundance hepatocyte markers: *Alb*, *Fgb*, and *Cyp3a11*.

4.5.5 Library preparation

Ten-cell sequencing libraries were prepared by reamplification, purification, and tagmentation as described previously (180). Briefly, each poly(A) PCR cDNA sample was re-amplified by PCR within its exponential phase (typically 10 to 20 cycles). Re-amplified cDNA was then twice purified with Ampure Agencourt XP SPRI beads, and samples were quantified on a CFX96 real-time PCR instrument (Bio-Rad) using a Qubit BR Assay Kit (Thermo Fisher). Samples were diluted to 0.2 ng/μl and tagmented with the Nextera XT DNA Library Preparation Kit (Illumina). Bulk KP libraries were prepared from 500 ng of total RNA by the Genome Analysis and Technology Core at the University of Virginia using mRNA oligo dT-purified with the NEB Next Ultra RNA library preparation kit (NEB).

4.5.6 RNA sequencing

10cRNA-seq data were sequenced and aligned as previously described (180). Ten-cell samples were multiplexed at an equimolar ratio, and 1.3 pM of the multiplexed pool was sequenced on a NextSeq 500 instrument with NextSeq 500/550 Mid/high Output v1/v2/v2.5 kits (Illumina) to obtain 75-bp paired-end reads. Bulk KP RNA samples were sequenced on a NextSeq 500 to obtain 50-bp single-end reads. Adapters were trimmed using fastq-mcf in the EAutils package (version ea-utils.1.1.2-779) with the following options: -q 10 -t 0.01 -k 0 (quality threshold 10, 0.01% occurrence frequency, no nucleotide skew causing cycle removal). Quality checks were performed with FastQC (version 0.11.8) and multiqc (version 1.7). Mouse datasets were aligned to the mouse transcriptome (GRCm38.82), reference sequences for ERCC spike-ins, and pLX302-EGFP by using RSEM (version 1.3.0) and Bowtie 2 (version 2.3.4.3). RSEM processing of the 10cRNA-seq data also included the following options: --single-cell-prior --paired-end. Counts from RSEM processing were converted to transcripts per million (TPM) by

dividing each value by the total read count for each sample and multiplying by 10^6. Total read count for TPM normalization did not include mitochondrial genes or ERCC spike-ins.

4.5.7 RNA FISH

A 150-bp fragment of human *SOX4* was cloned into pcDNA3, used as a template for in vitro transcription of a digoxigenin-labeled riboprobe for RNA FISH, and imaged as previously described (129). Loading-control riboprobes for *GAPDH*, *HINT1*, and *PRDX6* were previously reported (201).

4.5.8 Immunohistochemistry

Staining was performed by the Biorepository and Tissue Research Facility at the University of Virginia with 4 μm paraffin sections. For F4/80, antigen retrieval and deparaffinization were performed in PT Link (Dako) using low pH EnVision FLEX Target Retrieval Solution (Dako) for 20 minutes at 97°C. Staining was performed on a robotic platform (Autostainer, Dako). Endogenous peroxidases were blocked with peroxidase and alkaline phosphatase blocking reagent (Dako) before incubating the sections with F4/80 antibody (AbD Serotech, #MCA497R) at 1:200 dilution for 60 minutes at room temperature. Antigen–antibody complex was detected by using rabbit anti-rat biotin and streptavidin-HRP (Vector Laboratories) followed by incubation with 3,3'-diaminobenzidine tetrahydrochloride (DAB+) chromogen (Dako). For Cd3, sections were deparaffinized using EZ Prep solution (Ventana), and staining was performed on a robotic platform (Ventana Discover Ultra Staining Module). A heat-induced antigen retrieval protocol set for 64 min was carried out using Cell Conditioner 1 (Ventana). Endogenous peroxidases were blocked with peroxidase inhibitor (CM1) for 8 minutes before incubating the section with CD3 antibody (Dako, #A0452) at 1:300 dilution for 60 minutes at room temperature. Antigen-antibody complex was detected using DISCOVERY OmniMap anti-rabbit multimer RUO detection system and DISCOVERY ChromoMap DAB Kit (Ventana). All slides were counterstained with hematoxylin, dehydrated, cleared, and mounted for assessment. For both F4/80 and Cd3 stains, cells were counted visually and reported as the

average of multiple 10x-field images surrounding individual KP1 cell colonies in liver sections obtained from athymic nude and C57/B6 x 129S F_1 hybrid mice.

4.5.9 Immunoblot analysis

Quantitative immunoblotting was performed as previously described (202). Primary antibodies recognizing the following proteins or epitopes were used: Notch2 (Cell Signaling #5732, 1:1000), vinculin (Millipore #05-386, 1:10,000), GAPDH (Ambion #AM4300, 1:20,000), tubulin (Abcam #ab89984, 1:20,000).

4.5.10 Mouse-to-human ortholog mapping

Human orthologs for mouse genes were obtained from the Ensembl biomart in R using the getAttributes function. For genes with multiple human-ortholog mappings, we used expression characteristics of the human datasets considered [MCF10A-5E (180) and human SCLC (342)] to favor more-reliable clustering afterwards. For mouse genes with two human mappings, the human ortholog with higher expression variance in the corresponding human dataset was retained. For mouse genes with greater than two human mappings, two orthologs with the highest expression correlation were identified. From these, the ortholog with the higher expression variance was retained, as in the two-mapping case. Any remaining mouse gene names were capitalized in accordance with human gene symbol conventions.

4.5.11 Overdispersion-based stochastic profiling

Stochastic profiling with 10cRNA-seq data was performed exactly as described in Chapter 3.

4.5.12 Robust identification of transcriptional heterogeneities through subsampling

To minimize the contribution of outliers to the overdispersion analysis in samples collected from liver colonies, we generated 100 subsampled simulations for overdispersion-based stochastic profiling. After sample selection (see above), there were 33 10-cell samples plus 35 pool-and-split controls for nude liver colonies and 31 10-cell samples plus 24 pool-and-split controls for C57/B6 x 129S F_1 hybrid liver colonies. For each dataset, overdispersion-

based stochastic profiling was performed 100 times with random downsampling to 28 10-cell samples and 20 pool-and-split controls, as in Chapter 3. Only genes that recurred as candidates in >75% of simulations were evaluated further as candidate heterogeneously expressed genes.

4.5.13 Filtering out hepatocyte contamination in heterogeneous expressed genes

Among overdisperse transcripts in liver colonies, we further excluded genes that might vary because of residual hepatocyte capture during LCM. For each 10-cell sample, we calculated the geometric mean abundance of 11 liver-specific markers (*Alb*, *Fgb*, *Cyp3a11*, *Ambp*, *Apoh*, *Hamp*, *Ass1*, *Cyp2f2*, *Glul*, *Hal*, and *Pck1)* from published studies (332,349–354). Candidates that were significantly correlated with the mean liver signature ($p < 0.05$ by Fisher Z-transformed Spearman ρ correlation) were removed from further consideration for the in vivo study.

4.5.14 Continuous overdispersion analysis

Overdispersion values from the 2007 transcripts identified as candidate heterogeneities in either the KP1 spheres or in vivo conditions were recorded for 100 subsampling iterations. For the KP1 in vitro spheroids, 100 iterations of leave-one-out crossvalidation were performed as detailed in co-submitted work (271). Transcripts were retained if the 5th percentile of overdispersion in the C57/B6 x 129S F_1 hybrid condition was greater than the 95th percentiles of the other two conditions. If a gene was not expressed in a condition, the 5th and 95th percentiles were set to zero, and the gene was assigned to the overall median overdispersion during clustering.

4.5.15 Statistics

Sample sizes for stochastic profiling were determined by Monte Carlo simulation (157). Significance of overlap between candidate genes in KP1 spheroids and MCF10A-5E spheroids was evaluated using the hypergeometric test using the “phyper” function in R and a background of 20,000 genes. Pearson correlation between pairs of transcripts detected in both KP1 and MCF10A-5E spheroids were assessed using the “cor.test” function. Significant increases in

number of candidate genes between different conditions were assessed by the binomial test using the “binom.test” function in R. Spearman ρ correlation between overdisperse transcripts and liver markers was calculated using the “cor.test” function. Spearman ρ correlations were Fisher Z-transformed using the “FisherZ” function from the R package “DescTools” (version 0.99.31). Co-occurrence of transcript fluctuations was evaluated by hypergeometric test after binning 10cRNA-seq above or below the geometric mean of the two transcripts compared. Differences in cell number by immunohistochemistry were assessed by the Wilcoxon rank sum test using “wilcox.test”. Significance of overlaps between candidate genes identified in spheroids, nude mice, and C57/B6 x 129S F_1 hybrid mice were assessed by Monte Carlo simulations and corrected for multiple hypothesis testing as described in Chapter 3. Significance of differences in protein abundance by immunoblotting were assessed by one-way ANOVA with Tukey HSD post-hoc test. Hierarchical clustering was performed using “pheatmap” with standardized values, Euclidean distance, and “ward.D2” linkage or non-standardized values, Pearson distance, and “ward.D2” linkage. Gene set enrichment analyses were performed through the Molecular Signatures Database (355). Overlaps between gene lists and hallmark gene sets were computed using a hypergeometric test with false-discovery rate correction for multiple comparisons.

4.5.16 Data availability

Bulk and 10cRNA-seq data from this study is available through the NCBI Gene Expression Omnibus (GSE147358, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147358 Reviewer token: ufczqoiwzbwjnet). Other RNA-seq datasets were obtained from the Gene Expression Omnibus (GEO): MCF10A-5E 10cRNA-seq (GSE120261), AdCMV-Cre and AdCalca-Cre GEMM (GSE116977), and human SCLC tumor (GSE60052).

5 Discussion and Future Directions

5.1 Dissertation discussion

This dissertation focused on characterizing recurrent regulatory variations between epithelial cancer cells in multiple tumor models and microenvironments. We presented 10cRNA-seq - a method for obtaining in situ transcriptomic measurements from 10-cell samples isolated from tumor samples (Chapter 2). We extended the theory of stochastic profiling (Chapter 1) to 10cRNA-seq measurements using an abundance-dependent dispersion statistic (Chapter 3). We combined 10cRNA-seq and stochastic profiling to identify early-stage regulatory variations in cancer cells in clinical cases of luminal biopsies (Chapter 3). To systematically measure microenvironmental influences on regulatory variations of cancer cells, we applied 10cRNA-seq profiling to *in vitro* and murine models of small cell lung cancer (Chapter 4). In Chapters 3 and 4, we identified **r**ecurrent **h**eterogeneously **e**xpressed **g**enes (RHEGs) to focus on generalizable markers of cancer cell variations. For luminal breast cancers profiled in Chapter 3, we defined RHEGs as genes identified to be significantly overdispersed in the majority of tumors sampled. We found that luminal breast cancer RHEGs did not reflect common sources of cell-state variations, but were instead enriched for genes known to be drivers of multiple tumor types. In SCLC cells and metastases in Chapter 4, we defined different types of RHEGs to decode the relative influences of different microenvironments on cancer cell states. Comparing *in vitro* RHEGs to *in vivo* RHEGs we identified SCLC cell-intrinsic regulatory variations, as well as an expanded set of regulatory variations that arise from cell extrinsic influences. The next two subsections discuss the impact and relevance of our approach and findings.

5.1.1 Implications of cancer cell heterogeneity for treatment response

Both studies of cancer cell heterogeneity presented in this dissertation assessed samples prior to any treatment, raising questions about how the identified regulatory variations would influence treatment response. Individual cancer cells within tumors are either intrinsically drug resistant or able to regulate their cell-states to acquire resistance. Intrinsic treatment resistant populations marked by high expression of *AXL* have been identified in many melanoma tumors as a route to disease recurrence (20,104). In a recent study of triple-negative breast cancer, post-treatment cancer cells had high expression of growth-supporting cancer genes like *MYC,* which were also detected in a small proportion of cells in the same patient prior to treatment (19). However, pre- and post-treatment gene expression patterns in triple-negative breast cancer were highly patient specific, limiting the ability to make generalizable conclusions (19).

RHEGs identified in luminal breast cancer in Chapter 3 may present genes that are markers of intrinsic drug-resistant cell-states across multiple tumors (**Figure 5.1A**). This hypothesis is supported by the enrichment for both pan-EMT and pan-cancer driver genes in luminal breast cancer RHEGs (**Figure 3.13**), as alterations in these genes have been shown to support tumor growth in numerous other cancers (260–262,280). Multiple luminal RHEG drivers –*RET, PRKCZ,* and *MLST8*–converge upon AKT/mTOR related growth signaling, a pathway that is frequently activated as a route to both hormone and chemotherapy resistance in breast and ovarian tumors (356–360). A luminal cancer RHEG driver, *COL1A1*, was also detected as a potential marker of intrinsic treatment resistance in one triple-negative breast cancer tumor in the study described above (19). Further, partial mesenchymal states reflected by RHEG EMT markers in Chapter 3 (**Figure 3.13** and **Figure 3.14**), and *in vivo* RHEG stromal markers in Chapter 4 (**Figure 4.8**), have been associated with multiple drug resistance in epithelial tumors of the lung and pancreas (280,361–363). RHEG driver genes in luminal cancer also include

transcriptional regulators including *KAT8*, a histone acetyltransferase, indicating specific cancer cells that are differentially poised to activate survival pathways in response to treatment (364–366). Combination therapies with drugs that target histone modifiers have shown efficacy in models of both breast and lung cancer (316,367–369).

We observed the ability of cancer cells to differentially alter their regulatory states in Chapter 4 in response to heterotypic interactions and microenvironments. KP1 cells acquired heterogeneous expression of genes *Cd74*, *Lyz2*, *Bgn*, *Sparc*, and *Mgp* (**Figure 4.10**) after colonizing the liver. Similar phenotypic plasticity of SCLC cells could occur in response to drug treatments, leading to acquired drug resistance (**Figure 5.1B**). EMT-related gene upregulation was recently uncovered as a form of acquired therapy resistance in SCLC (109), indicating that despite having different selective pressures, the metastatic process and drug treatment may share regulators of phenotypic plasticity. Bioinformatically predicted transcription factors that regulate *in vivo* RHEGs include IRF1, SP1, and E2F1, which have all been implicated in chemoresistance in SCLC (370).

Reversible, dynamic responses enable additional strategies for mitigating resistance by designing dosing schedules that include drug-holidays to allow cells to revert to non-resistance states (371). Clinically, this would be immensely useful for SCLC treatment, where tumors start as chemo-sensitive but become chemo-resistant to result in ~5% overall survival (171,175). To tackle heterogeneity in treatment response, combination therapies are a promising strategy if candidate targets for treatment response can be evaluated a priori. To fully test this, matched measurements and functional tests of specific genes pre- and post-treatment are necessary and are discussed later in this chapter.

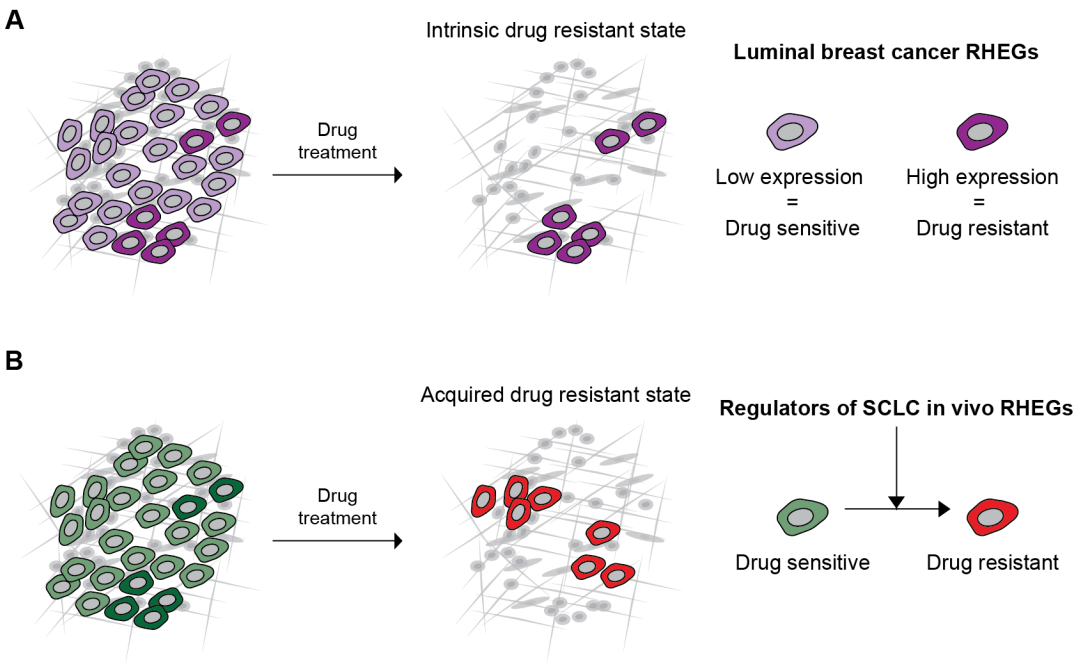

Figure 5.1 Potential roles for breast and lung cancer RHEGs in treatment resistance
(A) Schematic for intrinsic drug resistance to treatments. Luminal breast cancer RHEGs represent pre-treatment cell-states to be tested for intrinsic drug resistance.
(B) Schematic for acquired drug resistance to treatments. SCLC in vivo RHEGs represent differential epigenetic reprogramming capacity. Upstream regulators could represent mediators of acquired drug resistance.

5.1.2 Convergence of regulatory variations across multiple tumors and models

Another important question that remains is whether our approaches identify epithelial cancer cell variations that are shared across multiple tumor types (372). While luminal breast cancers and small cell lung cancer differ in cancer-initiating driver mutations (Chapter 1), it is possible that they may share regulatory variations to give rise to shared phenotypes, such as the hallmarks of cancer identified across human tumors (3,372,373). In Chapter 4 we observed cross-species and cross-tissue convergence in regulatory variations when comparing *in vitro* 3D cell cultures of KP1 cells (SCLC) and MCF10A-5E cells (breast epithelia). Coordination was observed for the same genes involved in protein transport and nuclear shuttling as well as cellular longevity in both sets of samples (**Figure 4.2**). A starting point for evaluating shared regulatory variations of SCLC cells with luminal breast cancer cells is to compare different categories of RHEGs identified in Chapters 3 and 4.

In Chapter 4, we identified a set of core RHEGs that were heterogeneously expressed in SCLC cells in all 3 evaluated contexts (**Figure 4.8**). Of these 26 core RHEGs, 5 genes were shared with the luminal breast cancer RHEGs identified in Chapter 3 (**Figure 5.2A**). However, core RHEGs include contributions from *in vitro* measurements, and may not reflect heterotypic interactions experienced *in vivo*. In Chapter 4, we observed that *in vivo* RHEGs stratified human cases of SCLC while core RHEGs did not. When compared to 149 *in vivo* RHEGs, UVABC RHEGs shared 13 genes that spanned diverse functions (**Figure 5.2B**). One of these genes is *CD74*, which encodes for the invariant chain of the MHC Class II molecule (374). Initially thought to only be expressed on antigen presenting cells like macrophages and dendritic cells, many studies have demonstrated CD74 expression in many other cell types (375). CD74 expression has been detected in epithelial cells at baseline (**Figure 5.2C**), with increased expression during inflammation (174,376), as well as in multiple tumor types in both human and murine studies (377–379). The protein CD74 is a cell surface receptor for its ligand,

macrophage inhibitory factor (*MIF*). Upon MIF binding, the CD74 receptor activates growth and proliferation through the MAPK pathway, leading to increased phagocytic activity in macrophages (380). In tumor cells, signaling through cell-surface CD74 causes increased growth and improved cancer cell survival (379,381). Cell-to-cell heterogeneity in CD74 protein expression in observed in normal tissue samples (**Figure 5.2C**), indicating that expression variation of CD74 is likely due to differential regulation and not genetic alteration. An immediate next step is to test the functional role of cancer cell expression of CD74 in our model of SCLC (section 5.3).

All shared genes that display recurrent heterogeneous expression in both studies have to be evaluated for technical and biological explanations (**Figure 5.2A and B**). Cross-study recurrent genes could represent the detection of similar cell-states across multiple epithelial cell types, enabled by our deep interrogation of carcinoma cells. Alternatively, the heterogeneous detection of these genes could be due to specific sequence features or genomic structure that causes measurement artefacts, and that many studies might find these genes to be heterogeneously expressed. Such cross-study differential expression has been observed in bulk studies, where several genes are always detected as differentially expressed regardless of the hypothesis being tested (382). Identifying such genes in 10cRNA-seq data could help refine future analytical approaches for evaluating transcriptional heterogeneity.

Previous studies have discovered that cancer cells converge on similar phenotypes by regulating expression of different sets of genes, depending on the mutational and epigenetic landscape of the cancer type (124,383). In addition to comparing specific marker genes of cell-states as done here, future comparisons made by grouping genes into their signaling pathways would be another way to assess generalizability across cancer types (373).

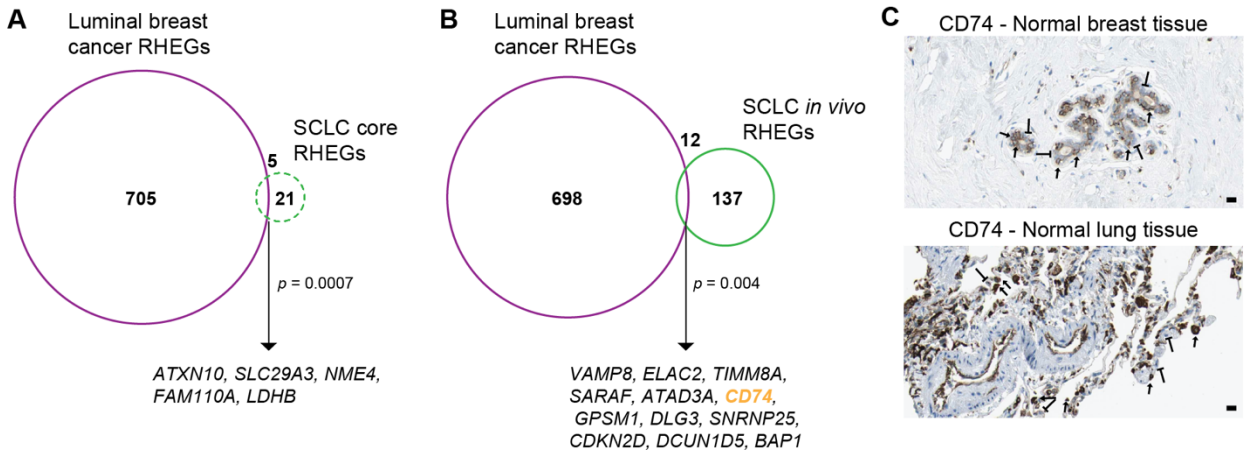

Figure 5.2 Shared RHEGs provide insight into regulatory variations across cell types
(A) Intersection of five shared RHEGs between luminal breast cancers profiled in Chapter 3 and all SCLC conditions profiled in Chapter 4.
(B) Intersection of twelve shared RHEGs between luminal breast cancers profiled in Chapter 3 and *in vivo* SCLC samples profiled in Chapter 4 includes CD74.
(C) Immunohistochemistry from Human Protein Atlas (HPA) showing cells with no CD74 expression (flat arrows) adjacent to cells that display expression of CD74 (arrowheads) in normal tissue from breast (upper) and lung (lower). Scale bar is 25μm.
For **(A), (B)**, statistical significance was assessed using the hypergeometric test.

5.2 Future studies of RHEGs identified in Chapters 3 and Chapter 4

In Chapter 3 we profiled five cases of luminal breast tumor biopsies and identified thousands of heterogeneously expressed genes within individual tumors (**Figure 3.10**). While these individual datasets have several remaining opportunities for computational analysis, the next few sections focus on validating and testing a prioritized list of RHEGs that emerged across multiple luminal breast tumors.

Similarly, we will focus on SCLC RHEGs that were shared amongst the two *in vivo* liver colonies profiled in Chapter 4. I will present plans for testing the phenotypic effects and regulation of *in vivo* RHEGs to understand their implications for disease progression.

5.2.1 Experimental tests for luminal breast cancer RHEG drivers

Luminal breast cancer RHEGs had a significant overlap with genes identified as drivers for many cancer types (**Figure 3.13**). We identified 46 “RHEG drivers” that comprise a diverse array of functions including growth signaling, stress tolerance, and transcriptional regulation (355). The next two subsections discuss approaches to further validate and characterize the functional value of these RHEG drivers and their protein products.

5.2.1.1 Validating heterogeneous expression of RHEG drivers

Heterogeneous gene predictions by stochastic profiling have been experimentally validated by RNA FISH in matched samples in Chapter 4 (**Figure 4.3**) and published studies (76,129,158). Therefore, we expect experimental measurements of RNA transcripts to have a high rate of validation for RHEGs. To relate gene-expression based predictions to cellular phenotype, the next step is to validate RHEG driver expression heterogeneity at the protein level.

A component of the mTOR signaling complex, *MLST8* was a recurrent heterogeneity in all 5 UVABC tumors profiled in Chapter 3 (**Figure 5.3A**) (358,384). I observed heterogeneity of MLST8 protein expression in immunohistochemistry measurement of an ER+ breast tumor available through the Human Protein Atlas (HPA) (**Figure 5.3B**).The HPA has IHC data for 44 of the 46 proteins encoded by RHEG drivers in normal breast tissue and breast tumors, however, comparing bulk changes in these proteins showed no consistent trends (**Figure 5.3C**), supporting a need to evaluate these proteins at the single-cell level in cancers. A next step is to obtain high-resolution images of IHC data, identify and digitally segment individual cancers cells to quantify intra-sample variation in RHEG driver expression in both normal and tumor tissues.

Additionally, validated antibodies obtained through the HPA enable us to measure single-cell coordination of RHEG driver proteins in matched UVABC samples. In a pilot study, I confirmed that immunofluorescence can be performed on cryosections of patient tissue used for 10cRNA-seq in Chapter 3 (**Figure 5.4**, see Materials and Methods). Since the *MLST8* gene is heterogeneously expressed in all 5 tumors, I would expect MLST8 protein expression to confirm this pattern. Next, I would prioritize multi-color immunofluorescence measurements of co-varying RHEGs (*CDKN2D*, *KLF4*, and *CDKN1A*; *TP73* and *MAD1L1*; *NFATC4* and *TNFSF10*; *GDF15* and its receptor *RET*) and quantify degree of single-cell covariation through quantitative image processing. Single-cell coordination of protein expression would confirm that co-varying transcripts detected in 10cRNA-seq data reflect differentially regulated expression states.

Protein expression studies will validate RHEG classification, but they will not provide further information regarding the mechanism of heterogeneous regulation. While we ruled out large scale copy number changes as a cause of RHEG classification (**Figure 3.9**), ruling out single nucleotide variants is impossible due to lack of full-length reads in 10cRNA-seq data. It remains possible that heterogeneous expression of RHEG driver genes is the result of mutational changes. To test mechanisms of regulation and functional consequences of RHEGs

beyond the UVABC cohort, appropriate experimental models are necessary. The next section will discuss testing RHEG predictions with experimental models.

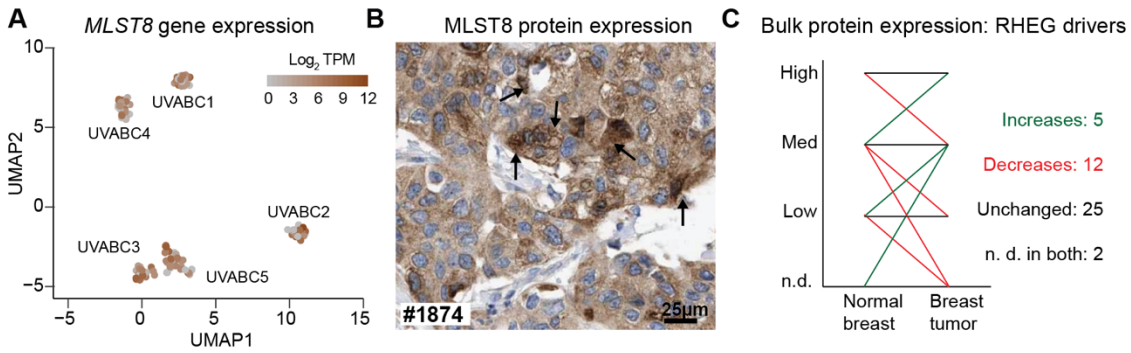

Figure 5.3 Variable protein expression of RHEG drivers in Human Protein Atlas
(A) UMAP of all UVABC tumors colored by expression of mTOR component gene *MLST8* showing heterogeneous expression in all 5 tumors
(B) Immunohistochemistry for MLST8 protein in an ER+ breast tumor (Case #1874) from the HPA, depicting heterogeneous expression of MLST8 in breast cancer cells. Arrows highlight cells with higher expression (375).
(C) Semi-quantitative comparison of total protein expression of 44 RHEG drivers between normal breast tissue and breast tumors shows inconsistent trends

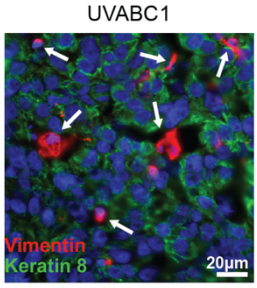

Figure 5.4 Heterogeneous expression of Vimentin protein confirmed in matched tissue from UVABC1.
Two-color immunofluorescence performed in matched patient tissue cryosections confirming heterogeneous expression of Vimentin (red, white arrows), an overdispersed candidate gene in 2 tumor samples, compared to Keratin 8 (green), which is a homogenously expressed luminal marker. Scale bar is 20µm.

5.2.1.2 3D organoid models for experimental tests of RHEG drivers

Tumor derived 3D organoids are an appropriate model to test RHEG drivers in a clinically relevant system that can be propagated and perturbed experimentally. A biobank of 3D organoids derived from ~95 luminal human breast tumors was recently described, along with matched bulk genomic and transcriptomic measurements (385). In luminal organoids which express RHEG drivers in bulk transcriptomic data, a next step would be to measure protein level of RHEG drivers using iterative-IF or imaging mass cytometry (**Figure 5.5A**) (288,289). This experiment would serve two purposes: first, validating RHEG driver heterogeneity in an independent cohort at the protein level, and second, providing a quantitative and spatial assessment of expression of RHEG driver proteins at the single-cell level. The next experiment would be to test if pre-treatment heterogeneous expression of RHEG drivers represents intrinsically resistant cell-states. Treatment of 3D tumor organoids with the anti-estrogen drugs tamoxifen and fulvestrant have been shown to demonstrate differential responses (385,386). Here we would further correlate the extent of response with RHEG driver expression pre- and post-treatment (**Figure 5.5B**). Proportions of resistant cells will be matched quantitatively to the frequency of RHEG driver expression prior to treatment, to narrow down the putative resistant markers. For a potential marker of intrinsic resistance like MLST8, I expect organoids with higher expression pre-treatment to display more resistance to tamoxifen. Since organoid models can be perturbed genetically, individual candidate gene expression will be induced exogenously to measure their ability to confer resistance in organoids that are otherwise sensitive to tamoxifen (**Figure 5.5C**).

Additionally, establishing protocols for testing drug responses in organoids can be extended to testing patient-specific models from UVA's Breast Care Clinic (385,387,388). This would allow for personalized treatment modeling, and testing dynamic responses to different dosing schedules. Ideally, organoids established from patients would be treated with different

combinations in different dosing frequencies to identify the most effective strategy. Organoid based treatment models would then be used as a personalized regimen for that patient.

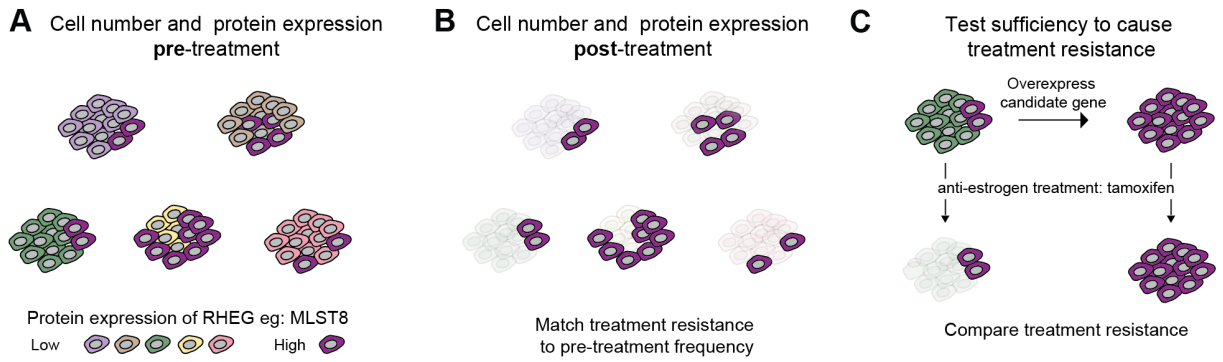

Figure 5.5 Experimental plan to test for intrinsic drug resistance in 3D organoids of luminal breast cancer
(A) 3D organoids of luminal breast cancer established from existing biobank or from patients at UVA. Viable cells and protein expression of RHEG drivers are measured across multiple organoids.
(B) Cells are treated with appropriate drug (eg. 1μm Tamoxifen or 0.5 μm Fulvestrant) and viable cells and protein expression of RHEG drivers are measured across multiple organoids. To identify RHEG drivers that confer intrinsic drug resistance, post-treatment persistor cells frequency will be matched to pre-treatment expression of RHEG drivers.
(C) The ability of specific RHEG drivers to confer drug resistance will be directly evaluated by comparing treatment responses in between wildtype organoids and organoids with exogenous RHEG driver overexpression.

5.2.2 Testing a potential role for oxidative stress in regulating variations between luminal cancer cells

A stress-responsive gene, *NQO1*, is a luminal breast cancer RHEG and a readout of the activity of pan-cancer driver gene *NFE2L2* (NRF2) (**Figure 3.13**). NQO1 protein is a detoxifying quinone reductase whose expression is induced by NFR2 in response to oxidative stress (389–391). The metabolism of estrogen to reactive quinones creates oxidative stress in luminal cancers, and could explain variable detection of NQO1 (375,392). Other RHEGs also included *CDKN1A* and *GDF15* that are transcriptionally regulated by the protein p53. Recent work in our lab has shown that oxidative stress responses are heterogeneous in breast epithelial cells and triple-negative breast tumors, and that these responses are mediated by NRF2 in coordination with p53 (328). Together, this suggests oxidative stress mediated activation of NRF2 and p53 may regulate heterogeneity between luminal cancer cells.

In previous work, a predictive model of oxidative stress response coordinated by NRF2 and p53 was developed (328). This model can be extended to 10cRNA-seq data from luminal cancers. I confirmed expression of all model inputs in luminal breast cancer 10cRNA-seq data from Chapter 3 (**Figure 5.6**). Transcript abundances indicate differential activation of these pathways within and across tumors, but model predictions will yield quantitative assessments of NRF2-p53 coordination and stress tolerance that cannot be not inferred by evaluating gene-expression changes alone. 10cRNA-seq data will be used to adjust the initial conditions to build individual stress dynamic models for each 10-cell measurement within tumors. I expect a subset of 10-cell samples to show high ROS tolerance and coordinated NRF2-p53 signaling. For these 10-cell samples, NRF2 and p53 would be candidates for transcriptional regulators of cell-states, which would then be tested *in vitro* using the model systems described in the previous section. 10-cell samples that do not show coordinated NRF2 and p53 mediated ROS tolerance would represent cell-states that might be vulnerable to increased oxidative stress.

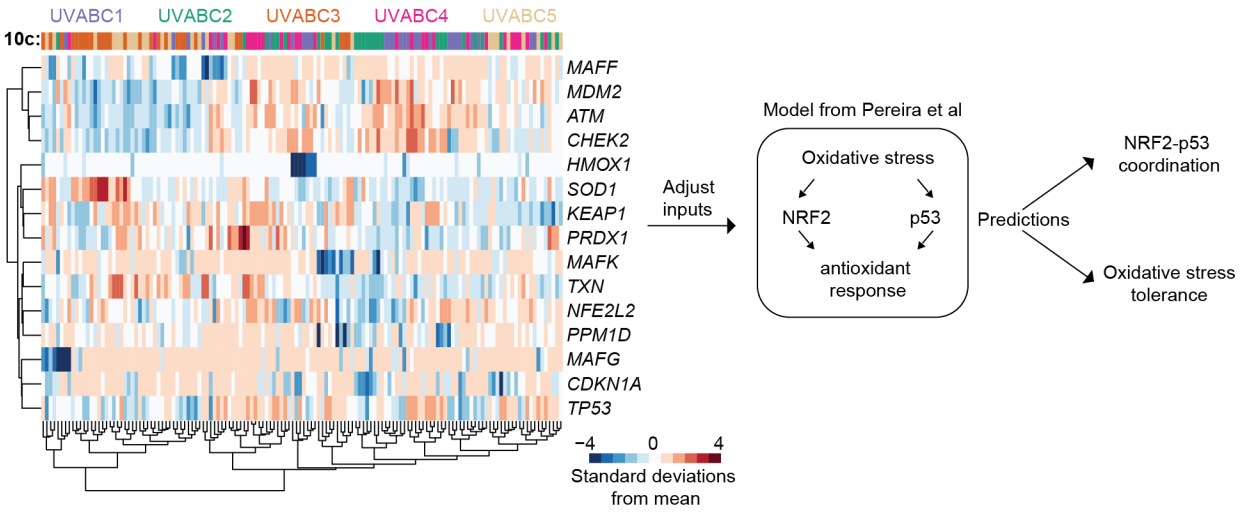

Figure 5.6 Oxidative stress as a cause of luminal breast cancer regulatory variations
Hierarchical clustering of 10-cell samples from UVABC tumors for transcripts that represent species in the NRF2-p53 coordinate model for oxidative stress tolerance shows heterogeneous expression within and across tumors. Expression values will be used to adjust model inputs and generate predictions for NRF2-p53 coordination and oxidative stress tolerance for each 10-cell sample.

5.2.3 Functional roles for RHEGs in SCLC cells measured in murine liver colonies

In Chapter 4, we observed that regulatory variations in KP1 SCLC cells expand dramatically when cancer cells are grown as liver colonies in both immunodeficient and immunocompetent mice. To focus on patterns of cancer cell variation that occur upon heterotypic interactions, we defined genes that were overdispersed in KP1 liver colonies of both mice as a set of *in vivo* RHEGs. The next subsections discuss approaches to further characterize the role of in vivo SCLC RHEGs.

5.2.3.1 Requirement for ATII markers: Cd74 and Lyz2

We observed heterogeneous expression of ATII cell type markers, *Cd74* and *Lyz2*, in KP1 cells after liver colonization, whereas their expression was were never detected in isolated 3D spheroids of KP1 cells (**Figure 4.10**). ATII differentiation occurs from pulmonary neuroendocrine cells (the cell of origin for SCLC) following injury and inflammation (174). Additionally, Cd74 is expressed by several epithelial cancer cells to promote survival signaling and cell growth (379,393). This suggests that *Cd74* and *Lyz2* expression and ATII differentiation may be necessary for liver colonization and metastatic growth for SCLC. Since Cd74 can be expressed by multiple cell-types, I confirmed that GFP+ SCLC cells in liver colonies also expressed Cd74 protein level in a pilot experiment (**Figure 5.7A**).

To test if ATII cell-type marker expression is necessary for liver colonization, I would utilize the CRISPR-Cas9 gene editing system to engineer KP1 cells that either lack functional *Cd74* or *Lyz2* genes. Since KP1 cells grown in culture have undetectable expression of *Cd74* and *Lyz2*, I would not expect the knockouts to have any phenotypic consequences *in vitro*. Control and knockout KP1 cells would then be assessed for liver colonization in C57/B6 x 129S F_1 mice with intact immune systems as described in Chapter 4. If PNEC differentiation into ATII serves to resolve injury and inflammation and promote growth, deleting these genes should

impair the metastatic process. Therefore, we would expect fewer liver colonies formed by KP1 cells that have *Cd74* and *Lyz2* knockouts as compared to parental cells. This would result in overall less metastatic disease in these mice, as well as longer survival (**Figure 5.7B**).

Cd74 is a cell surface receptor for its ligand Mif, and Mif binding induces Cd74 mediated signaling as well as transcriptional regulation (380). Interestingly, KP1 cells express Mif at abundant levels in vitro and in vivo (**Figure 4.10**), suggesting a ligand-receptor signaling interaction that may selectively play a role in the metastatic process of liver colonization. Although there were instances of 10-cell samples that show expression coordination for both Cd74 and Mif (**Figure 4.5**), 10-cell pooling limits our ability to determine if this interaction occurs through autocrine, paracrine, or juxtracrine signaling. To understand the mode of signaling for this ligand-receptor pair, I would perform multi-color immunofluorescence for Cd74 and Mif in KP1 liver colonies to assess if their expression is co-localized in single-cells or in adjacent cells, and whether it is restricted to KP1 cells or other cell-types in the liver. These experiments would confirm that RHEGs identified from 10cRNA-seq data reflect cell-state heterogeneities that are important for tumor growth and progression.

5.2.3.2 Regulation of KP1 cell de-differentiation

The mechanisms that induce the fragmented differentiation states in KP1 liver colonies remain unclear. In previous single-cell studies of SCLC GEMMs, non-neuroendocrine differentiation of PNECs has been associated with activation of the Notch pathway, but we did not see any evidence to suggest Notch pathway activation (Chapter 4, section 4.3.6). Further, the diverse, uncoordinated changes in marker expression by KP1 cells *in vivo* are suggestive of widespread changes in chromatin accessibility and transcriptional regulation. To begin exploring upstream regulators, I utilized bioinformatic tools to identify potential transcription factors (TFs) associated with the expression of all in vivo RHEGs. As different TF tools yield variable results due to differences in motif analysis methods, I used the intersection of three TF searching tools

to obtain a list of 21 candidate transcription factors (394–396). From these 21 bioinformatically derived candidate TFs, I would next narrow down to TFs that show gene-expression correlation with *in vivo* RHEGs. An additional confirmation would be to manually mine existing data from chromatin immunoprecipitation sequencing (ChIP-seq) experiments of the candidate TFs for evidence of binding events to coding sequences of RHEGs.

Further confirmation of regulation at the level of single genes and single transcription factors would require further experimental testing. One approach is performing ChIP-seq experiments with pull down of each candidate TF, and compare binding events between in vitro KP1 spheroids and in vivo KP1 colonies; this would provide confirmatory evidence at the level of a single candidate regulator. A whole-genome alternative is assay for transposase-accessible chromatin sequencing (ATAC-seq) measurements of in vitro KP1 spheroids and in vivo KP1 colonies to detect total changes in chromatin accessibility and occupancy (397). Compiling both levels of data will yield insights into differential chromatin occupancy as well as the most likely regulators of phenotypic plasticity in SCLC cells.

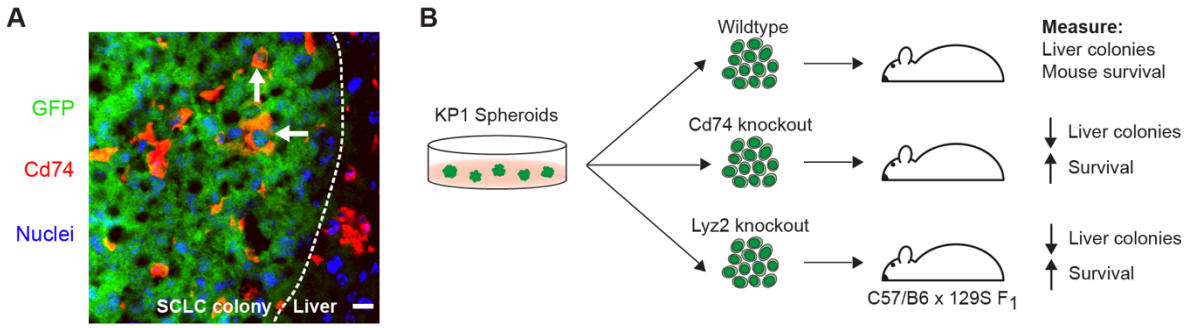

Figure 5.7 Testing the role of ATII-like SCLC cells in liver colonization
(A) Two-color immunofluorescence performed in C75/B6 F_1 liver cryosections confirming heterogeneous expression of ATII marker Cd74 (red, white arrows) by GFP-labeled KP1 cells (green). Colocalization of Cd74 and GFP is indicated by white arrows. Border of SCLC colony and liver tissue is marked by white dashed line. Scale bar is 20μm.
(B) Experimental plan to test functional role of ATII marker genes *Cd74* and *Lyz2* in liver colonization by KP1 cells. Both marker genes will be separately knocked out from KP1 cells using CRISPR-Cas9 mediated gene editing. Cells from all conditions will be delivered via tail vein to C75/B6 F_1 mice and assessed for liver colonization and overall survival to quantify metastatic capacity.

5.3 Future application of approaches in Chapters 2-4: risk stratification in breast premalignancies

Lobular carcinoma in situ (LCIS) is a non-invasive lesion of the breast with highly uncertain management and prognosis (398). CIS in the breast can arise as cells grow abnormally in the ducts (DCIS) or lobules (LCIS). DCIS lesions are considered obligate precursors to invasive cancer and therefore routinely excised. However, LCIS lesions are non-obligate precursors and their management is often institution dependent (399,400). LCIS lesions arise bilaterally and multifocally in patients, and have been associated with 8-11x increases in breast cancer risk of either breast despite a lack of association with germline mutations (399–402). Following a finding of LCIS, patients have to be monitored continually and consider drastic options like double mastectomies due to lack of prognostic biomarkers (400,403). The lack of clarity regarding the relation of LCIS to eventual tumors has created a clinical need for tools that stratify LCIS patients (399).

The molecular trajectory of normal cells to multifocal LCIS and cancers remains unclear. Previous work has identified recurrent losses in chromosomes 1q and 16q associated with loss of *CDH1* that persist across multiple LCIS foci and are retained in subsequent invasive tumors (400,404,405). These findings in multifocal disease suggest an early genomic event in normal mammary development that results in cells that are differentially poised to undergo malignant transformation. To understand tumor initiation, spatially resolved molecular characterizations of normal, premalignant, and malignant cells from patients diagnosed with LCIS are needed. 10cRNA-seq is uniquely suited to target cells that comprise different pathological features with single-cell precision, without losing cells to dissociation or sorting. Recently, we have shown the success of these approaches in characterizing trajectories of premalignant cells in a GEMM of gliomagenesis (271).

An ideal clinical sample for this study would include normal cells, LCIS foci, as well as invasive tumors, which co-occur in ~12% of diagnosed invasive breast cancers (406). For this

study, sample processing is likely to be the most challenging step. A large amount of breast tissue will have to be cryoembedded and sectioned because different pathological features will be spatially separated (**Figure 5.8A**). Since LCIS lesions and invasive tumors can be differentiated based on cell morphology, we can use nuclear stains to target epithelial cells for laser capture. We will obtain multiple 10-cell measurements from cells in normal lobules, cells in LCIS foci, as well as cancer cells in the same patient. Use analytical methods that can align cells in "psuedotime" based on transcriptional signatures, 10-cell transcriptomes will be used create a molecular trajectory for normal cells to malignancy (**Figure 5.8B**) (407–410). Further, heterogeneously expressed genes will be identified using stochastic profiling for all conditions. Combining heterogeneous expression with trajectory inference will enable a tracking of cell-states that are heterogeneous prior to malignancy (in normal and LCIS cells) but become selected for homogenous expression as cells become invasive (**Figure 5.8C**). Future extensions to stochastic profiling that enabe matched genomic measurements from cells will also allow us to relate transcriptional trajectories to genomic evolution in these lesions (411).

While this detailed characterization will be patient-specific, it is an important step in understanding this complex disease. LCIS foci also display inter-patient variations, and future measurements will have to include morphological variants of LCIS (400,412). Candidate genes identified using the outlined prospective approach can then be measured in banked retrospective samples to assess generalizability. Further, LCIS has been identified in pathological assessments of breast tumors profiled by the TCGA. Since TCGA expression data has matched clinical outcomes, it can be used to relate the expression of candidate genes to overall survival to identify prognostic biomarkers. Together, these approaches will suggest biomarkers to predict higher risk of invasive disease and enable better patient stratification at diagnosis.

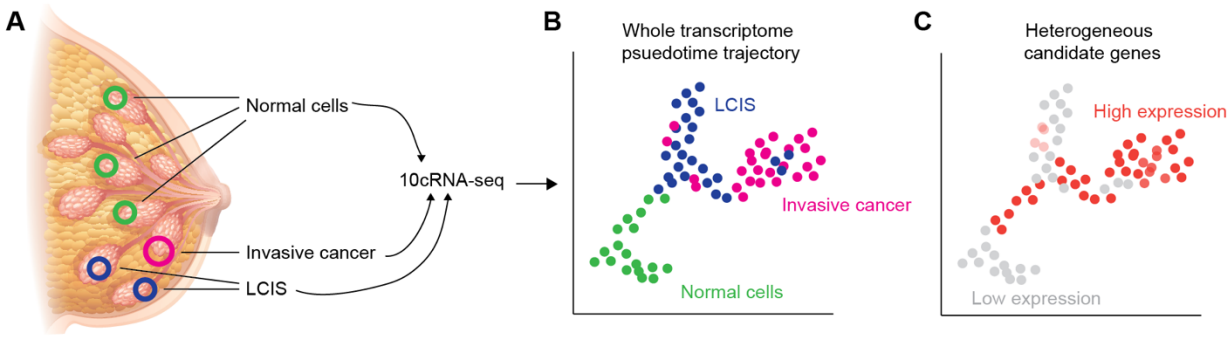

Figure 5.8 Using 10cRNA-seq to profile the malignant trajectory of cells in cases of co-occurring LCIS and invasive breast cancer
(A) Schematic of a human breast depicting that normal cells, foci of LCIS, and invasive cancers can be co-occurring but spatially separated across the breast
(B) Schematic of psuedotime analyses generated from 10-cell transcriptomes of normal, LCIS, and cancer cells from the same patient.
(C) Schematic of expression of a normal and LCIS heterogeneous gene projected on the psuedotime analysis from (B). Such genes that get selected for expression over the malignant trajectory would be potential biomarkers for patient stratification.

5.4 Concluding remarks

Cancer cell behavior is the result of a complex integration of internal and external cues, creating phenotypic diversity beyond what is observed in normal cells. The breadth of regulatory variations in human tumors is only beginning to be uncovered and is the first step in relating molecular and cellular variations to overall patient outcomes. In this dissertation, we developed complimentary experimental and bioinformatic approaches to identify markers and modulators of heterogenous cancer cell behaviors within isolated cells, solid tumors, and metastases. By identifying expression signatures that are unique to some sample types and shared across multiple others, we begin to decode the complex network of interactions that influence cellular states. The findings presented in this dissertation provide novel insight into the transcriptional landscapes of breast cancer and lung cancer cells, towards the goal of understanding differential outcomes for patients with these diseases. The approaches for in situ profiling presented here are readily applicable to other tumor models, particularly those where the microenvironmental determinants of cellular states remain unknown.

5.5 Materials and methods

Immunofluorescence on cryosections was performed as previously described with minor modifications (201). Briefly, cryosectioned slides stored at –80˚C were immediately fixed in 3.7% para-formaldehyde solution in PBS for 15m. Slides were then rinsed three times with PBS for 5m, prior to 1 hour blocking with 1X Western Blocking Reagent (Roche #11921673001) diluted in PBS + 0.3% Tween-20. The following primary antibodies were used: KRT8 (Millipore MAB3414, 1:200), Vimentin (Abcam ab16700, 1:200), GFP (Abcam ab13970, 1:1000), Cd74 (BD Bioscience 555317, 1:200). Slides were incubated with primary antibody at room temperature overnight and with secondary antibody (1:200) for 1hr. Slides were counterstained with 0.5µg/ml DAPI to visualize nuclei and imaged the same day.

6 References

1. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976 Oct 1;194(4260):23–8.

2. Weinberg RA. The genetic origins of human cancer. Cancer. 1988;61(10):1963–8.

3. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011 Mar 4;144(5):646–74.

4. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013 Sep 19;501(7467):338–45.

5. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012 Oct;490(7418):61–70.

6. George J, Lim JS, Jang SJ, Cun Y, Ozretić L, Kong G, et al. Comprehensive genomic profiles of small cell lung cancer. Nature. 2015 Aug;524(7563):47–53.

7. Pan-cancer analysis of whole genomes. Nature. 2020 Feb;578(7793):82–93.

8. Garraway LA, Sellers WR. Lineage dependency and lineage-survival oncogenes in human cancer. Nature Reviews Cancer. 2006 Aug;6(8):593–602.

9. Allinen M, Beroukhim R, Cai L, Brennan C, Lahti-Domenici J, Huang H, et al. Molecular characterization of the tumor microenvironment in breast cancer. Cancer Cell. 2004 Jul;6(1):17–32.

10. Visvader JE. Cells of origin in cancer. Nature. 2011 Jan;469(7330):314–22.

11. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS. 2001 Sep 11;98(19):10869–74.

12. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J Clin Oncol. 2009 Mar 10;27(8):1160–7.

13. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, et al. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin Cancer Res. 2004 Aug 15;10(16):5367–74.

14. Hammond MEH, Hayes DF, Dowsett M, Allred DC, Hagerty KL, Badve S, et al. American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Immunohistochemical Testing of Estrogen and Progesterone Receptors in Breast Cancer. Arch Pathol Lab Med. 2010 Jul;134(7):e48-72

15. Oh D-Y, Bang Y-J. HER2-targeted therapies - a role beyond breast cancer. Nat Rev Clin Oncol. 2020 Jan;17(1):33–48.

16. Sestak I, Dowsett M, Zabaglo L, Lopez-Knowles E, Ferree S, Cowens JW, et al. Factors Predicting Late Recurrence for Estrogen Receptor–Positive Breast Cancer. J Natl Cancer Inst. 2013 Oct 2;105(19):1504–11.

17. Wangchinda P, Ithimakin S. Factors that predict recurrence later than 5 years after initial treatment in operable breast cancer. World Journal of Surgical Oncology. 2016 Aug 24;14(1):223.

18. Levsky JM, Singer RH. Gene expression and the myth of the average cell. Trends Cell Biol. 2003 Jan;13(1):4–6.

19. Kim C, Gao R, Sei E, Brandt R, Hartman J, Hatschek T, et al. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. Cell. 2018 May 3;173(4):879-893.e13.

20. Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature. 2017 15;546(7658):431–5.

21. Fidler IJ. Tumor heterogeneity and the biology of cancer invasion and metastasis. Cancer Res. 1978 Sep;38(9):2651–60.

22. Mao Y, Keller ET, Garfield DH, Shen K, Wang J. Stroma Cells in Tumor Microenvironment and Breast Cancer. Cancer Metastasis Rev. 2013 Jun;32(0):303–15.

23. Chanmee T, Ontong P, Konno K, Itano N. Tumor-Associated Macrophages as Major Players in the Tumor Microenvironment. Cancers. 2014 Sep;6(3):1670–90.

24. Woo EY, Chu CS, Goletz TJ, Schlienger K, Yeh H, Coukos G, et al. Regulatory CD4+CD25+ T Cells in Tumors from Patients with Early-Stage Non-Small Cell Lung Cancer and Late-Stage Ovarian Cancer. Cancer Res. 2001 Jun 15;61(12):4766–72.

25. Hanahan D, Coussens LM. Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. Cancer Cell. 2012 Mar 20;21(3):309–22.

26. Chaudhuri O, Koshy ST, Branco da Cunha C, Shin J-W, Verbeke CS, Allison KH, et al. Extracellular matrix stiffness and composition jointly regulate the induction of malignant phenotypes in mammary epithelium. Nat Mater. 2014 Oct;13(10):970–8.

27. Reik W, Dean W, Walter J. Epigenetic Reprogramming in Mammalian Development. Science. 2001 Aug 10;293(5532):1089–93.

28. Rokicki W, Rokicki M, Wojtacha J, Dżeljijli A. The role and importance of club cells (Clara cells) in the pathogenesis of some respiratory diseases. Kardiochir Torakochirurgia Pol. 2016 Mar;13(1):26–30.

29. Liu Z, Liao F, Scozzi D, Furuya Y, Pugh KN, Hachem R, et al. An obligatory role for club cells in preventing obliterative bronchiolitis in lung transplants. JCI Insight. 2019 Apr; 5.124732

30. Nikolić MZ, Sun D, Rawlins EL. Human lung development: recent progress and new challenges. Development. 2018 Aug;145(16).

31. Franks TJ, Colby TV, Travis WD, Tuder RM, Reynolds HY, Brody AR, et al. Resident Cellular Components of the Human Lung. Proc Am Thorac Soc. 2008 Sep 15;5(7):763–6.

32. Crystal RG, Randell SH, Engelhardt JF, Voynow J, Sunday ME. Airway Epithelial Cells. Proc Am Thorac Soc. 2008 Sep 15;5(7):772–7.

33. Rock JR, Randell SH, Hogan BLM. Airway basal stem cells: a perspective on their roles in epithelial homeostasis and remodeling. Disease Models & Mechanisms. 2010 Sep 1;3(9–10):545–56.

34. Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med. 2018;24(8):1277–89.

35. Johnson SK, Kerr KM, Chapman AD, Kennedy MM, King G, Cockburn JS, et al. Immune cell infiltrates and prognosis in primary carcinoma of the lung. Lung Cancer. 2000 Jan 1;27(1):27–35.

36. Lavin Y, Kobayashi S, Leader A, Amir E-AD, Elefant N, Bigenwald C, et al. Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. Cell. 2017 04;169(4):750-765.e17.

37. Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat Med. 2015 Aug;21(8):938–45.

38. Djenidi F, Adam J, Goubar A, Durgeau A, Meurice G, Montpréville V de, et al. CD8+CD103+ Tumor–Infiltrating Lymphocytes Are Tumor-Specific Tissue-Resident Memory T Cells and a Prognostic Factor for Survival in Lung Cancer Patients. The Journal of Immunology. 2015 Apr 1;194(7):3475–86.

39. Jain P, Jain C, Velcheti V. Role of immune-checkpoint inhibitors in lung cancer. Ther Adv Respir Dis. 2018 Dec;12:1753465817750075.

40. Kakimi K, Matsushita H, Murakawa T, Nakajima J. γδ T cell therapy for the treatment of non-small cell lung cancer. Transl Lung Cancer Res. 2014 Feb;3(1):23–33.

41. Perica K, Varela JC, Oelke M, Schneck J. Adoptive T cell immunotherapy for cancer. Rambam Maimonides medical journal. 2015 Jan;6(1).

42. Anagnostou VK, Brahmer JR. Cancer Immunotherapy: A Future Paradigm Shift in the Treatment of Non–Small Cell Lung Cancer. Clin Cancer Res. 2015 Mar 1;21(5):976–84.

43. Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, et al. Safety, Activity, and Immune Correlates of Anti–PD-1 Antibody in Cancer. N Engl J Med. 2012 Jun 28;366(26):2443–54.

44. He Y, Rozeboom L, Rivard CJ, Ellison K, Dziadziuszko R, Yu H, et al. PD-1, PD-L1 Protein Expression in Non-Small Cell Lung Cancer and Their Relationship with Tumor-Infiltrating Lymphocytes. Med Sci Monit. 2017 Mar 9;23:1208–16.

45. Gandhi L, Rodríguez-Abreu D, Gadgeel S, Esteban E, Felip E, De Angelis F, et al. Pembrolizumab plus Chemotherapy in Metastatic Non-Small-Cell Lung Cancer. N Engl J Med. 2018 May 31;378(22):2078–92.

46. Diaz Bessone MI, Gattas MJ, Laporte T, Tanaka M, Simian M. The Tumor Microenvironment as a Regulator of Endocrine Resistance in Breast Cancer. Front Endocrinol. 2019 Aug;10:547

47. Kalluri R. The biology and function of fibroblasts in cancer. Nat Rev Cancer. 2016 Aug 23;16(9):582–98.

48. Bhowmick NA, Neilson EG, Moses HL. Stromal fibroblasts in cancer initiation and progression. Nature. 2004 Nov 18;432(7015):332–7.

49. Tyan S-W, Kuo W-H, Huang C-K, Pan C-C, Shew J-Y, Chang K-J, et al. Breast cancer cells induce cancer-associated fibroblasts to secrete hepatocyte growth factor to enhance breast tumorigenesis. PLoS ONE. 2011 Jan 13;6(1):e15313.

50. Orimo A, Gupta PB, Sgroi DC, Arenzana-Seisdedos F, Delaunay T, Naeem R, et al. Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated SDF-1/CXCL12 secretion. Cell. 2005 May 6;121(3):335–48.

51. Martinez-Outschoorn UE, Goldberg AF, Lin Z, Ko Y-H, Flomenberg N, Wang C, et al. Anti-estrogen resistance in breast cancer is induced by the tumor microenvironment and can be overcome by inhibiting mitochondrial function in epithelial cancer cells. Cancer Biology & Therapy. 2011 Nov 15;12(10):924–38.

52. Brechbuhl HM, Finlay-Schultz J, Yamamoto TM, Gillen AE, Cittelly DM, Tan A-C, et al. Fibroblast Subtypes Regulate Responsiveness of Luminal Breast Cancer to Estrogen. Clin Cancer Res. 2017 Apr 1;23(7):1710–21.

53. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009 Apr 9;458(7239):719–24.

54. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012 Mar 8;366(10):883–92.

55. Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C, et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nature Medicine. 2016 Jan;22(1):105–13.

56. Bhang HC, Ruddy DA, Krishnamurthy Radhakrishna V, Caushi JX, Zhao R, Hims MM, et al. Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. Nat Med. 2015 May;21(5):440–8.

57. Casasent AK, Schalck A, Gao R, Sei E, Long A, Pangburn W, et al. Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. Cell. 2018 Jan 11;172(1):205-217.e12.

58. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012 Jan;481(7381):306–13.

59. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011 Apr;472(7341):90–4.

60. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014 Aug;512(7513):155–60.

61. Zhang J, Fujimoto J, Zhang J, Wedge DC, Song X, Zhang J, Seth S, Chow CW, Cao Y, Gumbs C, Gold KA. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science. 2014 Oct 10;346(6206):256-9.

62. Navin NE. The first five years of single-cell cancer genomics and beyond. Genome Res. 2015 Oct;25(10):1499–507.

63. Lim B, Lin Y, Navin N. Advancing Cancer Research and Medicine with Single-Cell Genomics. Cancer Cell. 2020 Apr 13;37(4):456–70.

64. Ciriello G, Sinha R, Hoadley KA, Jacobsen AS, Reva B, Perou CM, et al. The molecular diversity of Luminal A breast tumors. Breast Cancer Res Treat. 2013 Oct;141(3):409–20.

65. Lawson DA, Kessenbrock K, Davis RT, Pervolarakis N, Werb Z. Tumour heterogeneity and metastasis at single-cell resolution. Nature Cell Biology. 2018 Dec;20(12):1349.

66. Haber DA, Settleman J. Drivers and passengers. Nature. 2007 Mar;446(7132):145–6.

67. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational Prediction of Driver Missense Mutations. Cancer Res. 2009 Aug 15;69(16):6660–7.

68. Bozic I, Antal T, Ohtsuki H, Carter H, Kim D, Chen S, et al. Accumulation of driver and passenger mutations during tumor progression. Proceedings of the National Academy of Sciences. 2010 Oct 26;107(43):18545–50.

69. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 2013 Mar 29;339(6127):1546–58.

70. Janes KA. Single-cell states versus single-cell atlases — two classes of heterogeneity that differ in meaning and method. Current Opinion in Biotechnology. 2016 Jun 1;39:120–5.

71. Bian S, Hou Y, Zhou X, Li X, Yong J, Wang Y, Wang W, Yan J, Hu B, Guo H, Wang J. Single-cell multiomics sequencing and analyses of human colorectal cancer. Science. 2018 Nov 30;362(6418):1060-3.

72. Ali HR, Jackson HW, Zanotelli VRT, Danenberg E, Fischer JR, Bardwell H, et al. Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. Nature Cancer. 2020 Feb;1(2):163–75.

73. Jackson HW, Fischer JR, Zanotelli VRT, Ali HR, Mechera R, Soysal SD, et al. The single-cell pathology landscape of breast cancer. Nature. 2020 Jan 20;1–6.

74. Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, Richman AR, Silverbush D, Shaw ML, Hebert CM, Dewitt J. An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell. 2019 Aug 8;178(4):835-49.

75. Gupta PB, Fillmore CM, Jiang G, Shapira SD, Tao K, Kuperwasser C, et al. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. Cell. 2011 Aug 19;146(4):633–44.

76. Wang C-C, Bajikar SS, Jamal L, Atkins KA, Janes KA. A time- and matrix-dependent TGFBR3-JUND-KRT5 regulatory circuit in single breast epithelial cells and basal-like premalignancies. Nat Cell Biol. 2014 Apr;16(4):345–56.

77. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, et al. Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. MBoC. 2002 Mar 21;13(6):1977–2000.

78. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016 Apr 8;352(6282):189–96.

79. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014 Jun 20;344(6190):1396–401.

80. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nature Reviews Cancer. 2012 May;12(5):323–34.

81. Tlsty TD, Coussens LM. Tumor stroma and regulation of cancer development. Annu Rev Pathol. 2006;1:119–50.

82. Robertson-Tessi M, Gillies RJ, Gatenby RA, Anderson ARA. Impact of Metabolic Heterogeneity on Tumor Growth, Invasion, and Treatment Outcomes. Cancer Res. 2015 Apr 15;75(8):1567–79.

83. Kim J, Tchernyshyov I, Semenza GL, Dang CV. HIF-1-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia. Cell Metab. 2006 Mar;3(3):177–85.

84. Helmlinger G, Yuan F, Dellian M, Jain RK. Interstitial pH and pO2 gradients in solid tumors in vivo: high-resolution measurements reveal a lack of correlation. Nat Med. 1997 Feb;3(2):177–82.

85. Papandreou I, Cairns RA, Fontana L, Lim AL, Denko NC. HIF-1 mediates adaptation to hypoxia by actively downregulating mitochondrial oxygen consumption. Cell Metab. 2006 Mar;3(3):187–97.

86. Lorusso G, Rüegg C. The tumor microenvironment and its contribution to tumor evolution toward metastasis. Histochem Cell Biol. 2008 Dec 1;130(6):1091–103.

87. Quail DF, Joyce JA. Microenvironmental regulation of tumor progression and metastasis. Nature Medicine. 2013 Nov;19(11):1423–37.

88. Cheung KJ, Ewald AJ. Illuminating breast cancer invasion: diverse roles for cell-cell interactions. Curr Opin Cell Biol. 2014 Oct;30:99–111.

89. Patsialou A, Wyckoff J, Wang Y, Goswami S, Stanley ER, Condeelis JS. Invasion of human breast cancer cells in vivo requires both paracrine and autocrine loops involving the colony-stimulating factor-1 receptor. Cancer Res. 2009 Dec 15;69(24):9498–506.

90. Wyckoff J, Wang W, Lin EY, Wang Y, Pixley F, Stanley ER, et al. A paracrine loop between tumor cells and macrophages is required for tumor cell migration in mammary tumors. Cancer Res. 2004 Oct 1;64(19):7022–9.

91. Müller A, Homey B, Soto H, Ge N, Catron D, Buchanan ME, et al. Involvement of chemokine receptors in breast cancer metastasis. Nature. 2001 Mar;410(6824):50–6.

92. Brooks PC, Strömblad S, Sanders LC, von Schalscha TL, Aimes RT, Stetler-Stevenson WG, et al. Localization of matrix metalloproteinase MMP-2 to the surface of invasive cells by interaction with integrin alpha v beta 3. Cell. 1996 May 31;85(5):683–93.

93. Wang W, Goswami S, Lapidus K, Wells AL, Wyckoff JB, Sahai E, et al. Identification and testing of a gene expression signature of invasive carcinoma cells within primary mammary tumors. Cancer Res. 2004 Dec 1;64(23):8585–94.

94. Wyckoff JB, Pinner SE, Gschmeissner S, Condeelis JS, Sahai E. ROCK- and Myosin-Dependent Matrix Deformation Enables Protease-Independent Tumor-Cell Invasion In Vivo. Current Biology. 2006 Aug 8;16(15):1515–23.

95. Nguyen-Ngoc K-V, Cheung KJ, Brenot A, Shamir ER, Gray RS, Hines WC, et al. ECM microenvironment regulates collective migration and local dissemination in normal and malignant mammary epithelium. Proc Natl Acad Sci U S A. 2012 Sep 25;109(39):E2595–604.

96. Le Magnen C, Shen MM, Abate-Shen C. Lineage Plasticity in Cancer Progression and Treatment. Annu Rev Cancer Biol. 2018 Mar;2:271–89.

97. Thiery JP. Epithelial–mesenchymal transitions in tumour progression. Nature Reviews Cancer. 2002 Jun;2(6):442–54.

98. Nieto MA, Huang RY-J, Jackson RA, Thiery JP. EMT: 2016. Cell. 2016 Jun 30;166(1):21–45.

99. Wei SC, Fattet L, Tsai JH, Guo Y, Pai VH, Majeski HE, et al. Matrix stiffness drives epithelial–mesenchymal transition and tumour metastasis through a TWIST1–G3BP2 mechanotransduction pathway. Nature Cell Biology. 2015 May;17(5):678–88.

100. Mu P, Zhang Z, Benelli M, Karthaus WR, Hoover E, Chen C-C, et al. SOX2 promotes lineage plasticity and antiandrogen resistance in TP53- and RB1-deficient prostate cancer. Science. 2017 Jan 6;355(6320):84–8.

101. Davies AH, Beltran H, Zoubeidi A. Cellular plasticity and the neuroendocrine phenotype in prostate cancer. Nature Reviews Urology. 2018 May;15(5):271–86.

102. Hangauer MJ, Viswanathan VS, Ryan MJ, Bole D, Eaton JK, Matov A, et al. Drug-tolerant persister cancer cells are vulnerable to GPX4 inhibition. Nature. 2017 Nov;551(7679):247–50.

103. Viswanathan VS, Ryan MJ, Dhruv HD, Gill S, Eichhoff OM, Seashore-Ludlow B, et al. Dependency of a therapy-resistant state of cancer cells on a lipid peroxidase pathway. Nature. 2017 27;547(7664):453–7.

104. Sharma SV, Lee DY, Li B, Quinlan MP, Takahashi F, Maheswaran S, et al. A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. Cell. 2010 Apr 2;141(1):69–80.

105. Patten DK, Corleone G, Győrffy B, Perone Y, Slaven N, Barozzi I, et al. Enhancer mapping uncovers phenotypic heterogeneity and evolution in patients with luminal breast cancer. Nature Medicine. 2018 Jul 23;1.

106. Hein SM, Haricharan S, Johnston AN, Toneff MJ, Reddy JP, Dong J, et al. Luminal epithelial cells within the mammary gland can produce basal cells upon oncogenic stress. Oncogene. 2016 Mar 17;35(11):1461–7.

107. Lloyd MC, Cunningham JJ, Bui MM, Gillies RJ, Brown JS, Gatenby RA. Darwinian Dynamics of Intratumoral Heterogeneity: Not Solely Random Mutations but Also Variable Environmental Selection Forces. Cancer Research. 2016 Jun 1;76(11):3136–44.

108. Junttila MR, de Sauvage FJ. Influence of tumour micro-environment heterogeneity on therapeutic response. Nature. 2013 Sep;501(7467):346–54.

109. Stewart CA, Gay CM, Xi Y, Sivajothi S, Sivakamasundari V, Fujimoto J, et al. Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. Nature Cancer. 2020 Apr;1(4):423–36.

110. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The Technology and Biology of Single-Cell RNA Sequencing. Molecular Cell. 2015 May 21;58(4):610–20.

111. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols. 2018 Apr;13(4):599–604.

112. Tirosh I, Suvà ML. Deciphering Human Tumor Biology by Single-Cell Expression Profiling. Annual Review of Cancer Biology. 2019;3(1):151–66.

113. Svensson V, Beltrame E da V. A curated database reveals trends in single cell transcriptomics. bioRxiv. 2019 Aug 21;742304.

114. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotech. 2016 Nov;34(11):1145–60.

115. Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, Specht MC, Bernstein BE, Michor F, Ellisen LW. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nature communications. 2018 Sep 4;9(1):1-0.

116. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell. 2018 Aug 23;174(5):1293-1308.e36.

117. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, Ryu HS, Kim S, Lee JE, Park YH, Kan Z. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nature communications. 2017 May 5;8(1):1-2.

118. Yin J, Li Z, Yan C, Fang E, Wang T, Zhou H, et al. Comprehensive analysis of immune evasion in breast cancer by single-cell RNA-seq. bioRxiv. 2018 Jul 16;368605.

119. Qiu S, Hong R, Zhuang Z, Li Y, Zhu L, Lin X, et al. A Single-Cell Immune Atlas of Triple Negative Breast Cancer Reveals Novel Immune Cell Subsets. bioRxiv. 2019 Jul 5;566968.

120. Savas P, Virassamy B, Ye C, Salim A, Mintoff CP, Caramia F, et al. Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. Nature Medicine. 2018 Jul;24(7):986–93.

121. Guo X, Zhang Y, Zheng L, Zheng C, Song J, Zhang Q, et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. Nature Medicine. 2018 Jul;24(7):978–85.

122. Peng J, Sun B-F, Chen C-Y, Zhou J-Y, Chen Y-S, Chen H, et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res. 2019 Sep;29(9):725–38.

123. Zheng C, Zheng L, Yoo J-K, Guo H, Zhang Y, Guo X, et al. Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. Cell. 2017 Jun 15;169(7):1342-1356.e16.

124. Suvà ML, Tirosh I. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. Mol Cell. 2019 11;75(1):7–12.

125. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell. 2017 Dec 14;171(7):1611-1624.e24.

126. Konieczkowski DJ, Johannessen CM, Abudayyeh O, Kim JW, Cooper ZA, Piris A, et al. A melanoma cell state distinction influences sensitivity to MAPK pathway inhibitors. Cancer Discov. 2014 Jul;4(7):816–27.

127. Sade-Feldman M, Yizhak K, Bjorgaard SL, Ray JP, de Boer CG, Jenkins RW, et al. Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. Cell. 2018 Nov 1;175(4):998-1013.e20.

128. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su M-J, Melms JC, et al. A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. Cell. 2018 Nov 1;175(4):984-997.e24.

129. Janes KA, Wang C-C, Holmberg KJ, Cabral K, Brugge JS. Identifying single-cell molecular programs by stochastic profiling. Nat Methods. 2010 Apr;7(4):311–7.

130. Letai A. Functional precision cancer medicine-moving beyond pure genomics. Nat Med. 2017 Sep 8;23(9):1028–35.

131. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009 May;6(5):377–82.

132. See P, Lum J, Chen J, Ginhoux F. A single-cell sequencing guide for immunologists. Frontiers in immunology. 2018 Oct 23;9:2425.

133. Nguyen QH, Pervolarakis N, Nee K, Kessenbrock K. Experimental considerations for single-cell RNA sequencing approaches. Frontiers in cell and developmental biology. 2018 Sep 4;6:108.

134. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Molecular Cell. 2017 Feb;65(4):631-643.e4.

135. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Medicine. 2017 Aug 18;9:75.

136. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nature Protocols. 2014 Jan;9(1):171–81.

137. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature Biotechnology. 2012 Aug;30(8):777–82.

138. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types. Science. 2014 Feb 14;343(6172):776–9.

139. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015 May 21;161(5):1202–14.

140. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell. 2015 May 21;161(5):1187–201.

141. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017 Jan 16;8(1):1–12.

142. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods. 2014 Feb;11(2):163–6.

143. van den Brink SC, Sage F, Vértesy Á, Spanjaard B, Peterson-Maduro J, Baron CS, et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat Methods. 2017 Sep 29;14(10):935–6.

144. Nichterwitz S, Chen G, Aguila Benitez J, Yilmaz M, Storvall H, Cao M, et al. Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. Nat Commun. 2016 08;7:12139.

145. Chen J, Suo S, Tam PP, Han J-DJ, Peng G, Jing N. Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. Nat Protoc. 2017;12(3):566–80.

146. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013 Nov;10(11):1093–5.

147. Bhargava V, Head SR, Ordoukhanian P, Mercola M, Subramaniam S. Technical variations in low-input RNA-seq methodologies. Sci Rep. 2014 Jan 14;4:3678.

148. Eberwine J, Sul J-Y, Bartfai T, Kim J. The promise of single-cell sequencing. Nat Methods. 2014 Jan;11(1):25–7.

149. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014 Jun;11(6):637–40.

150. Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, Macaulay IC, et al. Power analysis of single-cell RNA-sequencing experiments. Nat Meth. 2017 Apr;14(4):381–7.

151. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nature Reviews Genetics. 2015 Mar;16(3):133–45.

152. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology. 2019 Jan;37(1):38–44.

153. Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nature Methods. 2016 Mar;13(3):241–4.

154. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nature Methods. 2014 Jul;11(7):740–2.

155. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biology. 2015 Nov 2;16(1):241.

156. Maaten L van der, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008;9(Nov):2579–605.

157. Wang L, Janes KA. Stochastic profiling of transcriptional regulatory heterogeneities in tissues, tumors and cultured cells. Nat Protoc. 2013 Feb;8(2):282–301.

158. Bajikar SS, Fuchs C, Roller A, Theis FJ, Janes KA. Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles. Proc Natl Acad Sci USA. 2014 Feb 4;111(5):E626-635.

159. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. World J Clin Oncol. 2014 Aug 10;5(3):412–24.

160. Ring A, Dowsett M. Mechanisms of tamoxifen resistance. Endocrine-Related Cancer. 2004 Dec 1;11(4):643–58.

161. Ciruelos Gil EM. Targeting the PI3K/AKT/mTOR pathway in estrogen receptor-positive breast cancer. Cancer Treatment Reviews. 2014 Aug 1;40(7):862–71.

162. Makhoul I, Atiq M, Alwbari A, Kieber-Emmons T. Breast cancer immunotherapy: An update. Breast cancer: basic and clinical research. 2018 May 30;12:1178223418774802.

163. Zhu B, Tse LA, Wang D, Koka H, Zhang T, Abubakar M, et al. Immune gene expression profiling reveals heterogeneity in luminal breast tumors. Breast Cancer Research. 2019 Dec 19;21(1):147.

164. Verret B, Cortes J, Bachelot T, Andre F, Arnedos M. Efficacy of PI3K inhibitors in advanced breast cancer. Ann Oncol. 2019 Dec;30(Suppl 10):x12–20.

165. Sung H, Garcia-Closas M, Chang-Claude J, Blows FM, Ali HR, Figueroa J, et al. Heterogeneity of luminal breast cancer characterised by immunohistochemical expression of basal markers. Br J Cancer. 2016 Feb 2;114(3):298–304.

166. Finn RS, Martin M, Rugo HS, Jones S, Im S-A, Gelmon K, et al. Palbociclib and Letrozole in Advanced Breast Cancer. New England Journal of Medicine. 2016 Nov 17;375(20):1925–36.

167. Park K-S, Liang M-C, Raiser DM, Zamponi R, Roach RR, Curtis SJ, et al. Characterization of the cell of origin for small cell lung cancer. Cell Cycle. 2011 Aug 15;10(16):2806–15.

168. Sutherland KD, Proost N, Brouns I, Adriaensen D, Song J-Y, Berns A. Cell of Origin of Small Cell Lung Cancer: Inactivation of Trp53 and Rb1 in Distinct Cell Types of Adult Mouse Lung. Cancer Cell. 2011 Jun 14;19(6):754–64.

169. Song H, Yao E, Lin C, Gacayan R, Chen MH, Chuang PT. Functional characterization of pulmonary neuroendocrine cells in lung development, injury, and tumorigenesis. Proc Natl Acad Sci U S A. 2012 Oct 23;109(43):17531–6.

170. Ouadah Y, Rojas ER, Riordan DP, Capostagno S, Kuo CS, Krasnow MA. Rare Pulmonary Neuroendocrine Cells Are Stem Cells Regulated by Rb, p53, and Notch. Cell. 2019 Oct 3;179(2):403-416.e23.

171. Byers LA, Rudin CM. Small cell lung cancer: Where do we go from here? Cancer. 2015 Mar 1;121(5):664–72.

172. Yoshida T, Kakegawa J, Yamaguchi T, Hantani Y, Okajima N, Sakai T, et al. Identification and characterization of a novel chemotype MEK inhibitor able to alter the phosphorylation state of MEK1/2. Oncotarget. 2012 Dec;3(12):1533–45.

173. Lim JS, Ibaseta A, Fischer MM, Cancilla B, O'Young G, Cristea S, et al. Intratumoral heterogeneity generated by Notch signaling promotes small cell lung cancer. Nature. 2017 May 18;545(7654):360–4.

174. Marsh LM, Cakarova L, Kwapiszewska G, von Wulffen W, Herold S, Seeger W, et al. Surface expression of CD74 by type II alveolar epithelial cells: a potential mechanism for macrophage migration inhibitory factor-induced epithelial repair. Am J Physiol Lung Cell Mol Physiol. 2009 Mar;296(3):L442-52.

175. Rudin CM, Poirier JT, Byers LA, Dive C, Dowlati A, George J, et al. Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. Nat Rev Cancer. 2019 May;19(5):289–97.

176. Oser MG, Fonseca R, Chakraborty AA, Brough R, Spektor A, Jennings RB, et al. Cells Lacking the RB1 Tumor Suppressor Gene Are Hyperdependent on Aurora B Kinase for Survival. Cancer Discov. 2019;9(2):230–47.

177. Wooten DJ, Groves SM, Tyson DR, Liu Q, Lim JS, Albert R, et al. Systems-level network modeling of Small Cell Lung Cancer subtypes identifies master regulators and destabilizers. PLoS Comput Biol. 2019 Oct;15(10):e1007343.

178. Udyavar AR, Wooten DJ, Hoeksema M, Bansal M, Califano A, Estrada L, et al. Novel Hybrid Phenotype Revealed in Small Cell Lung Cancer by a Transcription Factor Network Model That Can Explain Tumor Heterogeneity. Cancer Res. 2017 01;77(5):1063–74.

179. Park K-S, Martelotto LG, Peifer M, Sos ML, Karnezis AN, Mahjoub MR, et al. A crucial requirement for Hedgehog signaling in small cell lung cancer. Nature Medicine. 2011 Nov;17(11):1504–8.

180. Singh S, Wang L, Schaff DL, Sutcliffe MD, Koeppel AF, Kim J, et al. In situ 10-cell RNA sequencing in tissue and tumor biopsy samples. Sci Rep. 2019 Mar 20;9(1):1–15.

181. Allinen M, Beroukhim R, Cai L, Brennan C, Lahti-Domenici J, Huang H, et al. Molecular characterization of the tumor microenvironment in breast cancer. Cancer Cell. 2004 Jul;6(1):17–32.

182. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000 Feb 3;403(6769):503–11.

183. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. Nat Genet. 2003 Jan;33(1):49–54.

184. Place AE, Jin Huh S, Polyak K. The microenvironment in breast cancer progression: biology and implications for treatment. Breast Cancer Res. 2011;13(6):227.

185. Carmona-Fontaine C, Bucci V, Akkari L, Deforet M, Joyce JA, Xavier JB. Emergence of spatial structure in the tumor microenvironment due to the Warburg effect. Proc Natl Acad Sci U S A. 2013 Nov 26;110(48):19402–7.

186. Cai DL, Jin LP. Immune Cell Population in Ovarian Tumor Microenvironment. J Cancer. 2017;8(15):2915–23.

187. Yuan Y. Spatial Heterogeneity in the Tumor Microenvironment. Cold Spring Harb Perspect Med. 2016 Aug 1;6(8).

188. Maley CC, Galipeau PC, Finley JC, Wongsurawat VJ, Li X, Sanchez CA, et al. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nat Genet. 2006 Apr;38(4):468–73.

189. Proia TA, Keller PJ, Gupta PB, Klebba I, Jones AD, Sedic M, et al. Genetic predisposition directs breast cancer phenotype by dictating progenitor cell fate. Cell Stem Cell. 2011 Feb 4;8(2):149–63.

190. Adam M, Potter AS, Potter SS. Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development. Development. 2017 Oct 1;144(19):3625–32.

191. Peng G, Suo S, Chen J, Chen W, Liu C, Yu F, et al. Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. Dev Cell. 2016 Mar 21;36(6):681–97.

192. Chen J, Suo S, Tam PP, Han J-DJ, Peng G, Jing N. Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. Nature Protocols. 2017 Mar;12(3):566–80.

193. Pereira M, Birtele M, Shrigley S, Benitez JA, Hedlund E, Parmar M, et al. Direct Reprogramming of Resident NG2 Glia into Neurons with Properties of Fast-Spiking Parvalbumin-Containing Interneurons. Stem Cell Reports. 2017 Sep 12;9(3):742–51.

194. Bhargava V, Head SR, Ordoukhanian P, Mercola M, Subramaniam S. Technical variations in low-input RNA-seq methodologies. Sci Rep. 2014 Jan 14;4:3678.

195. Picelli S, Bjorklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013 Nov;10(11):1096–8.

196. Islam S, Kjallquist U, Moliner A, Zajac P, Fan JB, Lonnerberg P, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011 Jul;21(7):1160–7.

197. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, et al. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. Genome Res. 2014 Mar 1;24(3):496–510.

198. Narayanan M, Martins AJ, Tsang JS. Robust Inference of Cell-to-Cell Expression Variations from Single- and K-Cell Profiling. PLOS Computational Biology. 2016 Jul 20;12(7):e1005016.

199. Martins AJ, Narayanan M, Prüstel T, Fixsen B, Park K, Gottschalk RA, et al. Environment Tunes Propagation of Cell-to-Cell Variation in the Human Macrophage Gene Network. Cell Syst. 2017 Apr 26;4(4):379-392.e12.

200. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013 Jun 13;498(7453):236–40.

201. Wang L, Brugge JS, Janes KA. Intersection of FOXO- and RUNX1-mediated gene expression programs in single breast epithelial cells during morphogenesis and tumor progression. Proc Natl Acad Sci U S A. 2011 Oct 4;108(40):E803-12.

202. Bajikar SS, Wang C-C, Borten MA, Pereira EJ, Atkins KA, Janes KA. Tumor-Suppressor Inactivation of GDF11 Occurs by Precursor Sequestration in Triple-Negative Breast Cancer. Dev Cell. 2017 20;43(4):418-435.e13.

203. Picelli S, Bjorklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 2014 Dec;24(12):2033–40.

204. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014;9(1):e78644.

205. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell. 2015 May 21;58(4):610–20.

206. Brady G, Iscove NN. Construction of cDNA libraries from single cells. Methods Enzymol. 1993;225:611–23.

207. Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, et al. Laser capture microdissection. Science. 1996 Nov 8;274(5289):998–1001.

208. Espina V, Wulfkuhle JD, Calvert VS, VanMeter A, Zhou W, Coukos G, et al. Laser-capture microdissection. Nat Protoc. 2006;1(2):586–603.

209. Kreklywich CN, Smith PP, Jones CB, Cornea A, Orloff SL, Streblow DN. Fluorescence-based laser capture microscopy technology facilitates identification of critical in vivo cytomegalovirus transcriptional programs. Methods Mol Biol. 2014;1119:217–37.

210. Murakami H, Liotta L, Star RA. IF-LCM: laser capture microdissection of immunofluorescently defined cells for mRNA analysis rapid communication. Kidney Int. 2000 Sep;58(3):1346–53.

211. Shaner NC, Campbell RE, Steinbach PA, Giepmans BN, Palmer AE, Tsien RY. Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. Nat Biotechnol. 2004 Dec;22(12):1567–72.

212. Galvao RP, Kasina A, McNeill RS, Harbin JE, Foreman O, Verhaak RG, et al. Transformation of quiescent adult oligodendrocyte precursor cells into malignant glioma through a multistep reactivation process. Proc Natl Acad Sci U S A. 2014 Oct 7;111(40):E4214-23.

213. Zong H, Espinosa JS, Su HH, Muzumdar MD, Luo L. Mosaic analysis with double markers in mice. Cell. 2005 May 6;121(3):479–92.

214. Liu C, Sage JC, Miller MR, Verhaak RG, Hippenmeyer S, Vogel H, et al. Mosaic analysis with double markers reveals tumor cell of origin in glioma. Cell. 2011 Jul 22;146(2):209–21.

215. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57–63.

216. Cronin M, Ghosh K, Sistare F, Quackenbush J, Vilker V, O'Connell C. Universal RNA reference materials for gene expression. Clin Chem. 2004 Aug;50(8):1464–71.

217. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, et al. The External RNA Controls Consortium: a progress report. Nat Methods. 2005 Oct;2(10):731–4.

218. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011 Sep;21(9):1543–51.

219. Lusk RW. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. PLoS One. 2014;9(10):e110808.

220. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016 Feb 17;17:29.

221. Seqc Maqc-Iii Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol. 2014 Sep;32(9):903–14.

222. Kurimoto K, Yabuta Y, Ohinata Y, Ono Y, Uno KD, Yamada RG, et al. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. Nucleic Acids Res. 2006;34(5):e42.

223. Schaffer BE, Park KS, Yiu G, Conklin JF, Lin C, Burkhart DL, et al. Loss of p130 accelerates tumor development in a mouse model for human small-cell lung carcinoma. Cancer Res. 2010 May 15;70(10):3877–83.

224. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 2010;11(12):R119.

225. DeAngelis MM, Wang DG, Hawkins TL. Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Res. 1995 Nov 25;23(22):4742–3.

226. Chen L, Sun F, Yang X, Jin Y, Shi M, Wang L, et al. Correlation between RNA-Seq and microarrays results using TCGA data. Gene. 2017 Sep 10;628:200–4.

227. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, Manno GL, Juréus A, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015 Mar 6;347(6226):1138–42.

228. Marques S, Zeisel A, Codeluppi S, van Bruggen D, Mendanha Falcao A, Xiao L, et al. Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. Science. 2016 Jun 10;352(6291):1326–9.

229. Montoro DT, Haber AL, Biton M, Vinarsky V, Lin B, Birket SE, et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. Nature. 2018 Aug;560(7718):319–24.

230. Nguyen QH, Pervolarakis N, Blake K, Ma D, Davis RT, James N, et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. Nature Communications. 2018 May 23;9(1):2028.

231. Bose S, Wan Z, Carr A, Rizvi AH, Vieira G, Pe'er D, et al. Scalable microfluidics for single-cell RNA printing and sequencing. Genome Biol. 2015 Jun 6;16:120.

232. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Aug 4;12:323.

233. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. Elife. 2017 Dec 5;6.

234. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017 Aug 18;357(6352):661–7.

235. Halpern KB, Shenhav R, Matcovitch-Natan O, Toth B, Lemze D, Golan M, et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. Nature. 2017 Feb 16;542(7641):352–6.

236. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015 Aug 19;

237. Fincher CT, Wurtzel O, de Hoog T, Kravarik KM, Reddien PW. Cell type transcriptome atlas for the planarian Schmidtea mediterranea. Science. 2018 May 25;360(6391).

238. Plass M, Solana J, Wolf FA, Ayoub S, Misios A, Glazar P, et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science. 2018 May 25;360(6391).

239. Hunter F, Xie J, Trimble C, Bur M, Li KCP. Rhodamine-RCA in vivo labeling guided laser capture microdissection of cancer functional angiogenic vessels in a murine squamous cell carcinoma mouse model. Mol Cancer. 2006 Feb 3;5:5.

240. Kretzschmar K, Watt FM. Lineage tracing. Cell. 2012 Jan 20;148(1–2):33–45.

241. Fend F, Emmert-Buck MR, Chuaqui R, Cole K, Lee J, Liotta LA, et al. Immuno-LCM: laser capture microdissection of immunostained frozen sections for mRNA analysis. Am J Pathol. 1999 Jan;154(1):61–6.

242. Steu S, Baucamp M, von Dach G, Bawohl M, Dettwiler S, Storz M, et al. A procedure for tissue freezing and processing applicable to both intra-operative frozen section diagnosis and tissue banking in surgical pathology. Virchows Arch. 2008 Mar;452(3):305–12.

243. Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002 Dec 5;420(6915):520–62.

244. Shedlock AM, Okada N. SINE insertions: powerful tools for molecular systematics. Bioessays. 2000 Feb;22(2):148–60.

245. Tsirigos A, Rigoutsos I. Alu and b1 repeats have been selectively retained in the upstream and intronic regions of genes of specific functional classes. PLoS Comput Biol. 2009 Dec;5(12):e1000610.

246. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015 May 21;161(5):1202–14.

247. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012 Sep 27;2(3):666–73.

248. Nam DK, Lee S, Zhou G, Cao X, Wang C, Clark T, et al. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. Proc Natl Acad Sci U S A. 2002 Apr 30;99(9):6152–6.

249. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018 Aug;560(7719):494–8.

250. Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2012 Jan;9(1):72–4.

251. Miller-Jensen K, Janes KA, Brugge JS, Lauffenburger DA. Common effector processing mediates cell-specific responses to stimuli. Nature. 2007 Aug 2;448(7153):604–8.

252. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008 Jul 1;24(13):1547–8.

253. Singh S, Sutcliffe MD, Repich K, Atkins KA, Harvey JA, Janes KA. Pan-cancer Drivers are Recurrent Transcriptional Regulatory Heterogeneities in Early-stage Luminal Breast Cancer. bioRxiv. 2020 Apr 2;2020.03.30.017186.

254. Surveillance, Epidemiology, and End Results (SEER) Program.

255. Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science. 2015 Jan 2;347(6217):78–81.

256. Medina D. The mammary gland: a unique organ for the study of development and tumorigenesis. J Mammary Gland Biol Neoplasia. 1996 Jan;1(1):5–19.

257. Cerchiari AE, Garbe JC, Jee NY, Todhunter ME, Broaders KE, Peehl DM, et al. A strategy for tissue self-organization that is robust to cellular heterogeneity and plasticity. Proc Natl Acad Sci U S A. 2015 Feb 17;112(7):2287–92.

258. Davis A, Gao R, Navin N. Tumor evolution: Linear, branching, neutral or punctuated? Biochim Biophys Acta Rev Cancer. 2017 Apr;1867(2):151–61.

259. Machado L, Esteves de Lima J, Fabre O, Proux C, Legendre R, Szegedi A, et al. In Situ Fixation Redefines Quiescence and Early Activation of Skeletal Muscle Stem Cells. Cell Rep. 2017 Nov 14;21(7):1982–93.

260. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018 Apr 5;173(2):371-385 e18.

261. Mak MP, Tong P, Diao L, Cardnell RJ, Gibbons DL, William WN, et al. A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. Clin Cancer Res. 2016 Feb 1;22(3):609–20.

262. Icgc Tcga Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020 Feb;578(7793):82–93.

263. Matsuno RK, Anderson WF, Yamamoto S, Tsukuma H, Pfeiffer RM, Kobayashi K, et al. Early- and late-onset breast cancer types among women in the United States and Japan. Cancer Epidemiol Biomarkers Prev. 2007 Jul;16(7):1437–42.

264. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet. 2017 May;49(5):708–18.

265. Elyada E, Bolisetty M, Laise P, Flynn WF, Courtois ET, Burkhart RA, et al. Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts. Cancer Discov. 2019 Aug;9(8):1102–23.

266. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2018 Dec 3;

267. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000 Aug 17;406(6797):747–52.

268. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell. 2015 Oct 8;163(2):506–19.

269. Gendoo DMA, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics. 2016 Apr 1;32(7):1097–9.

270. Schaff DL, Singh S, Kim K-B, Sutcliffe MD, Park K-S, Janes KA. Fragmentation of Small-cell Lung Cancer Regulatory States in Heterotypic Microenvironments. bioRxiv. 2020 Mar 31;2020.03.30.017210.

271. Sutcliffe MD, Galvao RP, Wang L, Kim J, Singh S, Zong H, et al. Single-cell Bottlenecks and Dead-ends During Glioma Premalignancy. bioRxiv. 2020 Mar 31;2020.03.30.017228.

272. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015 Feb;33(2):155–60.

273. Santos A, Wernersson R, Jensen LJ. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. Nucleic Acids Res. 2015 Jan;43(Database issue):D1140-4.

274. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The life history of 21 breast cancers. Cell. 2012 May 25;149(5):994–1007.

275. Shlien A, Raine K, Fuligni F, Arnold R, Nik-Zainal S, Dronov S, et al. Direct Transcriptional Consequences of Somatic Mutation in Breast Cancer. Cell Rep. 2016 Aug 16;16(7):2032–46.

276. Aitken SJ, Ibarra-Soria X, Kentepozidou E, Flicek P, Feig C, Marioni JC, et al. CTCF maintains regulatory homeostasis of cancer pathways. Genome Biol. 2018 Aug 7;19(1):106.

277. Takaku M, Grimm SA, Roberts JD, Chrysovergis K, Bennett BD, Myers P, et al. GATA3 zinc finger 2 mutations reprogram the breast cancer transcriptional network. Nat Commun. 2018 Mar 13;9(1):1059.

278. Hu Z, Artibani M, Alsaadi A, Wietek N, Morotti M, Shi T, et al. The Repertoire of Serous Ovarian Cancer Non-genetic Heterogeneity Revealed by Single-Cell Sequencing of Normal Fallopian Tube Epithelial Cells. Cancer Cell. 2020 Feb 10;37(2):226-242.e7.

279. Savci-Heijink CD, Halfwerk H, Hooijer GKJ, Koster J, Horlings HM, Meijer SL, et al. Epithelial-to-mesenchymal transition status of primary breast carcinomas and its correlation with metastatic behavior. Breast Cancer Res Treat. 2019 Apr;174(3):649–59.

280. Tan TZ, Miow QH, Miki Y, Noda T, Mori S, Huang RY, et al. Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. EMBO Mol Med. 2014 Oct;6(10):1279–93.

281. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature. 2020 Feb;578(7793):102–11.

282. Wang Q, Zhou Y, Weiss HL, Chow CW, Evers BM. NFATc1 regulation of TRAIL expression in human intestinal cells. PLoS One. 2011;6(5):e19882.

283. Zhang W, Geiman DE, Shields JM, Dang DT, Mahatan CS, Kaestner KH, et al. The gut-enriched Kruppel-like factor (Kruppel-like factor 4) mediates the transactivating effect of p53 on the p21WAF1/Cip1 promoter. J Biol Chem. 2000 Jun 16;275(24):18391–8.

284. Aksoy I, Giudice V, Delahaye E, Wianny F, Aubry M, Mure M, et al. Klf4 and Klf5 differentially inhibit mesoderm and endoderm differentiation in embryonic stem cells. Nat Commun. 2014 Apr 28;5:3719.

285. Santos Guasch GL, Beeler JS, Marshall CB, Shaver TM, Sheng Q, Johnson KN, et al. p73 Is Required for Ovarian Follicle Development and Regulates a Gene Network Involved in Cell-to-Cell Adhesion. iScience. 2018 Oct 26;8:236–49.

286. Fischer M. Census and evaluation of p53 target genes. Oncogene. 2017 Jul 13;36(28):3943–56.

287. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014 Aug 14;158(4):929–44.

288. Wagner J, Rapsomaniki MA, Chevrier S, Anzeneder T, Langwieder C, Dykgers A, et al. A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. Cell. 2019 May 16;177(5):1330-1345 e18.

289. Gut G, Herrmann MD, Pelkmans L. Multiplexed protein maps link subcellular organization to cellular states. Science. 2018 Aug 3;361(6401).

290. Stahl PL, Salmen F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016 Jul 1;353(6294):78–82.

291. Kumar MP, Du J, Lagoudas G, Jiao Y, Sawyer A, Drummond DC, et al. Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. Cell Reports. 2018 Nov 6;25(6):1458-1468.e4.

292. Li PX, Wong J, Ayed A, Ngo D, Brade AM, Arrowsmith C, et al. Placental transforming growth factor-beta is a downstream mediator of the growth arrest and apoptotic response of tumor cells to DNA damage and p53 overexpression. J Biol Chem. 2000 Jun 30;275(26):20127–35.

293. Mullican SE, Lin-Schmidt X, Chin CN, Chavez JA, Furman JL, Armstrong AA, et al. GFRAL is the receptor for GDF15 and the ligand promotes weight loss in mice and nonhuman primates. Nat Med. 2017 Oct;23(10):1150–7.

294. Chang GW, Hsiao CC, Peng YM, Vieira Braga FA, Kragten NA, Remmerswaal EB, et al. The Adhesion G Protein-Coupled Receptor GPR56/ADGRG1 Is an Inhibitory Receptor on Human NK Cells. Cell Rep. 2016 May 24;15(8):1757–70.

295. Lacher MD, Bauer G, Fury B, Graeve S, Fledderman EL, Petrie TD, et al. SV-BR-1-GM, a Clinically Effective GM-CSF-Secreting Breast Cancer Cell Line, Expresses an Immune Signature and Directly Activates CD4(+) T Lymphocytes. Front Immunol. 2018;9:776.

296. Yao HP, Zhou YQ, Zhang R, Wang MH. MSP-RON signalling in cancer: pathogenesis and therapeutic potential. Nat Rev Cancer. 2013 Jul;13(7):466–81.

297. Yang Y, Hsu JM, Sun L, Chan LC, Li CW, Hsu JL, et al. Palmitoylation stabilizes PD-L1 to promote breast tumor growth. Cell Res. 2019 Jan;29(1):83–6.

298. Tsujikawa T, Kumar S, Borkar RN, Azimi V, Thibault G, Chang YH, et al. Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis. Cell Rep. 2017 Apr 4;19(1):203–17.

299. van Seijen M, Lips EH, Thompson AM, Nik-Zainal S, Futreal A, Hwang ES, et al. Ductal carcinoma in situ: to treat or not to treat, that is the question. Br J Cancer. 2019 Aug;121(4):285–92.

300. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015 May 8;348(6235):648–60.

301. inferCNV of the Trinity CTAT Project.

302. van Es JH, Sato T, van de Wetering M, Lyubimova A, Yee Nee AN, Gregorieff A, et al. Dll1+ secretory progenitor cells revert to stem cells upon crypt damage. Nat Cell Biol. 2012 Oct;14(10):1099–104.

303. McEvoy J, Flores-Otero J, Zhang J, Nemeth K, Brennan R, Bradley C, et al. Coexpression of normally incompatible developmental pathways in retinoblastoma genesis. Cancer Cell. 2011 Aug 16;20(2):260–75.

304. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016 Nov 10;539(7628):309–13.

305. Yao M, Ventura PB, Jiang Y, Rodriguez FJ, Wang L, Perry JSA, et al. Astrocytic trans-Differentiation Completes a Multicellular Paracrine Feedback Loop Required for Medulloblastoma Tumor Growth. Cell. 2020 Feb 6;180(3):502-520 e19.

306. Robinson DR, Wu YM, Lonigro RJ, Vats P, Cobain E, Everett J, et al. Integrative clinical genomics of metastatic cancer. Nature. 2017 Aug 17;548(7667):297–303.

307. Garg A, Sui P, Verheyden JM, Young LR, Sun X. Consider the lung as a sensory organ: A tip from pulmonary neuroendocrine cells. Curr Top Dev Biol. 2019;132:67–89.

308. Kuo CS, Krasnow MA. Formation of a Neurosensory Organ by Epithelial Cell Slithering. Cell. 2015 Oct 8;163(2):394–405.

309. Dvorak HF. Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing. N Engl J Med. 1986 Dec 25;315(26):1650–9.

310. Peifer M, Fernández-Cuesta L, Sos ML, George J, Seidel D, Kasper LH, et al. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nat Genet. 2012 Oct;44(10):1104–10.

311. Jain AK, Allton K, Iacovino M, Mahen E, Milczarek RJ, Zwaka TP, et al. p53 regulates cell cycle and microRNAs to promote differentiation of human embryonic stem cells. PLoS Biol. 2012;10(2):e1001268.

312. Kareta MS, Gorges LL, Hafeez S, Benayoun BA, Marro S, Zmoos AF, et al. Inhibition of pluripotency networks by the Rb tumor suppressor restricts reprogramming and tumorigenesis. Cell Stem Cell. 2015 Jan 8;16(1):39–50.

313. Meuwissen R, Linn SC, Linnoila RI, Zevenhoven J, Mooi WJ, Berns A. Induction of small cell lung cancer by somatic inactivation of both Trp53 and Rb1 in a conditional mouse model. Cancer Cell. 2003 Sep;4(3):181–9.

314. Cui M, Augert A, Rongione M, Conkrite K, Parazzoli S, Nikitin AY, et al. PTEN is a potent suppressor of small cell lung cancer. Mol Cancer Res. 2014 May;12(5):654–9.

315. McFadden DG, Papagiannakopoulos T, Taylor-Weiner A, Stewart C, Carter SL, Cibulskis K, et al. Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. Cell. 2014 Mar 13;156(6):1298–311.

316. Jia D, Augert A, Kim DW, Eastwood E, Wu N, Ibrahim AH, et al. Crebbp Loss Drives Small Cell Lung Cancer and Increases Sensitivity to HDAC Inhibition. Cancer Discov. 2018 Nov;8(11):1422–37.

317. Yang D, Denny SK, Greenside PG, Chaikovsky AC, Brady JJ, Ouadah Y, et al. Intertumoral Heterogeneity in SCLC Is Influenced by the Cell Type of Origin. Cancer Discov. 2018 Oct;8(10):1316–31.

318. Denny SK, Yang D, Chuang C-H, Brady JJ, Lim JS, Grüner BM, et al. Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility. Cell. 2016 Jul 14;166(2):328–42.

319. Calbo J, van Montfort E, Proost N, van Drunen E, Beverloo HB, Meuwissen R, et al. A functional role for tumor cell heterogeneity in a mouse model of small cell lung cancer. Cancer Cell. 2011 Feb 15;19(2):244–56.

320. Kagohashi K, Satoh H, Ishikawa H, Ohtsuka M, Sekizawa K. Liver metastasis at the time of initial diagnosis of lung cancer. Med Oncol. 2003;20(1):25–8.

321. Ponath P, Menezes D, Pan C, Chen B, Oyasu M, Strachan D, et al. A Novel, Fully Human Anti-fucosyl-GM1 Antibody Demonstrates Potent In Vitro and In Vivo Antitumor Activity in Preclinical Models of Small Cell Lung Cancer. Clin Cancer Res. 2018 Oct 15;24(20):5178–89.

322. Maretto S, Cordenonsi M, Dupont S, Braghetta P, Broccoli V, Hassan AB, et al. Mapping Wnt/beta-catenin signaling during mouse development and in colorectal tumors. Proc Natl Acad Sci U S A. 2003 Mar 18;100(6):3299–304.

323. Kutay U, Bischoff FR, Kostka S, Kraft R, Gorlich D. Export of importin alpha from the nucleus is mediated by a specific nuclear transport factor. Cell. 1997 Sep 19;90(6):1061–71.

324. Giordano F, Saheki Y, Idevall-Hagren O, Colombo SF, Pirruccello M, Milosevic I, et al. PI(4,5)P(2)-dependent and Ca(2+)-regulated ER-PM interactions mediated by the extended synaptotagmins. Cell. 2013 Jun 20;153(7):1494–509.

325. Chen LY, Redon S, Lingner J. The human CST complex is a terminator of telomerase activity. Nature. 2012 Aug 23;488(7412):540–4.

326. Haigis MC, Guarente LP. Mammalian sirtuins--emerging roles in physiology, aging, and calorie restriction. Genes Dev. 2006 Nov 1;20(21):2913–21.

327. Poulsen EG, Steinhauer C, Lees M, Lauridsen AM, Ellgaard L, Hartmann-Petersen R. HUWE1 and TRIP12 collaborate in degradation of ubiquitin-fusion proteins and misframed ubiquitin. PLoS One. 2012;7(11):e50548.

328. Pereira EJ, Burns JS, Lee CY, Marohl T, Calderon D, Wang L, Atkins KA, Wang CC, Janes KA. Sporadic activation of an oxidative stress–dependent NRF2-p53 signaling network in breast epithelial spheroids and premalignancies. Science Signaling. 2020 Apr 14;13(627).

329. Desai TJ, Brownfield DG, Krasnow MA. Alveolar progenitor and stem cells in lung development, renewal and cancer. Nature. 2014 Mar 13;507(7491):190–4.

330. Das KC, Lewis-Molock Y, White CW. Thiol modulation of TNF alpha and IL-1 induced MnSOD gene expression and activation of NF-kappa B. Mol Cell Biochem. 1995 Jul 5;148(1):45–57.

331. Mackay F, Majeau GR, Hochman PS, Browning JL. Lymphotoxin beta receptor triggering induces activation of the nuclear factor kappaB transcription factor in some cell types. J Biol Chem. 1996 Oct 4;271(40):24934–8.

332. Tabula Muris C, Overall coordination, Logistical coordination, Organ collection, processing, Library preparation, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018 Oct;562(7727):367–72.

333. Mikulak J, Bruni E, Oriolo F, Di Vito C, Mavilio D. Hepatic Natural Killer Cells: Organ-Specific Sentinels of Liver Immune Homeostasis and Physiopathology. Front Immunol. 2019;10:946.

334. Budzynski W, Radzikowski C. Cytotoxic cells in immunodeficient athymic mice. Immunopharmacol Immunotoxicol. 1994 Aug;16(3):319–46.

335. Burns K, Clatworthy J, Martin L, Martinon F, Plumpton C, Maschera B, et al. Tollip, a new component of the IL-1RI pathway, links IRAK to the IL-1 receptor. Nat Cell Biol. 2000 Jun;2(6):346–51.

336. Shibuya H, Yamaguchi K, Shirakabe K, Tonegawa A, Gotoh Y, Ueno N, et al. TAB1: an activator of the TAK1 MAPKKK in TGF-beta signal transduction. Science. 1996 May 24;272(5265):1179–82.

337. Yasuda K, Muto T, Kawagoe T, Matsumoto M, Sasaki Y, Matsushita K, et al. Contribution of IL-33-activated type II innate lymphoid cells to pulmonary eosinophilia in intestinal nematode-infected mice. Proc Natl Acad Sci U S A. 2012 Feb 28;109(9):3451–6.

338. Pichery M, Mirey E, Mercier P, Lefrancais E, Dujardin A, Ortega N, et al. Endogenous IL-33 is highly expressed in mouse epithelial barrier tissues, lymphoid organs, brain, embryos, and inflamed tissues: in situ analysis using a novel Il-33-LacZ gene trap reporter strain. J Immunol. 2012 Apr 1;188(7):3488–95.

339. Yao Y, Nowak S, Yochelis A, Garfinkel A, Bostrom KI. Matrix GLA protein, an inhibitory morphogen in pulmonary vascular development. J Biol Chem. 2007 Oct 12;282(41):30131–42.

340. Liu W, Morgan KM, Pine SR. Activation of the Notch1 Stem Cell Signaling Pathway during Routine Cell Line Subculture. Front Oncol. 2014;4:211.

341. Bray SJ. Notch signalling in context. Nat Rev Mol Cell Biol. 2016 Nov;17(11):722–35.

342. Jiang L, Huang J, Higgs BW, Hu Z, Xiao Z, Yao X, et al. Genomic Landscape Survey Identifies SRSF1 as a Key Oncodriver in Small Cell Lung Cancer. PLOS Genetics. 2016 Apr 19;12(4):e1005895.

343. Frechin M, Stoeger T, Daetwyler S, Gehin C, Battich N, Damm EM, et al. Cell-intrinsic adaptation of lipid composition to local crowding drives social behaviour. Nature. 2015 Jul 2;523(7558):88–91.

344. Schafer ZT, Grassian AR, Song L, Jiang Z, Gerhart-Hines Z, Irie HY, et al. Antioxidant and oncogene rescue of metabolic defects caused by loss of matrix attachment. Nature. 2009 Sep 3;461(7260):109–13.

345. Shin SI, Freedman VH, Risser R, Pollack R. Tumorigenicity of virus-transformed cells in nude mice is correlated specifically with anchorage independent growth in vitro. Proc Natl Acad Sci U S A. 1975 Nov;72(11):4435–9.

346. Rackley CR, Stripp BR. Building and maintaining the epithelium of the lung. J Clin Invest. 2012 Aug;122(8):2724–30.

347. Combs SE, Hancock JG, Boffa DJ, Decker RH, Detterbeck FC, Kim AW. Bolstering the case for lobectomy in stages I, II, and IIIA small-cell lung cancer using the National Cancer Data Base. J Thorac Oncol. 2015 Feb;10(2):316–23.

348. Benson RE, Rosado-de-Christenson ML, Martinez-Jimenez S, Kunin JR, Pettavel PP. Spectrum of pulmonary neuroendocrine proliferations and neoplasms. Radiographics. 2013 Oct;33(6):1631–49.

349. Zhang F, Xu X, Zhou B, He Z, Zhai Q. Gene expression profile change and associated physiological and pathological effects in mouse liver induced by fasting and refeeding. PLoS One. 2011;6(11):e27553.

350. Fish RJ, Neerman-Arbez M. Fibrinogen gene regulation. Thromb Haemost. 2012 Sep;108(3):419–26.

351. Yanagimoto T, Itoh S, Sawada M, Kamataki T. Mouse cytochrome P450 (Cyp3a11): predominant expression in liver and capacity to activate aflatoxin B1. Arch Biochem Biophys. 1997 Apr 15;340(2):215–8.

352. Salier JP, Chan P, Raguenez G, Zwingman T, Erickson RP. Developmentally regulated transcription of the four liver-specific genes for inter-alpha-inhibitor family in mouse. Biochem J. 1993 Nov 15;296 (Pt 1):85–91.

353. Leduc MS, Shimmin LC, Klos KL, Hanis C, Boerwinkle E, Hixson JE. Comprehensive evaluation of apolipoprotein H gene (APOH) variation identifies novel associations with measures of lipid metabolism in GENOA. J Lipid Res. 2008 Dec;49(12):2648–56.

354. Franzen O, Gan LM, Bjorkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database (Oxford). 2019 Jan 1;2019.

355. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015 Dec 23;1(6):417–25.

356. Segouffin-Cariou C, Billaud M. Transforming ability of MEN2A-RET requires activation of the phosphatidylinositol 3-kinase/AKT signaling pathway. J Biol Chem. 2000 Feb 4;275(5):3568–76.

357. Seto KK, Andrulis IL. Atypical protein kinase C zeta: potential player in cell survival and cell migration of ovarian cancer. PloS one. 2015;10(4).

358. Hwang Y, Kim LC, Song W, Edwards DN, Cook RS, Chen J. Disruption of the scaffolding function of mLST8 selectively inhibits mTORC2 assembly and function and suppresses mTORC2-dependent tumor growth in vivo. Cancer research. 2019 Jul 1;79(13):3178-84.

359. Kim S-H, Juhnn Y-S, Song Y-S. Akt involvement in paclitaxel chemoresistance of human ovarian cancer cells. Ann N Y Acad Sci. 2007 Jan;1095:82–9.

360. Tokunaga E, Kimura Y, Mashino K, Oki E, Kataoka A, Ohno S, et al. Activation of PI3K/Akt signaling and hormone resistance in breast cancer. Breast Cancer. 2006 Apr 1;13(2):137–44.

361. Wang Z, Li Y, Kong D, Banerjee S, Ahmad A, Azmi AS, et al. Acquisition of Epithelial-Mesenchymal Transition Phenotype of Gemcitabine-Resistant Pancreatic Cancer Cells Is

Linked with Activation of the Notch Signaling Pathway. Cancer Res. 2009 Mar 15;69(6):2400–7.

362. Nurwidya F, Takahashi F, Murakami A, Takahashi K. Epithelial Mesenchymal Transition in Drug Resistance and Metastasis of Lung Cancer. Cancer Res Treat. 2012 Sep;44(3):151–6.

363. Zheng X, Carstens JL, Kim J, Scheible M, Kaye J, Sugimoto H, et al. Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. Nature. 2015 Nov 26;527(7579):525–30.

364. Dong Z, Zou J, Li J, Pang Y, Liu Y, Deng C, et al. MYST1/KAT8 contributes to tumor progression by activating EGFR signaling in glioblastoma cells. Cancer Medicine. 2019;8(18):7793–808.

365. Patani N, Jiang WG, Newbold RF, Mokbel K. Histone-modifier gene expression profiles are associated with pathological and clinical outcomes in human breast cancer. Anticancer Res. 2011 Dec;31(12):4115–25.

366. Pfister S, Rea S, Taipale M, Mendrzyk F, Straub B, Ittrich C, et al. The histone acetyltransferase hMOF is frequently downregulated in primary breast carcinoma and medulloblastoma and constitutes a biomarker for clinical outcome in medulloblastoma. International Journal of Cancer. 2008;122(6):1207–13.

367. Jang ER, Lim S-J, Lee ES, Jeong G, Kim T-Y, Bang Y-J, et al. The histone deacetylase inhibitor trichostatin A sensitizes estrogen receptor α -negative breast cancer cells to tamoxifen. Oncogene. 2004 Mar;23(9):1724–36.

368. Légaré S, Basik M. Minireview: The Link Between ERα Corepressors and Histone Deacetylases in Tamoxifen Resistance in Breast Cancer. Mol Endocrinol. 2016 Sep 1;30(9):965–76.

369. Gardner EE, Lok BH, Schneeberger VE, Desmeules P, Miles LA, Arnold PK, et al. Chemosensitive Relapse in Small Cell Lung Cancer Proceeds through an EZH2-SLFN11 Axis. Cancer Cell. 2017 Feb 13;31(2):286–99.

370. Bao J, Li M, Liang S, Yang Y, Wu J, Zou Q, Fang S, Chen S, Guo L. Integrated high-throughput analysis identifies super enhancers associated with chemoresistance in SCLC. BMC medical genomics. 2019 Dec;12(1):67.

371. Konieczkowski DJ, Johannessen CM, Garraway LA. A Convergence-Based Framework for Cancer Drug Resistance. Cancer Cell. 2018 May 14;33(5):801–15.

372. Rozenblatt-Rosen O, Regev A, Oberdoerffer P, Nawy T, Hupalowska A, Rood JE, et al. The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell. 2020 Apr 16;181(2):236–49.

373. Solé X, Bonifaci N, López-Bigas N, Berenguer A, Hernández P, Reina O, et al. Biological Convergence of Cancer Signatures. PLOS ONE. 2009 Feb 20;4(2):e4544.

374. Schröder B. The multifaceted roles of the invariant chain CD74 — More than just a chaperone. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 2016 Jun 1;1863(6, Part A):1269–81.

375. Uhlén M, Björling E, Agaton C, Szigyarto CA-K, Amini B, Andersen E, et al. A human protein atlas for normal and cancer tissues based on antibody proteomics. Mol Cell Proteomics. 2005 Dec;4(12):1920–32.

376. Valiño-Rivas L, Cuarental L, Grana O, Bucala R, Leng L, Sanz A, et al. TWEAK increases CD74 expression and sensitizes to DDT proinflammatory actions in tubular cells. PLOS ONE. 2018 Jun 20;13(6):e0199391.

377. Borghese F, Clanchy FI. CD74: an emerging opportunity as a therapeutic target in cancer and autoimmune disease. Expert Opinion on Therapeutic Targets. 2011 Mar 1;15(3):237–51.

378. Ssadh HA, Abdulmonem WA. Immunophenotyping of the cluster of differentiation 74, migration inhibitory factor, and cluster of differentiation 44 expression on human breast cancer-derived cell lines. Int J Health Sci (Qassim). 2019;13(2):17–24.

379. Tanese K, Hashimoto Y, Berkova Z, Wang Y, Samaniego F, Lee JE, et al. Cell Surface CD74-MIF Interactions Drive Melanoma Survival in Response to Interferon-γ. J Invest Dermatol. 2015 Nov;135(11):2775–84.

380. Leng L, Metz CN, Fang Y, Xu J, Donnelly S, Baugh J, et al. MIF Signal Transduction Initiated by Binding to CD74. J Exp Med. 2003 Jun 2;197(11):1467–76.

381. Maharshak N, Cohen S, Lantner F, Hart G, Leng L, Bucala R, et al. CD74 is a survival receptor on colon epithelial cells. World J Gastroenterol. 2010 Jul 14;16(26):3258–66.

382. Crow M, Lim N, Ballouz S, Pavlidis P, Gillis J. Predictability of human differential gene expression. Proc Natl Acad Sci U S A. 2019 Mar 26;116(13):6491–500.

383. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DSA, Nobel AB, et al. Concordance among Gene-Expression–Based Predictors for Breast Cancer. New England Journal of Medicine. 2006 Aug 10;355(6):560–9.

384. Wang T, Blumhagen R, Lao U, Kuo Y, Edgar BA. LST8 regulates cell growth via target-of-rapamycin complex 2 (TORC2). Mol Cell Biol. 2012 Jun;32(12):2203–13.

385. Sachs N, de Ligt J, Kopper O, Gogola E, Bounova G, Weeber F, et al. A Living Biobank of Breast Cancer Organoids Captures Disease Heterogeneity. Cell. 2018 Jan 11;172(1):373-386.e10.

386. Walsh AJ, Cook RS, Sanders ME, Aurisicchio L, Ciliberto G, Arteaga CL, et al. Quantitative optical imaging of primary tumor organoid metabolism predicts drug response in breast cancer. Cancer Res. 2014 Sep 15;74(18):5184–94.

387. Goldhammer N, Kim J, Timmermans-Wielenga V, Petersen OW. Characterization of organoid cultured human breast cancer. Breast Cancer Research. 2019 Dec 11;21(1):141.

388. Rosenbluth JM, Schackmann RCJ, Gray GK, Selfors LM, Li CM-C, Boedicker M, et al. Organoid cultures from normal and cancer-prone human breast tissues preserve complex epithelial lineages. Nature Communications. 2020 Apr 6;11(1):1–14.

389. Favreau LV, Pickett CB. Transcriptional regulation of the rat NAD(P)H:quinone reductase gene. Identification of regulatory elements controlling basal level expression and inducible expression by planar aromatic compounds and phenolic antioxidants. J Biol Chem. 1991 Mar 5;266(7):4556–61.

390. Ross D, Siegel D. Functions of NQO1 in cellular protection and CoQ10 metabolism and its potential role as a redox sensitive molecular switch. Frontiers in physiology. 2017 Aug 24;8:595.

391. Oh E-T, Park HJ. Implications of NQO1 in cancer therapy. BMB Rep. 2015 Nov;48(11):609–17.

392. Yang Y, Zhang Y, Wu Q, Cui X, Lin Z, Liu S, et al. Clinical implications of high NQO1 expression in breast cancers. Journal of Experimental & Clinical Cancer Research. 2014 Feb 5;33(1):14.

393. McClelland M, Zhao L, Carskadon S, Arenberg D. Expression of CD74, the receptor for macrophage migration inhibitory factor, in non-small cell lung cancer. Am J Pathol. 2009 Feb;174(2):638–46.

394. Kwon AT, Arenillas DJ, Worsley Hunt R, Wasserman WW. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. G3 (Bethesda). 2012 Sep;2(9):987–1002.

395. Clarke DJB, Kuleshov MV, Schilder BM, Torre D, Duffy ME, Keenan AB, et al. eXpression2Kinases (X2K) Web: linking expression signatures to upstream cell signaling networks. Nucleic Acids Res. 2018 Jul 2;46(W1):W171–9.

396. Keenan AB, Torre D, Lachmann A, Leong AK, Wojciechowicz ML, Utti V, et al. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. Nucleic Acids Res. 2019 Jul 2;47(W1):W212–24.

397. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015;109:21 29 1-9.

398. Foote FW, Stewart FW. Lobular carcinoma in situ. Am J Pathol. 1941 Jul;17(4):491-496.3.

399. Shah V, Nowinski S, Levi D, Shinomiya I, Kebaier Ep Chaabouni N, Gillett C, et al. PIK3CA mutations are common in lobular carcinoma in situ, but are not a biomarker of progression. Breast Cancer Research. 2017 Jan 17;19(1):7.

400. Ginter PS, D'Alfonso TM. Current Concepts in Diagnosis, Molecular Features, and Management of Lobular Carcinoma In Situ of the Breast With a Discussion of Morphologic Variants. Archives of Pathology & Laboratory Medicine. 2017 Jun 2;141(12):1668–78.

401. Carniello JS, Giri D, De Brot M, Andrade V, King T. Multifocality and Bilaterality of Lobular Carcinoma In Situ in Women with Synchronous Breast Malignancies. American Journal of Clinical Pathology. 2016 Sep 1;146(suppl_1).

402. Petridis C, Arora I, Shah V, Moss CL, Mera A, Clifford A, et al. Frequency of Pathogenic Germline Variants in CDH1, BRCA2, CHEK2, PALB2, BRCA1, and TP53 in Sporadic Lobular Breast Cancer. Cancer Epidemiol Biomarkers Prev. 2019;28(7):1162–8.

403. King TA, Pilewskie M, Muhsen S, Patil S, Mautner SK, Park A, et al. Lobular Carcinoma in Situ: A 29-Year Longitudinal Experience Evaluating Clinicopathologic Features and Breast Cancer Risk. JCO. 2015 Sep 14;33(33):3945–52.

404. Sakr RA, Schizas M, Carniello JVS, Ng CKY, Piscuoglio S, Giri D, et al. Targeted capture massively parallel sequencing analysis of LCIS and invasive lobular cancer: Repertoire of somatic genetic alterations and clonal relationships. Mol Oncol. 2016 Feb;10(2):360–70.

405. Lee JY, Schizas M, Geyer FC, Selenica P, Piscuoglio S, Sakr RA, et al. Lobular Carcinomas *In Situ* Display Intralesion Genetic Heterogeneity and Clonal Evolution in the Progression to Invasive Lobular Carcinoma. Clinical Cancer Research. 2019 Jan 15;25(2):674–86.

406. Abner AL, Connolly JL, Recht A, Bornstein B, Nixon A, Hetelekidis S, et al. The relation between the presence and extent of lobular carcinoma in situ and the risk of local recurrence for patients with infiltrating carcinoma of the breast treated with conservative surgery and radiation therapy. Cancer. 2000 Mar 1;88(5):1072–7.

407. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. 2019 19;20(1):59.

408. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019 May;37(5):547–54.

409. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology. 2014 Apr;32(4):381–6.

410. Loeffler-Wirth H, Binder H, Willscher E, Gerber T, Kunz M. Pseudotime dynamics in melanoma single-cell transcriptomes reveals different mechanisms of tumor progression. Biology. 2018 Jun;7(2):23.

411. Casasent AK, Schalck A, Gao R, Sei E, Long A, Pangburn W, et al. Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. Cell. 2018 Jan 11;172(1):205-217.e12.

412. Andrade VP, Morrogh M, Qin L-X, Olvera N, Giri D, Muhsen S, et al. Gene expression profiling of lobular carcinoma in situ reveals candidate precursor genes for invasion. Molecular Oncology. 2015 Apr 1;9(4):772–82.