SINGLE-CELL REGULATORY HETEROGENEITIES IN BASAL-LIKE BREAST EPITHELIA

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 SAMEER SUBHASH BAJIKAR

May 2016

DEPARTMENT OF BIOMEDICAL ENGINEERING
UNIVERSITY *of* VIRGINIA

APPROVAL SHEET

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

Sameer S. Bajikar
Sameer S. Bajikar, Author

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

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

Jason A. Papin
Dr. Jason A. Papin
Committee Chair
Department of Biomedical Engineering

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

Jennifer M. Munson
Dr. Jennifer M. Munson
Committee Member
Department of Biomedical Engineering

David Wotton
Dr. David Wotton
Committee Member
Department of Biochemistry and Molecular Genetics

Accepted for the School of Engineering and Applied Science:

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

Abstract

Many biological measurements assume all the cells in a population are identical. However, no two cells are identical. How cell variation, or heterogeneity, influences biological outcomes is unclear. Cellular heterogeneity has been correlated with poorer prognoses in cancer, suggesting that heterogeneity plays a role in the progression of the disease. Understanding the role of heterogeneity is limited by the ability to identify which genes and proteins are different in the tumor population.

In this Dissertation, we build and extend computational tools to identify and quantify heterogeneously regulated genes. We apply these approaches to understand the role of heterogeneity in basal-like breast cancer, the most lethal form of breast cancer. We begin by designing a computational algorithm to quantify geometric properties of three-dimensional breast cancer spheroid in vitro assays. We use the quantitative characteristics to identify inter-cell line and intra-cell line differences in spheroid morphology. Furthermore, we use the approach to quantify the extent of heterogeneity present in basal-like breast cancer cell lines. Next, we build molecular and agent-based models to explain the heterogeneity we observe in a biomarker protein, JUND. These models helped us to understand what the triggers and consequences of JUND heterogeneity were in breast epithelial spheroids. Last, we build a statistical model of sampling to obtain single-cell level gene expression information from population measurements. We use this approach to globally quantify the frequency of cells expressing different transcriptional programs at an elevated level. We identified one gene (*PIK3CD*) that was rarely expressed, but critical to normal spheroid development.

In Part II of this thesis, we identified a heterogeneously expressed ligand, growth-differentiation factor 11 (GDF11). GDF11 has not been implicated in normal or pathologic breast biology. In this Dissertation, we discover that GDF11 improves the morphology of normal breast and breast cancer spheroids. We identify SMAD4 and ID2 as molecular mediators of GDF11 treatment. GDF11 blocks the ability of basal-like breast cancer cells to establish tumors in a xenograft model, suggesting the loss of GDF11 function could play a role in the progression of human basal-like breast cancer. Strikingly, we observe that human basal-like breast cancers have a defect in secretion of GDF11. This intracellular sequestration of a tumor suppressive ligand presents a new way in which cancer cells inactivate tumor suppressors.

Acknowledgements

Performing impactful science requires a team effort. Throughout this Dissertation, I refer to results “we” gathered. Truly, every experiment performed required help from someone else in some way. I am incredibly fortunate to have worked in such a fantastic environment.

I first want to thank my advisor, Dr. Kevin Janes. Your continued support and enthusiasm has been a source of inspiration. You have continually challenged me, and I am a much better scientist for it. I will always appreciate that you are deep in the trenches alongside the students; some of the best learning moments for me have come through these situations. To list all the lessons learned, both in life and science, would yield a document longer than this thesis. I hope to continue use these lessons to be a better mentor and scientist myself.

The encouragement of everyone at the University of Virginia makes the university a special place. My thesis committee has been an invaluable resource for my scientific work and personal development. I want to thank Drs. Jason Papin, Amy Bouton, Jenny Munson, and David Wotton for their continued support. Additionally, I would like to thank Dr. Hui Zong for allowing me to attend his lab meetings. I have thoroughly enjoyed learning about your science and the considerations of studying other types of cancer. I also want to thank the Women’s Oncology group, Particularly Dr. Jill Slack-Davis, and Systems Biology faculty, particular Drs. Jeff Saucerman and Shayn Peirce-Cottler, for scientific and career feedback over the years. Dr. Pat Pramoonjago and BTRF were critical in performing immunohistochemistry. I hope to be friends and colleagues with everyone in the future.

Collaboration is necessity in science research. I would like to thank Drs. Christiane Fuchs, Andreas Roller, and Fabian Theis for their collaboration on Chapter 4. I would also like to thank Dr. Kristen Atkins for her collaboration and support on Chapters 3 and 5. In these collaborations, the skills brought to the table helped make those Chapters scientifically strong.

The lab culture and environment has been so conducive to working at a high level. I want to thank Cheryl Borgman for keeping everything moving – there is no way I could do the work needed without your organization and help. In the same light, I want to thank Dr. Lixin Wang for her help and advice across multiple projects. I thank Dr. Chun-Chao Wang for being a great collaborator on Chapters 3 and 5, and a good sounding board for ideas. Dr. Karin Jensen was a fantastic role model to show what was required to be an outstanding graduate student. Zeinab Chitforoush and Millie Shah have been great friends throughout the Dissertation work. I wish Christian Smolko, Liz Pereira, and Sham Singh all the best with their thesis work. I want to especially acknowledge the work of Michael Borten in Chapters 2 and 5. You have been a pleasure to mentor, and I am proud of the progress you have made. I have no doubt you will succeed in your future endeavors.

The Dissertation is a long road. I am grateful for the lifelong friends I have made during graduate school. Through the efforts of friends like Scott Seaman, Edik Blais, Eric Greenwald, and Phil Yen, I maintained a healthy work-life balance. From happy hours to conferences to weddings, we’ve had a busy ride. I also want to thank the University of Virginia Men’s Basketball team for being an elite team during my graduate work. There are few better things than watching the ‘Hoos win (especially with free sideline student tickets)!

Beyond the science, the support I have in my personal life remains central to everything I do. I must thank my parents, Anagha and Subhash, for their love and support since day one. I am lucky that you guys are just a short drive away, and I know I can rely on you for anything I need. To my brother Vikaas, you were just a kid when I started and now you are a first year at

UVa! Its been a pleasure to watch you grow up and has been awesome to spend this year with you in Charlottesville. I'd like to thank family I acquired during graduate school, Ruth and Skip. Your support has been just as strong and as valuable as anyone.

My own family has grown during graduate school. To my daughter Parker, I *sincerely* want to thank you for not following the stereotypes of babies while I was writing this Dissertation. I promise to pay it forward to you in the future. The biggest lesson I want to convey to you is that there are always going to be external factors outside your control that will influence the results you desire (sometimes good, sometimes bad). The only thing you are truly in control of is your effort. Even through difficult times, make sure you always give your best effort in every little thing you do. Success will follow.

Finally, the person I want to thank the most is my do-it-all wife, Mary Beth. We've experienced so many things in six years, and yet, our adventure has only begun. Your love and support has given me the strength and drive to pursue and achieve my goals. There is no way I can communicate in writing how much you do for me, so I will simply say that **I love you**.

Sameer S. Bajikar
May 2016

My thesis work has been supported by the Graduate Research Fellowship (National Science Foundation) and Commonwealth and Sture Fellowships (University of Virginia). This work has been funded by the American Cancer Society, National Institutes of Health, and the University of Virginia.

Table of Figures

Table of Tables

1 Chapter 1 – Introduction and Background

1.1 Heterogeneity is intrinsic to biology

No two cells are identical (1). Despite having identical genetic composition and environmental cues, cells can be very different in their behaviors. Some of these cell differences are due to stochastic fluctuations in gene expression (2-4); however, these differences also occur due to heterogeneity in gene regulation (5, 6), resulting in distinct populations of cells with different behaviors (7). These differences in cell behaviors, or phenotypes, drive a number of biological processes. In this Chapter, we discuss the role of heterogeneity in normal and disease states, and the challenges in studying heterogeneity.

1.1.1 Heterogeneity in normal biology

Cellular heterogeneity occurs naturally in many normal contexts, but features prominently in development (8-11). In the simplest case, all multicellular organisms begin as single-cells that proliferate, diverge, and differentiate in phenotype. As all the cells share the same genetic background, the causes or cues for differences in phenotype, particularly early in development, are not readily apparent (12). The development and differentiation process has been extensively studied in stem cells. In a variety of studies, stem cells have been found to be highly heterogeneous at the mRNA (10, 13-15) and protein level (16, 17). These cell-to-cell differences are important for the organism to develop correctly (11). The particular emphasis placed on transcription factor abundance and activity suggests that heterogeneity in gene regulation is critical (10, 11, 16).

Heterogeneity is also present in adult tissue, exemplified by the immune system (18). Across mRNA (19) and protein (20) levels, immune cells are highly heterogeneous even within the same immune cell type. These cell-to-cell differences have a profound impact on how immune cells respond to stimulation (21) and drugs (20). Cell-to-cell differences, specifically in gene expression, have also been mapped to differences in cellular behavior in other lineages, further implicating the importance of heterogeneity in gene regulation to functional biological consequences (22, 23). The principles of cellular heterogeneity also occur in the context of disease, most notably in cancer.

1.1.2 Heterogeneity in cancer

Heterogeneity in cancer is studied at two different scales (24). The highest scale examines the differences in tumors that occur between patients, or intertumor heterogeneity. Patient-to-patient variability within the same cancer site is predominantly due to the differences in genetic perturbations in the tumors. Intertumor heterogeneity has been extensively catalogued through large sequencing and profiling efforts, like The Cancer Genome Atlas (25-31). Through these efforts, for example, we now know what the recurrent, both dominant and rare, mutations are for many cancer types. Intertumor heterogeneity is valuable for stratifying patients to identify approproate therapeutic options (32); however, we are quickly reaching saturation of information given the number of tumors that have already been sequenced. This suggests that additional information could be gleaned looking at a different biological scale.

Intratumor heterogeneity, the second scale of heterogeneity, describes the differences between the individual cells within the same tumor (33). Like intertumor heterogeneity, intratumor heterogeneity can also be genetically based. The genomics of intratumor heterogeneity can be studied at the macroscopic level through deep sequencing (34), regionally by sampling different parts of the tumor (35, 36), and at the single-cell level (37, 38). The genetic differences can catalogue the number of subpopulations (34, 39-41) in a tumor and help piece together a phylogenetic tree of tumor evolution to suggest which clones give rise to metastases or drug resistance (37, 42). Beyond a catalogue or tumor evolution, however, associating genomic changes to cancer phenotypes remains a limitation, given that the same mutation can manifest differently (43, 44). Cancer genomics provides a snapshot into what mutations are driving the uncontrolled proliferation, but not necessarily a snapshot into the tumor behavior.

Intratumor heterogeneity also occurs by non-genetic means. Differences in the tumor stroma can dictate tumor behavior. For example, increased mechanical stiffness drives malignant behavior in breast epithelial cells (45). Stromal cell interactions can also drive tumor behavior in different directions (46-48). Beyond host or microenvironment interactions, tumor cells can be transcriptionally distinct from one another (49, 50). These differences in transcriptional regulation can create subpopulations of cancer cells with different phenotypes (51-54). These subpopulations are dynamic and difficult to capture through conventional approaches. Regulatory heterogeneity remains an understudied aspect of cancer biology (see Section 1.4).

Tumor heterogeneity has been recognized for decades (55), yet heterogeneity continues to be a confounding and complicating factor in cancer research. In this Dissertation we put forth evidence that transcriptional regulatory heterogeneities are a major driver of cancer cell and overall tumor behavior. The evidence presented focuses on a cancer that is an exemplar of intratumor heterogeneity, basal-like breast cancer.

1.2 Basal-like breast cancer is a heterogeneous, aggressive disease

Many types of cancers have been catalogued and categorized through molecular profiling to identify subtypes of tumors and understand the extent of intertumor heterogeneity. These analyses have been applied to breast cancer, one of the most common and lethal malignancies in women. These analyses uncovered four predominant subtypes: luminal A, luminal B, human epidermal growth factor receptor 2 (HER2) amplified, and basal-like (56, 57). The "basal-like" moniker comes from the first set of molecular profiles, which identified a number of basal cytokeratins uniquely expressed in this group of tumors (56-60). Further molecular characterization has stratified basal-like into basal-like (Basal A) and Claudin-low (Basal B), with claudin-low tumors having mesenchymal-like and stem-cell characteristics (61). Of the four major subtypes, basal-like breast cancer has the poorest prognosis. Basal-like breast cancers are also more susceptible to relapse and metastases (60, 62). Basal-like breast cancer is a rare subtype, occurring in 10% of all breast cancer diagnoses, yet account for a disproportionate number of breast cancer deaths (63).

One contributing factor to the poor prognosis is that basal-like breast cancers are typically "triple-negative" for the therapeutically relevant receptors in breast cancer (57). "Triple-negative" status refers to no expression of the estrogen receptor (ER), progesterone receptor (PR), and no genomic amplification of HER2 (27, 41). These receptors and receptor pathways have specific inhibitors, used as the standard of care, if these receptors are expressed in the tumor (64). However, none of these options are viable for basal-like breast cancer. While the basal-like subtype has been known and studied for over a decade (57), no specific or targeted therapies have been uncovered (65).

The predominant genetic lesions that occur in basal-like breast cancer are loss of *TP53* and *BRCA1 (27)*. Both of these genes are important in the DNA damage response and critical in maintaining genomic stability (66). Due to the loss of these two genes, basal-like breast cancers do exhibit genomic instability (67, 68). However, the measured changes in genomic copy numbers may be a consequence rather than the cause of basal-like breast cancer lethality. For example, drugs targeting the DNA damage pathway (PARP inhibitors) had no significant effect on treatment of human basal-like breast cancer (69). Additionally, recent evidence demonstrates that mortality risk actually decreases with increased instability (34). Taken together, these data suggest that genomics alone cannot explain the aggressiveness of basal-like breast cancer.

Intratumor heterogeneity compounds the difficulty in understanding and treating basal-like breast cancer. Intratumor heterogeneity is widespread in basal-like breast cancer tumors. Many tumors show heterogeneous expression of basal cytokeratins (59) (Figure 1.1). Strikingly, studies have shown that tumor cell heterogeneity is highly dynamic, where cells are transitioning between different molecular states (52, 53). The heterogeneity we observe by immunohistochemistry or by other techniques simply shows a snapshot of the state the tumor cells were in at the time of fixation. The time-scale that tumor cells can transition between states strongly suggests that heterogeneity in gene regulation is critical to the final tumor behavior (70). This is supported by the observation that intratumor heterogeneity correlates with a poorer prognosis (71), suggesting heterogeneity contributes to the progression of the disease.

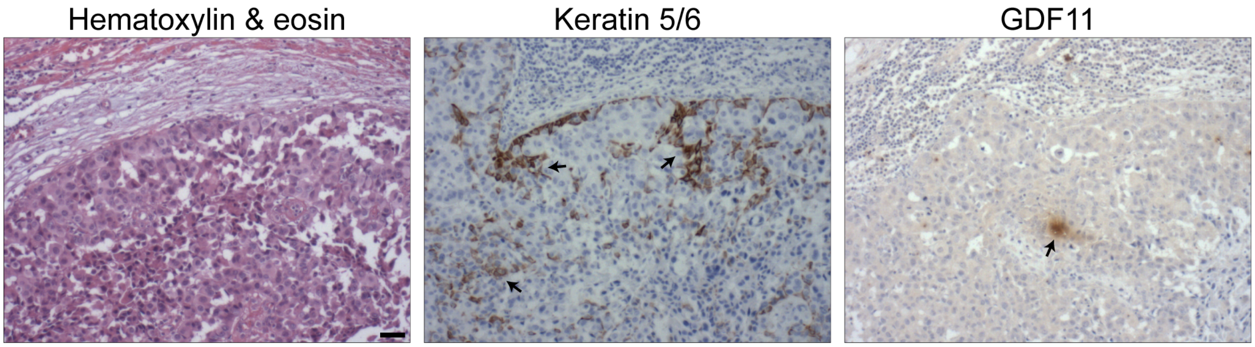


Figure 1.1 Heterogeneous expression of biomarkers in basal-like breast cancer. Serial sections from a clinical specimen of basal like breast cancer show heterogeneous expression of two different biomarkers: keratin 5/6 (middle; see Chapter 3) and GDF11 (right; see Chapter 5) despite the cells within the tumor having similar morphology (hematoxylin and eosin stain; left). Arrows highlight heterogeneous immunoreactivity for the respective biomarkers. Scale bar is 80 μm.

1.3 Challenges in studying heterogeneity

One key limitation in understanding the role of regulatory heterogeneities in cancer is in globally identifying which genes and proteins in the tumor are heterogeneous (72, 73). Many techniques rely on measuring protein or mRNA content from large numbers of cells. These techniques give one metric (e.g., expression of a gene or protein) that represents the behavior of the average cell in the population (Figure 1.2). However, whether any cells exhibit this average behavior is unclear (74). The most robust way to describe and quantify heterogeneity is to make measurements on every single cell in the population; however, single-cell techniques present difficult experimental considerations.

Profiling the proteome on the single-cell level is possible with labeling approaches. Flow cytometry based approaches can use antibody labeling to quantify percentages of cells that express a marker (75). Fluorescence based flow cytometry is limited by the spectral overlap of dyes, giving a finite number of markers the assay can probe (76). Modern mass spectrometry based cytometry, "mass cytometry", alleviates this limitation by replacing the dyes with heavy ions that are readily separated by mass spectrometry (20). However, mass cytometry is limited by the number of antigen-specific antibodies currently available with direct heavy ion conjugation. These limitations allow an experimentalist to segregate a population of cells based on a set of preselected marker proteins. Identification of new subpopulations that do not have a known combination of biomarkers is impossible with these approaches. Additionally, these methods typically require the tissue to be dissociated into single cells, removing the cells from their in situ context (77). Single-cell Western blotting is possible, but faces the same, if not more, limitations as cytometry (78). Immunofluorescence can provide important insight into the localization and abundance of a protein of interest in a single cell, but demonstrates the same

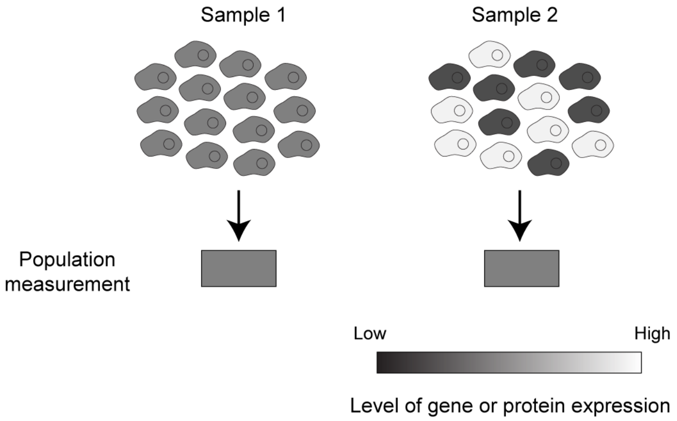


Figure 1.2 Population measurements blur cellular heterogeneity.
Population level measurements like mass spectrometry and RNA sequencing give a measurement of the abundance of a given protein or transcript. The bulk measurement averages out the differences between single cells, and the assumption is the population measurement is reflective of the average cell. The three samples displayed above have very different single-cell expression distributions; yet, a population level measurement would determine that the abundance of the hypothetical gene or protein is the same.

limitations as cytometry (79). Therefore, the proteome is unable to be globally profiled in single mammalian cells.

Profiling the transcriptome at the single-cell level is also possible with labeling and quantitative PCR (qPCR) approaches. RNA fluorescence in situ hybridization (RNA FISH) is a technique that fluorescently labels individual transcripts within a fixed cell (80, 81). Advances in FISH have allowed for the profiling of ~1000 transcripts per single-cell (82-84). While closer to a global measurement, only ~5% of the transcriptome can be covered in one experiment. Similarly, only several genes can be probed at a given time with single-cell qPCR (10, 85, 86). These transcriptomic approaches can measure gene expression changes for known markers, but the discover of new heterogeneities or new subpopulations is not possible.

The ideal method that would catalogue and categorize cellular transcriptional heterogeneity would be a global profile of the transcriptome in a single cell. Many approaches have been devised and tested (13, 19, 49, 87-93), and indeed, global profiles of single-cell gene expression can be constructed (94). However, there is a significant limitation in single-cell techniques to separate true biological variation from technical variation. Despite several iterations of improved sensitivity, the prevailing conclusion is that the starting material provided by a single cell is insufficient for quantitative transcriptomics (89). The difficulty in RNA purification and reverse transcription cDNA synthesis of limited biological material precludes quantitative assessment of the original single-cell transcriptome (72, 95). For example, our lab has taken RNA from a pool of cells and serially diluted the starting material down to the amount expected from a single cell. Given that the RNA source is from a pool of cells, we would expect any technical replicate to yield the same measurement for a given transcript. Yet, when we measure gene expression by qPCR, we observe a loss of quantitative accuracy and unreliable

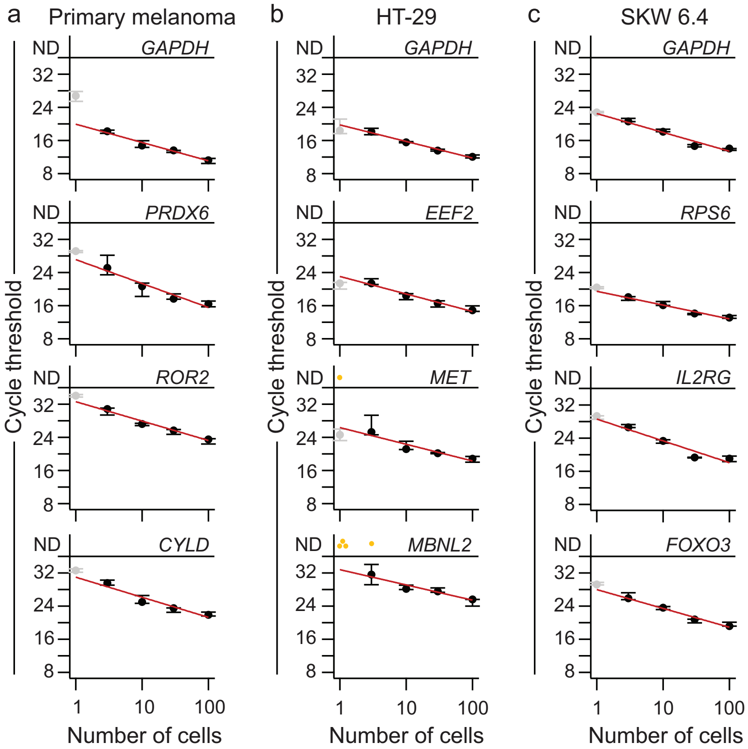


Figure 1.3 Noise in gene expression profiling increases with single-cell level input material. (a-c) 100-cell samples were serially diluted and amplified by poly(A) PCR under optimal conditions for (a) microdissected primary melanoma cells, (b) microdissected HT-29 colon adenocarcinoma, and (c) SKW 6.4 lymphoblastoid suspension cells. High- and low-abundance genes were monitored by qPCR, and data are shown as the median ± range of three replicate small-sample amplifications. Red lines show the log-linear fit of the 3–100-cell dilutions. Note that the one-cell amplifications (gray) often deviate from the log-linear fit or are frequently not detectable (yellow, ND). Reprinted from (72) with permission.

detection of genes when using a single-cell amount of material (Figure 1.3). Single-cell transcriptomics can be useful for inferring qualitative differences between cells, but given the technical variation, single-cell techniques are not suited to uncover regulatory heterogeneities.

1.4 Stochastic profiling as an approach to study heterogeneity

How can we then globally identify regulatory heterogeneities in an unbiased manner? To strike a compromise between population-level and single-cell level measurements, we developed a technique called stochastic profiling (73). In stochastic profiling, we randomly sample a population by collecting a small number (e.g., ten) of cells and measuring the gene expression of the pooled small-cell sample (72). By profiling more than a single cell, we gain additional starting material that enables quantitative reproducibility (Figure 1.3) (72, 89). Additionally, by limiting our samples to few cells (typically ten cells), we avoid averaging out any heterogeneously regulated transcripts. We then repeat the process a number of times to collect transcriptional profiles from several random collections of cells in the population. We can then look at the fluctuations between the gene expression measurements to predict which genes are homogenously or heterogeneously regulated.

We first assume that heterogeneously regulated genes are expressed at a basal, low level, and at a higher level in a subpopulation of the cells (Figure 1.4A and Chapter 4) (72, 73, 96). Genes that are homogenously regulated will exhibit minimal fluctuations between the small-cell samples; additionally, these fluctuations will be approximately normally distributed (Figure 1.4B, C). Conversely, heterogeneously regulated genes will exhibit substantial fluctuations from sample to sample. The variance between samples is attributed to selecting fewer or more cells from the population of cells expressing the given gene at a higher level (Figure 1.4B; see Chapter 4). We can statistically compare the distribution of measurements for both the homogenous and heterogeneous genes against a normal distribution (formally, with the Kolmogorov-Smirnov test). Genes who have a measurement distribution significantly different than a normal distribution are flagged as candidate heterogeneously regulated genes (Figure 1.4B, C).

Stochastic profiling offers many advantages to other contemporary transcriptomic techniques (see Section 1.3). First, we avoid the technical noise of single-cell techniques by using more cells without losing the ability to detect cell-to-cell differences (73). Second, the technique has been optimized for several different biological contexts: cells in suspension, cells in culture, and tissue (Figure 1.3) (72). The latter context is particularly important, because other population-level and single-cell techniques require tissue dissociation to collect the biological material (77). Stochastic profiling is capable of working with laser capture microdissected tissue, allowing the preservation of important in situ contexts for the cells of interest (73). More recently, the technique has been expanded to enable fluorescence guided laser capture microdissection, allowing the user to use genetically encoded fluorophores or fluorescent antibodies to specifically profile cells of interest.

Stochastic profiling was initially validated on an in vitro model of heterogeneity, three-dimensional (3D) breast epithelial morphogenesis. In this model, single normal breast epithelial cells are seeded on top of reconstituted basement membrane (97, 98). Over the course of two

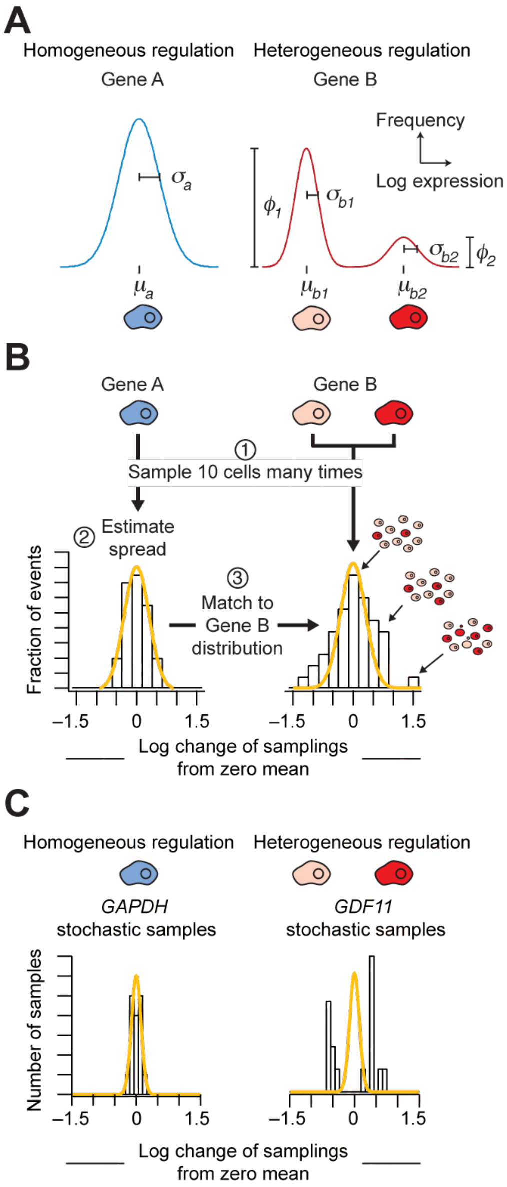


Figure 1.4 Stochastic profiling identifies regulatory heterogeneities in an unbiased manner.
(A) Schematic of homogenous and heterogeneous gene regulation. Genes that are homogenously regulated (Gene A) exhibit little cell-to-cell variability. Any differences in gene expression are due to noise. Conversely, heterogeneously regulated genes (Gene B) are expressed at a low, bsasl level and at a high level by a subpopulation of the cells (see Chapter 4). (B) Stochastic profiling samples a given population of cells and measures gene expression on this small pooled sample of cells. Genes that are homogenously regulated (Gene A) exhibit low variance in their small-cell pooled samples. The distribution of measurements is not statistically different from a normal distribution. Genes that are heterogeneously regulated (Gene B) exhibit high variance in their small-cell pooled samples. The distribution of measurements is statistically different from a normal distribution. (C) Example measurements of a homogenously regulated gene (*GAPDH*, left) and a heterogeneously regulated gene (*GDF11*, right, see Chapter 5) from (73).

weeks, the single breast cells proliferate, organize, and polarize into a spheroid. By two weeks, the cells attached to the matrix will proliferation arrest. Cells within the spheroid, not in contact with the extracellular matrix, undergo apoptosis and the spheroid hollows. During the development of the spheroid, the cells attached the matrix are highly heterogeneous in their cell state (see Chapters 3-5) (53, 99, 100).

Previous work in the lab used laser capture microdissection to randomly select ten cells from the matrix-attached population and profile the transcriptome in those samples (73). The analysis suggested that ~700 genes were heterogeneously regulated. Using RNA FISH, we validated the heterogeneous regulation of a number of transcripts expressed amongst the matrix-attached cells (73, 96). Importantly, transcripts that had similar fluctuations in their 10-cell samples were highly correlated at the single cell level, suggesting that stochastic profiling identifies heterogeneously co-regulated gene programs (73). The functions of these heterogeneities require experimental follow up to understand their role in breast epithelial morphogenesis (see Part II).

1.5 Goals of dissertation

Heterogeneity confounds many aspects of biological research. By acknowledging heterogeneity, we can build tools, like stochastic profiling, to reduce the complexity that heterogeneity adds (7, 73, 83, 96, 101). These tools can also uncover biology that would otherwise be masked by conventional techniques (53). Consequently, the new biology that is discovered must be experimentally explored and validated. The goals of this dissertation are to build computational tools to understand and quantify heterogeneity (Part I: Chapters 2-4) and to use stochastic profiling to uncover interesting biology in basal-like breast epithelia with potential clinical applications (Chapter 3 and Part II: Chapter 5). The central hypothesis to the Dissertation is that single-cell regulatory heterogeneities are not only correlative to poor prognosis, but play a functional role in the progression of basal-like breast cancers. Identifying, describing, and understanding the role of transcriptional regulatory heterogeneities provides important and translational insight into basal-like breast cancer biology.

Part I: Computational analyses of heterogeneity

In this Part of the Dissertation, we discuss several computational approaches that we have developed to understand the biological role of heterogeneity. Each Chapter within this Part builds a computational tool that helped us to simply the heterogeneity at different scales: patient-to-patient variability (Chapter 2), multicellular heterogeneity (Chapters 2 and 3), intercellular heterogeneity (Chapter 3), and molecular heterogeneity (Chapter 4).

2 Chapter 2 – Digital morphometry quantifies heterogeneity in three-dimensional phenotypes

2.1 Foreword

Breast epithelial cells form three-dimensional (3D) spheroids when grown on top basement membrane (97, 98). This in vitro assay can be used to study many different aspects of epithelial biology. In this Dissertation, we use 3D culture as a model of non-genetic heterogeneity (Chapters 3-5). The cells within the 3D spheroid are highly heterogeneous in their cell state (53, 73, 96). We perturb genes identified by stochastic profiling and observe changes in 3D phenotype to indicate biological function of the predict heterogeneity (Chapters 3-5). 3D culture is capable of generating a diverse set of phenotypes from the same experimental set up, providing the ability to observe subtle phenotypic differences (102, 103). Quantifying these phenotypes is time consuming and subject to bias. In this Chapter, we present a computational approach to use image segmentation techniques to extract morphometric properties of individual spheroids. We use this tool to show we can quantitatively probe differences in spheroid phenotypes from different basal-like breast cancer cell lines. The proof of concept shows that these morphometric properties provide a signature for the geometric state of a spheroid. In the future, we can use these data to reduce the time for quantification and allow for higher throughput 3D culture experimentation. This unpublished work was done in collaboration with a Janes Lab undergraduate, Michael Borten, who put significant effort towards in implementing the computational algorithms, data acquisition, and data analysis presented here.

2.2 Introduction

Cells in culture are very different from one another (1, 102). Cell lines derived from the same biological source differ from each other (104, 105). Individual cells within the same cell line also differ from each other (53, 96, 100, 102). How these differences impact cellular phenotypes is unclear (52, 106).

Three-dimensional culture in vitro assays provide an experimentally tractable approach to study both inter- and intra-cell line heterogeneity (97-99, 107). Single cells are seeded on top or within reconstituted basement membrane and proliferate to form three-dimensional (3D) spheroids (98). Heterogeneity of phenotypes between spheroids of different cell lines reflects the heterogeneity between cell lines (102). As each spheroid originates clonally, differences in spheroid phenotype within the same cell line or culture reflect the intra-cell line heterogeneity (53). 3D spheroids are diverse in their phenotypes, so even small differences in cells in 2D culture could propagate to significant differences in 3D culture (102).

Documenting the results of a 3D experiment is nontrivial. A common method is to simply present a representative brightfield image of a spheroid or a collection of spheroids to describe qualitative characteristics of the spheroid(s) (102, 107-110). Another method is to score the spheroids into phenotypic bins by manual inspection of each spheroid individually (53). Both means of documentation are prone to bias from the experimenter. Here, we present a computational approach to quantify spheroid phenotypes. Our method uses image segmentation approaches to quantify morphometric properties of individual spheroids. We applied this approach to basal-like breast cancer cell lines due to the extensive inter- and intra-cellular heterogeneity and clinical relevance of 3D culture (102-104, 111-114). We show that we can segregate 3D spheroid phenotypes from different basal-like breast cell lines. Additionally, using the morphometric signatures, we devise a classifier that quantifies the intra-cell line spheroid phenotype heterogeneity. Digital morphometry provides a means to perform quantitative 3D spheroid culture.

2.3 Results

2.3.1 Design of a segmentation pipeline for digital images of breast epithelial spheroids

In order to characterize the morphometric properties of dozens to thousands of spheroids, we needed to develop an experimental and computational pipeline to acquire digital images of spheroids and convert them to a set of quantitative metrics (Figure 2.1). First, we set up 3D cultures of a panel of basal-like breast cell lines with varying mutational and transcriptional profiles (102, 104, 105). The cultures are established by coating the wells of a chamber slide with reconstituted basement membrane (Matrigel, Corning) and allowing the basement membrane to gel (98). A suspension of single, dissociated cells and dilute basement membrane is overlaid on top of the gel (Figure 2.1; Step 1). The single cells then proliferate to each form multicellular spheroids, which are documented with digital microscopy (Figure 2.1; Step 2). We documented our panel of breast spheroids by serially and exhaustively imaging the culture every three to five days across four biological replicates for twenty days, generating a database of over 1000 images capturing the entire morphogentic process.

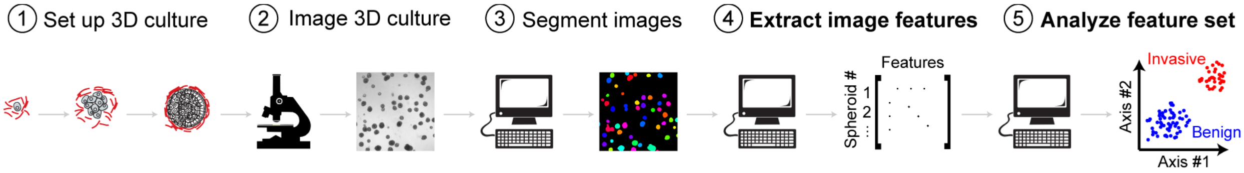


Figure 2.1 Experimental and computational pipeline to generate morphometric profiles of breast epithelial spheroids.

(1) 3D cultures are established by seeding single breast epithelial cells on top of reconstituted basement membrane. (2) Images of each culture are acquired with digital microscopy. Multiple images per well are acquired across multiple biological replicates. (3) Each image is segmented to identify spheroids (Figure 2.2). (4) Segmented spheroids are analyzed for morphometric properties like area and perimeter. (5) The signatures are analyzed to identify groups between and within the cell lines tested (Figure 2.6 and Figure 2.7).

The geometric features (e.g., spheroid area and spheroid perimeter) of each spheroid can be measured manually by tracing the spheroid and creating a region of interest (ROI) (115). The ROI contains information of the all pixels in the image for a single spheroid. For example, the number of pixels can be counted to measure the area of the spheroid. Other additional metrics (see below) are also easily obtained by analyzing the pixels within the ROI. However, the time required to individually trace spheroids, even within a single image, prohibits manual analysis of the number of images in our database.

We sought to leverage work in the areas of computer vision and image segmentation to design a computational algorithm that automatically defines the ROI for each spheroid (116-118). Image segmentation approaches have been successfully applied to fluorescent images of single-cell and whole organisms (117-119). In fluorescence microscopy, the signal is white pixels on a black background, leading to a high signal-to-background contrast. Image segmentation routines typically look for and enhance these contrasts to identify an ROI that encapsulates an object. For brightfield images, the signal is gray-to-black pixels on a gray background, yielding a very low signal-to-background contrast. The low contrast presents a challenge when using image segmentation algorithms.

Edge detection and pixel thresholding are two common approaches for segmenting objects from a background. Edge detection finds the boundaries of objects by looking for the highest contrasts in an image. These contrasts occur when there is a change in pixel intensity, and could represent the transition from an object to the background. After the contrasts are detected, they are connected to form an edge, which defines the ROI of an object. We tried a number of edge detection routines, but the variation in pixel intensities within a spheroid and within the basement membrane layer consistently created false edges. These false edges either prevented reliable detection and segmentation of the spheroids or the false edges created segmented objects where there were no spheroids (Bajikar and Borten, unpublished observations). We excluded edge detection based techniques to analyze our brightfield images.

Pixel thresholding creates a binary image from a grayscale image by setting the pixels of lower intensity than some threshold to background and all other pixels to foreground (signal). The binary image has high contrast between signal and background, resembling a black and white fluorescence image. The threshold value can be empirically determined or calculated based on the pixel values in the image. Minimizing the variance between the foreground and background pixels is a common first value to test as a threshold, an approach known as Otsu's Method. We applied Otsu's Method to our images and observed that some spheroids were segmented from the background, as evaluated by comparing the raw image to the segmented, binary image. However, the binary image consistently failed to capture entire regions of spheroids (Figure 2.2). Additionally, this behavior was insensitive to the exact threshold value (Borten, unpublished observations). Examination of these regions showed that the supposed background pixels had higher intensities than the threshold value. This was due to the non-uniformity in the basement membrane coating. The coating is of varying thickness throughout the chamber and causes the illumination to be uneven. The background pixel intensities are brighter in the areas of higher illumination and these pixels are incorrectly classified as foreground pixels. While some spheroids were correctly segmented, these results indicate that global thresholding approaches are inappropriate for brightfield images of spheroids.

The background pixel intensities were very similar in confined local regions, raising the possibility that pixel thresholding could be applied to subsets of the image. We used a publicly available thresholding routine that performs Otsu's Method on subsets of the image, which we refer to henceforth as “adaptive thresholding”. When we applied adaptive thresholding to our brightfield images, we observed that we now accurately segmented spheroids throughout the whole image (Figure 2.2). The key difference with adaptive thresholding to standard Otsu’s thresholding is the window of the image over which the algorithm is applied. We observed that different window sizes allowed us to capture bigger or smaller spheroids. The window size is a convenient free parameter that allows the algorithm to be used generally for many cell lines. Adaptive threshold accurately segments spheroids defining ROIs we can further analyze (Figure 2.1; Step 3).

Once the spheroids are segmented, we can use the information contained in the segmented ROI to calculate simple geometric measurements like area and perimeter (Figure 2.3). We can calculate additional metrics like Eccentricity, which quantifies the circularity of a segmented ROI. The calculation of metrics is computationally cheap, so we accrued every possible metric we could for each spheroid (Figure 2.1; Step 4). A full list of metrics can be found in Chapter 2 Methods. The set of metrics provide a signature that describes the morphometric state of a given spheroid (digital morphometry).

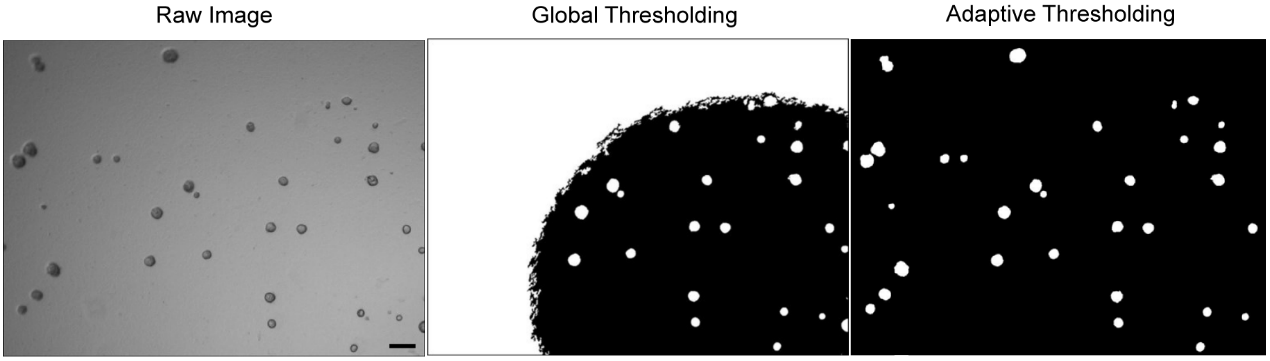
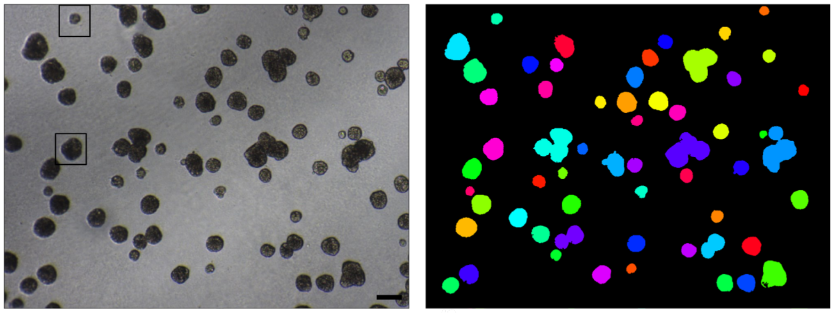


Figure 2.2 Adaptive thresholding accurately segments breast epithelial spheroids from brightfield, digital images.

Images of spheroids are acquired with brightfield, digital microscopy (left). Using traditional thresholding techniques (Otsu's Method), some spheroids are accurately segmented. However, uneven illumination causes entire regions of the image to be classified as foreground (white pixels) incorrectly center). Adaptive thresholding applies Otsu's Method on subsets of the image to avoid the effects of uneven illumination. With adaptive thresholding, spheroids across the entire image are accurately segmented (right). Scale bar is 200 μm.



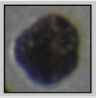
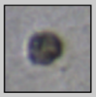
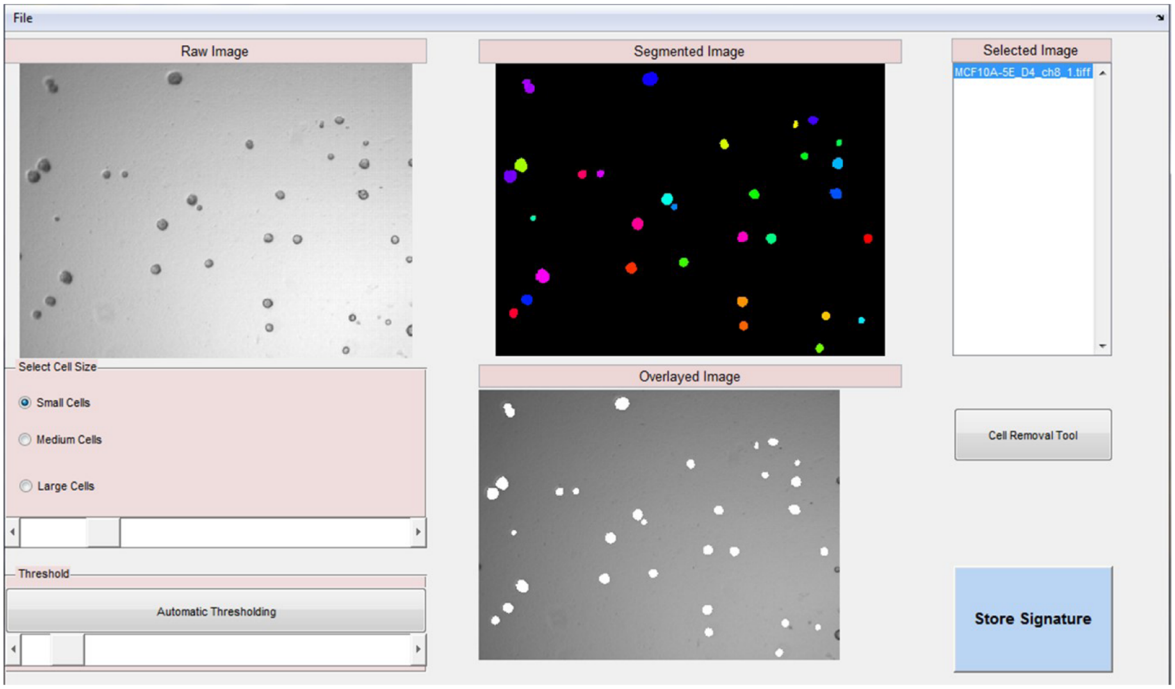
Raw	Segmented	Area (pixels^2)	Perimeter (pixels)	Eccentricity	Diameter (pixels)
		2206	168.647	0.4132	52.998
		655	89.674	0.4074	28.878

Figure 2.3 Morphometric properties obtained from segmented spheroids.
Raw digital images (top left) are segmented with adaptive thresholding (top right). In the segmented image, different colors represent individually identified spheroids. Once the spheroids are identified, we can use the ROI defined by the segmentation algorithm to calculate morphometric properties. Two example spheroids are shown, in the raw image (bottom left) and segmented image (bottom right). Example morphometric properties are shown in the table (bottom right). Scale bar is 200 μm.

2.3.2 Design of a graphical user interface (GUI) for the segmentation algorithm

Adaptive thresholding provides a means to deconstruct our library of images into quantitative signatures we can analyze (Figure 2.3). However, the utility of the algorithm was limited by the necessity of running MATLAB and interfacing with the MATLAB command line. Users that are unfamiliar or do not have access to MATLAB would not be able to use our approach. To generalize the utility for all users, we design a stand-alone graphical user interface (GUI) to house our segmentation routine (Figure 2.4). Users import their images into the GUI with a menu system. The user then defines the window size based on the size of the spheroid. The segmentation routine then runs on all the images imported. If the segmentation is not accurate, the user can adjust the threshold until a suitable segmentation result is achieved. Additionally, if a spheroid cannot be accurately segmented despite adjusting the threshold parameter, the user can remove that spheroid from further analysis. Lastly, the spheroid metrics can be exported to MATLAB or Excel for data processing and analysis. The GUI allows any user to quickly and automatically quantify characteristics of their spheroid cultures across multiple images.



MCF10A-5E_D4_ch8_1.tiff

● Small Cells
○ Medium Cells
○ Large Cells

Figure 2.4 Graphical user interface (GUI) to interactively segment images.
A GUI was created to house the segmentation routine, allowing users unfamiliar with or who do not have access to MATLAB to use the algorithm. Raw images are loaded through the File drop menu. Adatpive threshold window size is selected by either choosing one of three heuristically determined bins (small,, medium, and large) or can be manually adjusted with the top slider bar. The threshold value can be automatically calculated with the "Adaptive Thresholding" button or can be manually adjusted with the slider bar below. The window size and threshold are applied to all images loaded. The segmented image is shown in the upper center panel, and the overlay of the segmented image and raw image is shown in the center bottom. These two images help assess the quality of the segmentation. Parameters can be adjusted to better segment an image if needed. If a spheroid cannot be segmented, the spheroid can be removed from further analysis with the "Cell Removal Tool". Once the segmentation is correct, the morphometric parameters are extracted to MATLAB or Excel with the "Store Signature" tool.

2.3.3 Segmentation quantifies growth kinetics in breast epithelial spheroids

With an established segmentation routine and easy-to-use interface, we first tested if the extracted metrics could describe the growth characteristics of spheroids from different cell lines. We specifically focused on the metric of spheroid area, as we could qualitatively evaluate the fidelity of the results with visual inspection of the raw images. The changes in spheroid area are well characterized for one of our basal-like breast cell lines, MCF10A-5E. The MCF10A-5E cells are immortalized, but not transformed, and are used to model normal breast epithelial biology (97, 98). Additionally, the MCF10A cell line has been extensively used in 3D culture. Previous work has stereotyped the 3D spheroid morphogenesis of the MCF10A cell line (98, 120). Initially, the single cells actively proliferate and establish polarity during the early stages (until about ten days of culture) and subsequently growth arrest and hollow towards the latter stages of morphogenesis (between twelve and twenty days of culture) (99). These features of morphogenesis have been qualitatively described and supported through molecular analysis, but not quantified in a high throughput manner (99, 120).

We successfully segmented ~700 spheroids across five time points and assessed the population-level trajectory of spheroid growth. In the MCF10A-5E cell line, we observed a 1.5-fold increase in size from day four to day twelve in culture. Afterwards, the average spheroid size did not significantly change, as expected. The variance in sizes remained constant over time as well, suggesting that the population growth was stable and that the size variation is set at an early stage of morphogenesis. We then compared this growth trajectory to a Ras transformed variant of the MCF10A cells (MCF10ADCIS.COM) (121) and observed that Ras transformation drastically altered morphogenesis. Despite starting at a near equivalent size, the average MCF10ADCIS.COM spheroid grew nearly five times as large as the

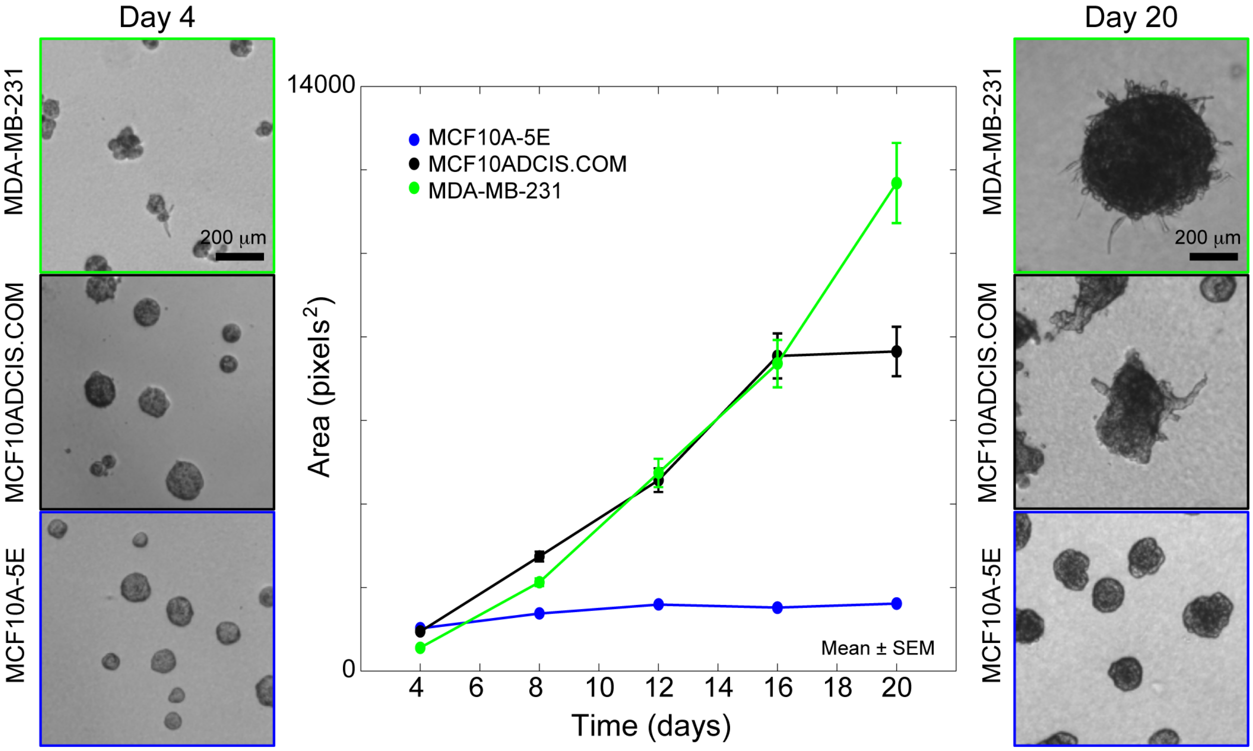


Figure 2.5 Digital morphometry quantifies spheroid growth trajectories.
Images of spheroids from three cell lines (MCF10A-5E, MCF10ADCIS.COM, MDA-MB-231) were grown in 3D for twenty days. Images were acquired every four days, and segmented with adaptive thresholding (Figure 2.2). The spheroid area for each segmented spheroid was accrued over each time point. The plot shows average spheroid area (mean ±SEM) for each cell line over time. Representative images of each cell line at day 4 and day 20 are shown to the left and right of the graph, respectively. Scale bar is 200 μm.

average MCF10A-5E spheroid. Similarly, the basal-like breast cancer cell line MDA-MB-231 formed spheroids that continued to proliferate for the duration of observation. The growth trajectory measured qualitatively matched the size distribution of spheroids in the raw images, validating the segmentation approach. These data indicate that the morphometric measurements captured known growth characteristics and separate cell lines with different growth trajectories.

2.3.4 Digital morphometry segregates spheroid phenotypes

Previous work has categorized a number of breast cancer cell line spheroids in three phenotypic bins: mass, grape-like, and stellate (102, 103). However, this phenotype classification uses immunofluorescent characterization of cell-to-cell adhesions, which requires the additional costs of antibodies, confocal microscopy, and importantly, time (102). To test if morphometric signatures from the brightfield images could segregate spheroid phenotypes, we segmented spheroids and accrued signatures for every cell line in our panel. We examined a single time point of morphogenesis for each cell line where the spheroids were clearly formed but not grown to the point of infringing upon each other. For each cell line, we segmented ~50 to ~800 spheroids at this time point of growth. This time point captures the representative spheroid phenotype of each cell line.

To compare cell line spheroids to one another, we generated a computational "average" spheroid for each cell line by taking the median of each metric. We clustered the average signatures together to identify predominant groups of signatures, suggesting a common phenotypic bin for those cell lines (Figure 2.6A). Clustering revealed three main groups of signatures; examining the raw images of each spheroid, we assigned three phenotypic bins to these groupings: non-invasive, partially invasive, and invasive (Figure 2.6B-D). Spheroid area was a key metric in segregating phenotypes as invasive spheroids were larger than partially or non-invasive spheroids (Figure 2.6A, D). Partially and non-invasive spheroids were distinguished by the Extent and Solidity metrics. Both measurements quantify the overlap of the ROI with a fitted polygon to the area. A non-invasive spheroid will have high Extent and Solidity as the spheroid will fit tightly within a bounding polygon (e.g., a circle fits tightly within a box); conversely, an invasive spheroid will have low Extent and Solidity (e.g., a star does not fit tightly within a box). Additionally, we observed a qualitative correlation between our phenotypic bins and those previously published (102), suggesting that brightfield image morphometrics are sufficient to categorize breast cancer spheroid phenotypes Figure 2.1; Step 5).

Within each spheroid phenotypic bin, there were additional subgroups, suggesting morphometry could identify finer gradations of spheroid phenotypes. For example, amongst the non-invasive cell lines, three cell lines (MDA-MB-436, Hs578T, and HCC1395) had increased values for the Eccentricity metric. Lower values for Eccentricity indicate more circular spheroids. This subgrouping of non-invasive cell lines formed noncircular spheroids distinct from the circular spheroid signature (HCC1937 and MCF10A-5E) (Figure 2.6A, B). Amongst the invasive cell lines, the SUM159PT had a lower value for Extent and Solidity, suggesting these spheroids had more invasive processes (Figure 2.6A, D). These data suggest that digital morphometry reveals subtle differences between spheroids in the same phenotypic class.

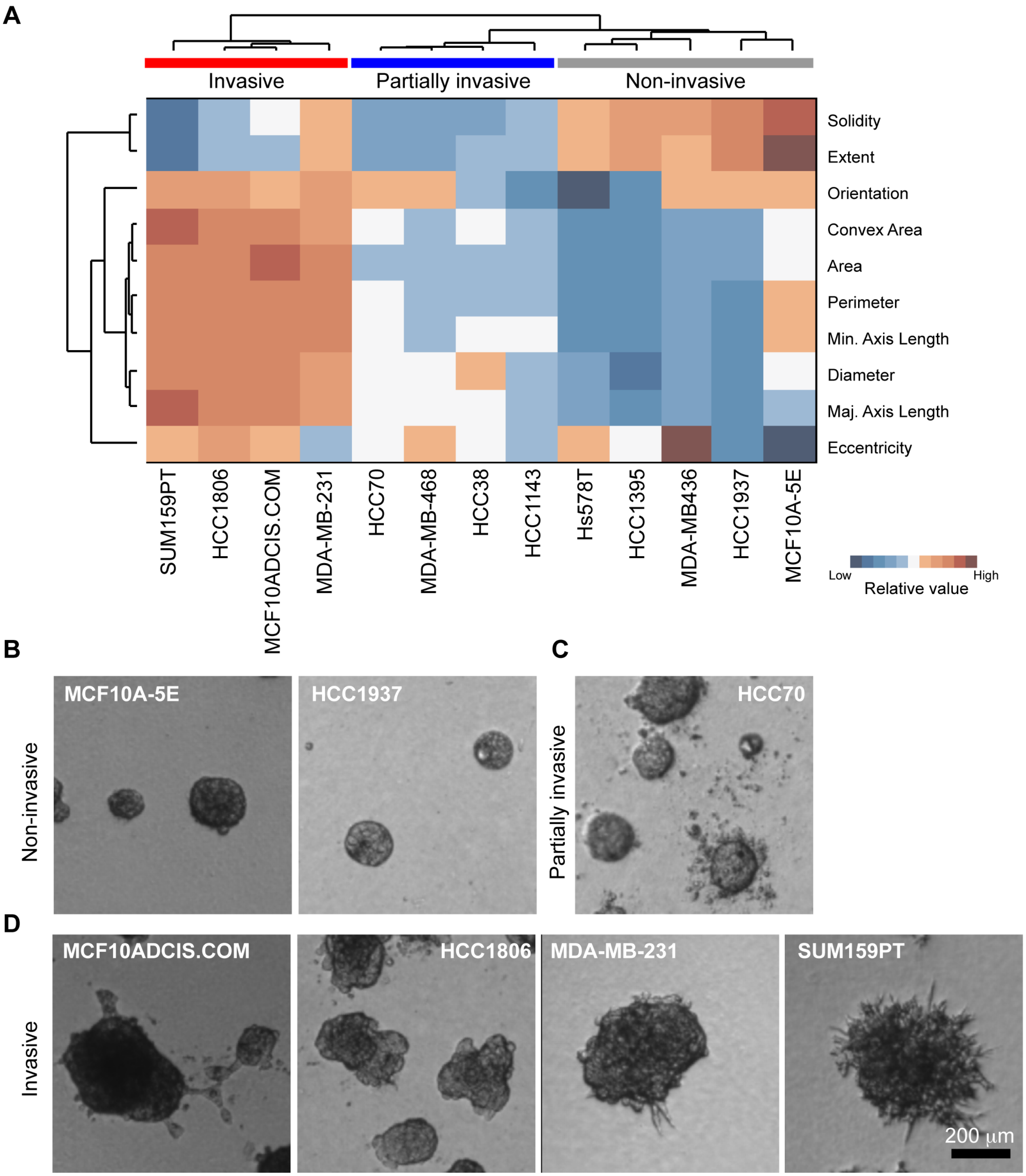


Figure 2.6 Morphometric signatures segregate spheroid phenotypes.
(A) Morphometric signatures were collected by segmenting ~50- ~800 spheroids from thirteen basal-like breast lines (x-asix). An average signature was generated from the median of each metric (y-axis). The average signatures were clustered into three predominant groups with Euclidean distance and ward linkage. (B) Representative images of MCF10A-5E and HCC1937 spheroids. (C) Representative image of HCC70 spheroids. (D) Representative images of MCF10ADCIS.COM, HCC1806, MDA-MB-231, and SUM159PT spheroids. Scale bar is 200 µm.

2.3.5 Digital morphometry quantifies spheroid phenotypic heterogeneity

Phenotype segregation with an aggregated signature categorized the inter-cell line heterogeneity. Could digital morphometry categorize, or ideally quantify, the intra-cell line spheroid heterogeneity? We hypothesized that the combination of metrics could similarly group different spheroid phenotypes within a cell line. Rather than biasedly selecting a subset of the metrics, we used principle component analysis (PCA) to reduce the dimensionality of the dataset to take all metrics into account. Briefly, PCA collapses correlated metrics into a new variable capturing the variance of the data amongst those metrics (122). By collapsing the metrics, we can turn our ten-metric signature into a two-metric signature that we can easily visualize. We chose four cell lines to test the proof of concept: MCF10A-5E (non-invasive), HCC70 (partially invasive), MDA-MB-231 (invasive), and SUM159PT (invasive). Importantly, the MCF10A-5E serves as a control as we observe little intra-spheroid heterogeneity in this cell line (Figure 2.7A).

Individual MCF10A-5E spheroids clustered tightly together, in stark contrast to the three basal-like breast cancer cell lines (Figure 2.7A). The partially invasive and invasive cell lines were also segregated in the principal component space (Figure 2.7B-D). For each breast cancer cell line, a subset of spheroids fell near the cluster of MCF10A-5E spheroids. We sought to quantify the percentage of "MCF10A-5E-like" spheroids as a surrogate for quantifying non-invasive spheroids. With a simple gating of the principle component values near the MCF10A-5E cluster, we created a classifier for non-invasive spheroids (Figure 2.7E-G). As a positive

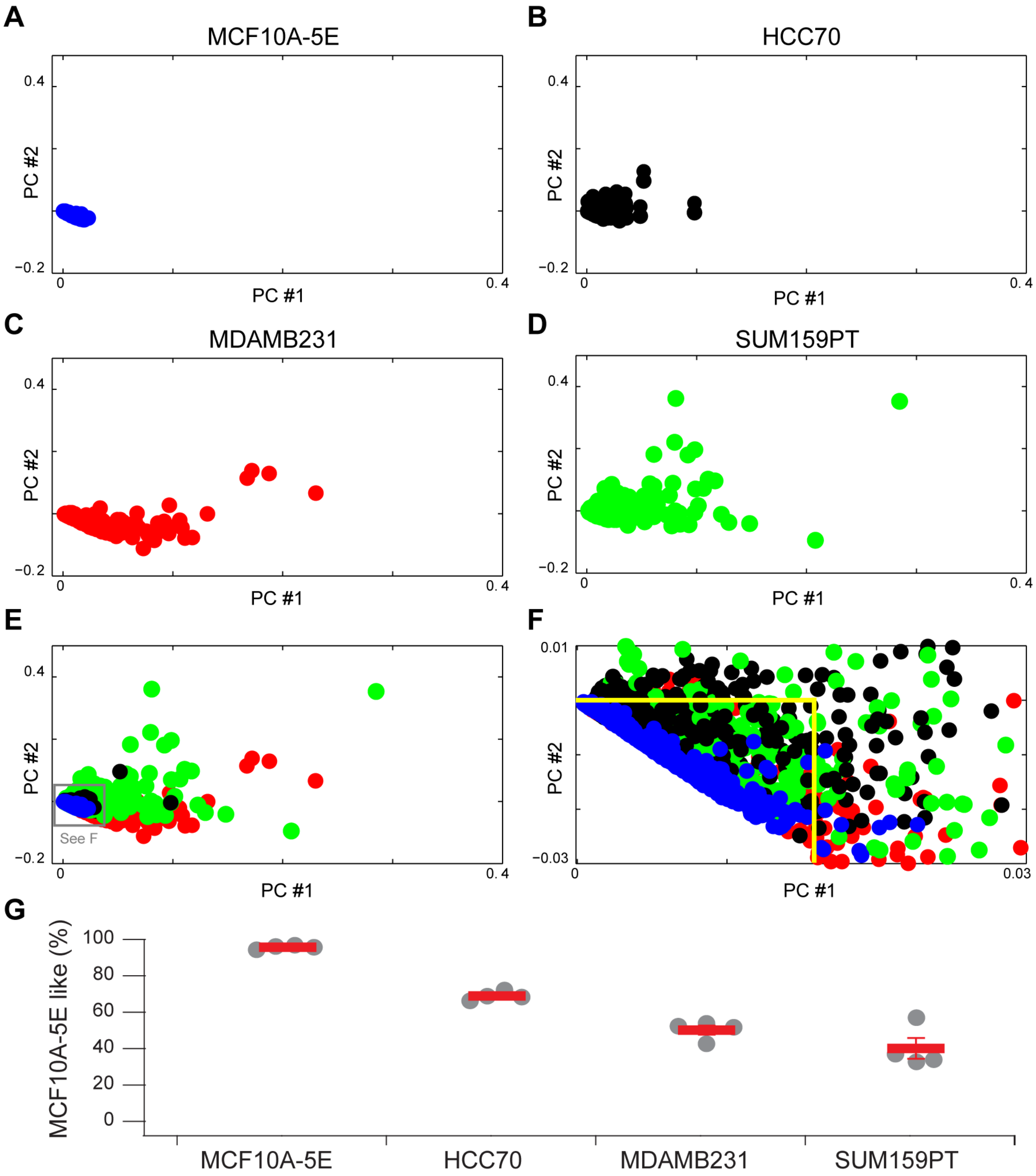


Figure 2.7 Analysis of morphometric signatures quantifies intra-cell line spheroid heterogeneity.

(A-E) Morphometric signatures for four cell lines (MCF10A-5E, HCC70, MDA-MB-231, SUM159PT) were generated for individual spheroids at a single time point. Principle component analysis (PCA) reduced the dimensionality of the 10-metric signature to a 2-dimensional metric. Reduced metrics were plotted along the first two principle components for the MCF10A-5E (A), HCC70 (B), MDA-MB-231 (C), SUM159PT (D), and all together (E). (F) Zoom of gray boxed region in E to highlight relationship between MCF10A-5E spheroids and the other three cell lines. Yellow line shows values of the principle components gated to quantify MCF10A-5E-like spheroids. (G) Quantification of percentage of MCF10A-5E-like spheroids in each cell line, determined through morphometric signatures. Values shown are mean ±SEM.

control to verify the gating, we identified 95% of the MCF10A-5E spheroids as MCF10A-5E-like. In the HCC70 partially invasive cell line, we identified ~70% spheroids with MCF10A-5E-like characteristics, while the invasive cell lines MDA-MB-231 and SUM159PT were ~50% and ~40% MCF10A-5E-like respectively. The MDA-MB-231 quantification qualitatively matches what we observe through manual classification of phenotypes (see Chapter 5, Figure 5.3), demonstrating that intra-cell line heterogeneity is both identified and quantified by digital morphometry.

2.4 Discussion

Digital morphometry provides quantitative metrics of individual spheroid shape. Using these metrics, we quantitatively categorize inter-cell line spheroid heterogeneity and quantify the extent of intra-cell line spheroid heterogeneity. By housing the algorithms in an easy-to-use interface, we have developed a platform for any user to quantify characteristics of their 3D spheroid cultures.

Quantifying inter-cell line spheroid signatures provides a non-molecular comparator for cell line phenotype. For example, the HCC1806 cell line has a Basal A transcriptional profile (123), yet, the HCC1806 cell line spheroids cluster tightly with several cell lines with a Basal B transcriptional profile (104). Reciprocally, there are Basal B cell lines, MDA-MB-436 and Hs578T, that cluster further away from the invasive cell lines (102, 104). This suggests that there may be other molecular markers outside of the basal-like markers that categorize tumor phenotype, or that there are multiple transcriptional states that lead to the same phenotypic outcome.

Using a simple gating strategy of the morphometric signatures, we quantified intra-cell line spheroid heterogeneity. We can build off of this proof of concept by using more sophisticated classifier algorithms like support vector machines (124) or naïve Bayes classifiers (125), which will be able to quantify phenotypes in an unbiased manner (126). Additionally, we can use image tracking techniques to quantify dynamic changes in individual spheroids (127), using the metrics to build predictive models (83, 101). Quantifying and cataloguing spheroid phenotypes is a major limitation in the throughput of spheroid assays. Using automated imaging systems coupled with our segmentation routine, we can significantly increase the scale of experiments. Given the clinical relevance of spheroid assays (107, 128, 129), high throughput drug or RNAi screens may yield higher quality hits than traditional two-dimensional culture screens.

3 Chapter 3 – Modeling intercellular regulatory heterogeneities

3.1 Foreword

In the previous chapter, we discussed a computational approach to analyze heterogeneity between and amongst breast cancer cell line spheroids. In this Chapter, we use computational approaches to better understand the role of intercellular regulatory heterogeneities in normal and malignant breast morphogenesis. Previous work in the lab identified two anticorrelated clusters of genes from the prior stochastic profiling data set (73). One cluster was highlighted by the TGF-β signaling co-receptor, *TGFBR3*, while the other cluster was highlighted by the transcription factor, *JUND*. We went on to show that the regulation of these two genes was important to breast epithelial morphogenesis and that their regulation was intricately connected. In this Chapter, we discuss the construction and analysis of an ordinary differential equation (ODE) model of the purported circuit relating *TGFBR3* and *JUND*. Furthermore, we show that clinical specimens of early triple-negative breast cancer exhibit anticorrelation between TGFBRIII and JUND. Interestingly, the anticorrelation was dependent on the extracellular matrix (ECM) local microenvironment, specifically the ECM protein tenascin C. In this Chapter, we discuss the construction and analysis of an agent-based model describing the role of tenascin C in establishing patterns of biomarker heterogeneity in clinical specimens. This work was done in collaboration with Chun-Chao Wang in the Janes lab. Additionally, this work appears in *Nature Cell Biology* (*Nat Cell* Biol, 16, 345-56), specifically in Figures 3, 7, and the Supplemental Note.

3.2 Introduction

Genetically identical cells often coexist in starkly different molecular states. Stochastic heterogeneities can drive cell fates in specific developmental contexts (130, 131). Within mature tissues, however, cell-autonomous heterogeneity is normally suppressed unless the molecular circuitry has been perturbed (22). Accordingly, cell-to-cell heterogeneity has been extensively described in solid tumors, and heterogeneity within carcinoma cell lines has been associated with drug resistance (132-134).

Heterogeneity cannot be entirely explained by random biological noise—there are substantial contributions from a cell's local environment and its history (2, 23). For most epithelial tissues, it is difficult to track cell-to-cell variability in time and space (135). Organotypic 3D cultures provide an opportunity to monitor heterogeneity by supporting cells in reconstituted basement membrane (98, 136, 137). The more-realistic geometry and ECM context can give rise to non-genetic variations in molecular state (73, 99). For instance, ECM-adhesion receptors comprise nearly all of the stem-progenitor markers for heterogeneity in breast tissue and breast cancer (138, 139). Organotypic heterogeneities might provide insight into clinical mechanisms of tissue-tumor heterogeneity that would otherwise be inaccessible.

Using 3D basement-membrane cultures of basal-like breast epithelia (97, 98), we have uncovered a dynamic heterogeneity that develops among ECM-attached cells during acinar formation. The overall expression circuit is composed of two anti-correlated transcriptional programs that establish a pair of expression states defined by *TGFBR3* and *JUND*. When this circuit is spontaneously excited, ECM-attached cells oscillate transiently and asynchronously between states, creating the static appearance of a cellular mosaic. Single-cell TGFBR3–JUND regulation tracks with heterogeneity of a diagnostic cytokeratin (KRT5) for ductal carcinomas *in situ* with basal-like features (basal-like DCIS). Remarkably, KRT5 correlations reverse upon detachment *in vitro* and in ECM-poor regions of basal-like DCIS. We link the reversal to a keratinization process that is maintained by expression of tenascin C (TNC). The dynamic and ECM-dependent transition of individual tumor cells between expression states may relate to the exceedingly poor prognosis of heterogeneous basal-like breast cancer (59, 71).

3.3 Results

3.3.1 TGFBR3–JUND heterogeneity is critical for normal acinar morphogenesis

We recently described a random-sampling approach that profiles statistical fluctuations to uncover cell-to-cell heterogeneities in gene-expression regulation (72, 73). Applying this "stochastic profiling" technique to a basal-like cell clone cultured in basement membrane, we cataloged 547 transcripts subject to strongly heterogeneous regulation. 17% of the transcripts fell into two clusters that were anticorrelated on a sampling-to-sampling basis. The first cluster included *TGFβ receptor III* (*TGFBR3*, a high-affinity TGFβ receptor (140)), *growth differentiation factor 11* (*GDF11*, a TGFβ-family ligand (141)), and *TGFβ-induced protein* (*TGFBI*, an ECM protein downstream of TGFβ-family signaling (142)). The co-occurrence of a TGFβ receptor, ligand, and marker protein suggested that the first cluster might be linked to TGFBR3-dependent signaling and gene expression.

The triplet of TGFβ-related genes was strongly anticorrelated with the *jun D proto-oncogene* (*JUND*), which was the only transcription factor in the second cluster comprised mostly of protein biosynthetic genes (73). We verified the single-cell anticorrelation by RNA

FISH and further showed that *JUND* and *TGFBR3* were expressed at reciprocal frequencies in ECM-attached cells. *TGFBR3* and *JUND* thus mark two states that basal-like cells spontaneously occupy when in contact with ECM.

TGFBR3 expression is strongly induced during organotypic culture (Figure 3.1a) (143). If *TGFBR3* upregulation occurred sporadically, then it could explain the heterogeneous pattern of expression observed among single ECM-attached cells (Figure 3.1d). To test whether *TGFBR3* induction was important for acinar morphogenesis, we knocked down TGFBR3 and verified specificity with an RNAi-resistant murine Tgfbr3 that is doxycycline (DOX) inducible (Tgfbr3 addback; Figure 3.1b). Inhibiting *TGFBR3* upregulation caused a profound ductal-branching phenotype in ~30% of shTGFBR3 acini (Figure 3.1c,d). Branching returned to baseline when Tgfbr3 was induced at day 4, the time when endogenous *TGFBR3* levels normally begin to rise (Figure 3.1a,c,d). Thus, *TGFBR3* upregulation is specifically important to suppress ductal branching, conceivably by sensitizing cells to TGFβ-family ligands (see Chapter 5) (140).

Unlike *TGFBR3*, *JUND* is easily detected under normal growth conditions and is frequently expressed in ECM-attached cells. To examine the role of sporadic *JUND* downregulation, we constitutively expressed HA-tagged JUND. This perturbation gave rise to stable cellular “bridges” across the acinar lumen, which are cytologically similar to the cribiform subtype of DCIS (144) (Figure 3.1e–g). Heterogeneous *JUND* downregulation remained critical until late in morphogenesis, because induction of HA-JunD at day 9 caused cribiform acini weeks later. To exclude artifacts caused by mild JUND overexpression, we coexpressed a stable shRNA against JUND together with an RNAi-resistant murine JunD that restored near-endogenous levels (Figure 3.1h). This homogenization of *JUND* expression also caused

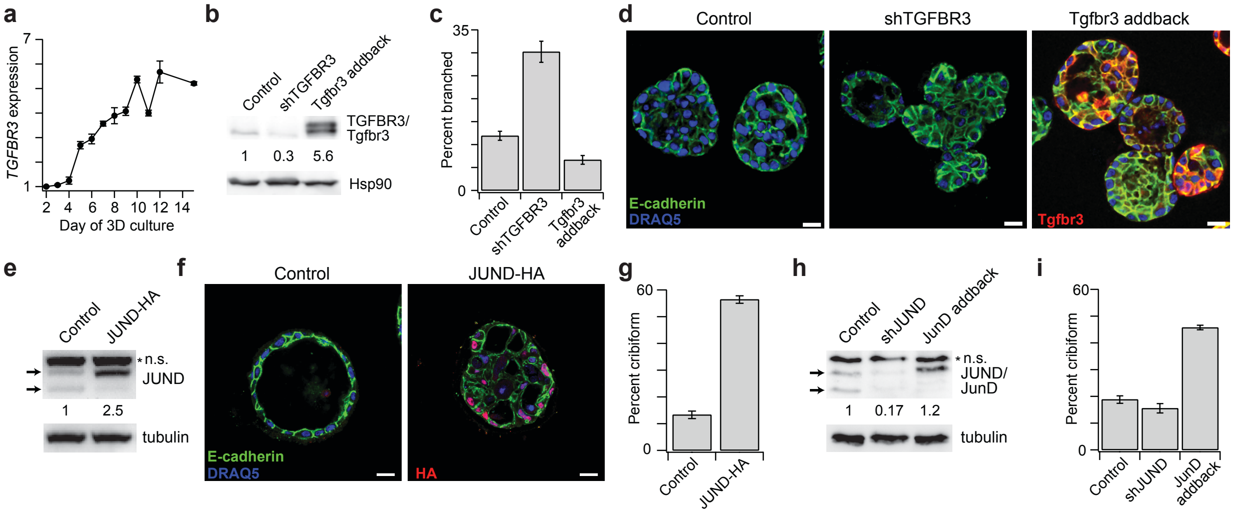


Figure 3.1 TGFBR3 and JUND are functionally important for 3D morphogenesis.

(a) Time-dependent expression of TGFBR3 during 3D morphogenesis(143). (b) Knockdown of TGFBR3 and inducible addback of murine RNAi-resistant Tgfbr3. TGFBR3/Tgfbr3 levels for cells cultured in the absence (Lane 1 and 2) or presence (Lane 3) of 1 µg/ml DOX for 24 hours were analyzed by immunoblotting. Hsp90 was used as a loading control. Densitometry of TGFBR3/Tgfbr3 abundance is shown normalized to the shGFP control. (c and d) Blocking TGFBR3 induction specifically elicits a ductal-branching phenotype. The MCF10A-5E lines described in (b) were placed in morphogenesis in the absence (control and shTGFBR3) or presence (Tgfbr3 addback) of 1 µg/ml DOX from day 4–10. Acini were fixed at day 10 of 3D culture, stained for E-cadherin (green) and HA-tagged Tgfbr3 (red), and analyzed by confocal immunofluorescence. Cells were counterstained with DRAQ5 (blue) to label nuclei. (e) Constitutive expression of HA-tagged JUND analyzed by immunoblotting. Densitometry of JUND abundance is shown normalized to pBabe vector control. (f and g) Constitutive JUND expression causes stable cribiform-like acinar structures. Acini from the MCF10A-5E lines described in (e) were placed in morphogenesis, fixed at day 28, stained for E-cadherin (green) and HA-tagged JUND (red), and analyzed by confocal immunofluorescence. Cells were counterstained with DRAQ5 (blue) to label nuclei. (h) Homogenization of JUND expression by knockdown of JUND and addback with murine RNAi-resistant JunD to near-endogenous expression levels. JUND/JunD levels were determined by immunoblotting. Densitometry of JUND/JunD abundance is shown normalized to the shGFP control. (i) Quantification of the cribiform-like phenotype at day 28 of 3D culture for the cells in (h). For (a), (c), (g), and (i), data are shown as the mean ± s.e.m. of three (a) or four (c, g, i) independent experiments. For (d) and (f), scale bar is 20 µm. For (e) and (h), tubulin was used as a loading control and n.s. denotes a non-specific band.

cribiform acini (Figure 3.1i). Therefore, heterogeneous regulation of *JUND* is critically important for acinar morphogenesis of basal-like cells.

3.3.2 TGFBR3–JUND signaling is oscillatory and dynamically coupled

To determine whether the TGFBR3–JUND clusters were functionally linked, we constitutively expressed TGFBR3 or JUND and then analyzed endogenous mRNA levels of the other cluster (Figure 3.2a–c). Constitutive JUND expression downregulated both *TGFBR3* (P = 0.0026, one-sided t test; Figure 3.2a) and *TGFBI* (P = 0.0027, one-sided t test; Fig. 3b), suggesting that JUND antagonizes expression of the *TGFBR3* cluster. Ectopic *TGFBR3* expression reciprocally inhibited *JUND* expression (P = 0.022, one-sided t test; Figure 3.2c), indicating that *JUND* does not simply act as an upstream repressor of the *TGFBR3* cluster. Mutual TGFBR3–JUND antagonism creates a double-negative (positive) feedback loop, which can establish two distinct molecular states (145).

There were also two other negative autoregulatory feedbacks in the overall wiring. Consistent with earlier reports (146, 147), we found that constitutive JUND expression caused downregulation of endogenous *JUND* (P = 0.043, one-sided t test; Figure 3.2c), and *TGFBR3* expression was acutely downregulated upon signaling from TGFβ-family ligands ($P = 1.4 \times 10^{-5}$, one-sided t test; Figure 3.2d). All together, we arrived at a hybrid signaling-transcriptional circuit comprised of one positive-feedback and two negative-feedback loops (Figure 3.2e).

Regulatory circuits with interlinked positive and negative feedback can oscillate between molecular states (145, 148). We thus developed a live-cell imaging procedure for monitoring *TGFBR3* and *JUND* activities simultaneously. Active TGFβ-family signaling (TGFBR3*) was tracked by RFP1-labeled Smad2 (Figure 3.2e). For *JUND*, we engineered a rapidly responsive fluorescent reporter of endogenous promoter activity (Figure 3.2e). We inserted ~2 kb of the *JUND* promoter (P_{JUND}) upstream of the fast-maturing YFP variant, Venus (149), which was destabilized by N-end rule fusion to ubiquitin C and C-terminal fusion to a PEST sequence (150, 151). Coexpression of ultradestabilized Venus (udsVenus) (P_{JUND}) and RFP1-Smad2 did not substantially perturb acinar morphogenesis relative to control cultures, suggesting that endogenous TGFBR3– JUND pathways were not dramatically affected. For 3D-culture experiments in which stable time-lapse imaging was successful, we repeatedly observed at-least one ECM-attached cell with coupled dual-reporter dynamics.

To compile two-color reporter activities across multiple experiments, we combined spectral filtering with algorithms from multiple-sequence alignment (see Chapter 3 Methods). The aggregate alignment revealed that both reporters exhibited transient peaks of activity separated by 5–10 hr (Figure 3.2f,g). When an ECM-attached cell remained in the optical plane long enough to observe two peaks, the second peak usually had smaller amplitude than the first, suggesting pathway damping (Figure 3.2f, upper rows; Figure 3.2g, middle rows). Strikingly, when the two reporters were compared within the same cell, dynamics were antiphase at nearly all time points (152). Asynchronous, antiphase dynamics within the TGFBR3–JUND circuit provide a mechanism for the static anticorrelation observed in fixed specimens.

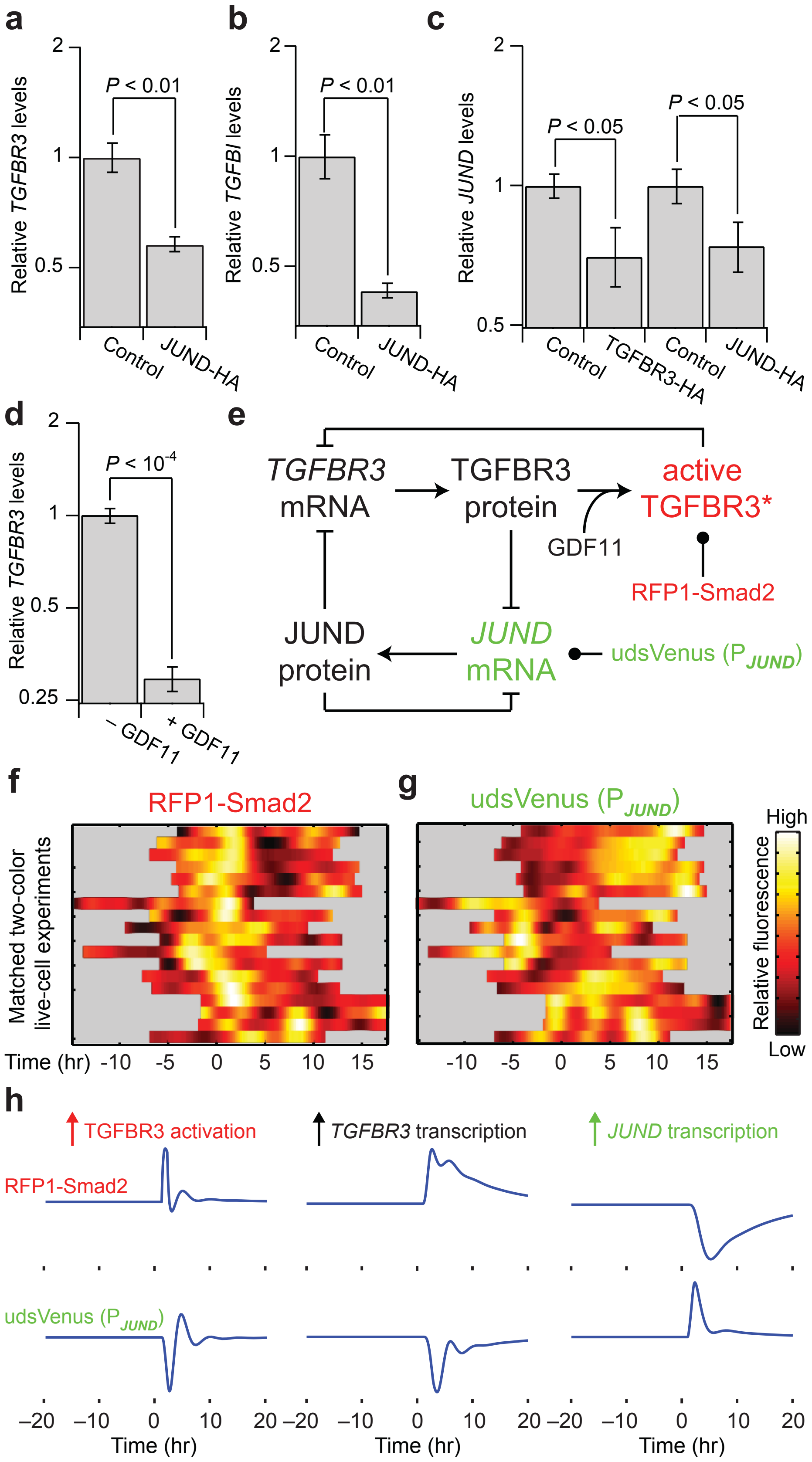

a
Relative *TGFBR3* levels
2
1
0.5
P < 0.01
Control
JUND-HA
b
Relative *TGFBI* levels
2
1
0.5
P < 0.01
Control
JUND-HA
c
Relative *JUND* levels
2
1
0.5
P < 0.05
P < 0.05
Control
TGFBR3-HA
Control
JUND-HA
d
Relative *TGFBR3* levels
2
1
0.5
0.25
P < 10-4
– GDF11
+ GDF11
e
TGFBR3 mRNA
TGFBR3 protein
active TGFBR3*
GDF11
RFP1-Smad2
JUND protein
JUND mRNA
udsVenus (P*JUND*)
f
RFP1-Smad2
Matched two-color live-cell experiments
Time (hr) -10 -5 0 5 10 15
g
udsVenus (P*JUND*)
-10 -5 0 5 10 15
High
Relative fluorescence
Low
h
TGFBR3 activation
TGFBR3 transcription
JUND transcription
RFP1-Smad2
udsVenus (P*JUND*)
–20 –10 0 10 20
Time (hr)
–20 –10 0 10 20
Time (hr)
–20 –10 0 10 20
Time (hr)

Figure 3.2 *JUND* transcription and TGFβ-family signaling activity are functionally and dynamically coupled.

(**a** and **b**) *TGFBR3* and *TGFBI* are repressed by constitutive JUND expression. (**c**) Endogenous *JUND* is repressed by constitutive expression of TGFBR3 or JUND. (**d**) *TGFBR3* is negatively regulated by TGFβ-family signaling. (**e**) Schematic of positive and negative feedback loops connecting *TGFBR3* and *JUND*. The arrows and flat markers indicate the positive and negative relationships from (**a**–**d**). Black circles indicate the two fluorescent reporters (RFP1-Smad2 and udsVenus (P_{JUND})) used to monitor the single-cell dynamics of TGFβ-family activity and *JUND* promoter activity. (**f** and **g**) Multiple alignment of dynamic single-cell fluorescence trajectories. Two-color live-cell confocal imaging was used to quantify the level of nuclear RFP1-Smad2 (left) and total udsVenus (P_{JUND}) expression (right) of ECM-attached cells at day 10 of morphogenesis. Gray indicates no data. (**h**) Damped oscillations in an ordinary differential equations model of the TGFBR3–JUND expression circuit induced by TGFBR3 activation (left; RFP1-Smad2 range: [11.5–15.7], udsVenus (P_{JUND}) range: [13.0–20.6]), *TGFBR3* upregulation (middle; RFP1-Smad2 range: [12.1–23.2], udsVenus (P_{JUND}) range: [0.745–18.4]), or *JUND* upregulation (right; RFP1-Smad2 range: [3.80–12.1], udsVenus (P_{JUND}) range: [18.4–65.2]). In the model, the basal transcription rate was 4 hr^{-1}, the basal translation rate was 100 $mRNA^{-1}$ hr^{-1}, the mRNA degradation rate was 0.23 hr^{-1}, the degradation of TGFBR3 protein was 3 hr^{-1}, the degradation of JUND protein was 0.37 hr^{-1}, the degradation of udsVenus was 2.8 hr^{-1}, and the activation rate of TGFBR3 was 1 hr^{-1} (Figure 3.3). For (**a**–**c**), MCF10A-5E cells stably expressing JUND-HA, TGFBR3–HA, or vector control were placed in 3D culture and analyzed at day 10 of morphogenesis by quantitative PCR for the indicated genes. Endogenous *JUND* was analyzed with primers specific for the 3’ UTR of *JUND*. For (**d**), MCF10A-5E cells were stimulated with 250 ng/ml GDF11 for 4 hr and analyzed for *TGFBR3* expression. Data are shown as the mean ± s.e.m. of four independent samples, and *P* values were calculated by Student’s one-sided *t* test.

We next used computational modeling to test whether the empirical circuit wiring could exhibit damped, antiphase responses like those observed in live cells (Figure 3.2f). The circuit was modeled as a system of ordinary differential equations containing JUND (mRNA and protein) and TGFBR3 (mRNA, protein, and ligand-bound protein – TGFBR3*). We modeled the dynamics of *TGFBR3* mRNA, *JUND* mRNA, TGFBR3 protein, JUND protein, active TGFBR3* (as read out by the RFP1-Smad2 reporter), and the udsVenus reporter through the following system of six ordinary differential equations:

$$\frac{d[\mathit{TGFBR3}]}{dt} = k_{basaltxn} - f_1 \frac{[\mathrm{TGFBR3^*}]^{nH}}{[\mathrm{TGFBR3^*}]^{nH} + (IC50_{\mathrm{TGFBR3}})^{nH}} - f_3 \frac{[\mathrm{JUND}]^{nH}}{[\mathrm{JUND}]^{nH} + (IC50_{\mathrm{JUND}})^{nH}} - k_{degRNA}[\mathit{TGFBR3}] \qquad (1)$$

$$\frac{d[\mathit{JUND}]}{dt} = k_{basaltxn} - f_2 \frac{[\mathrm{TGFBR3^*}]^{nH}}{[\mathrm{TGFBR3^*}]^{nH} + (IC50_{\mathrm{TGFBR3}})^{nH}} - f_3 \frac{[\mathrm{JUND}]^{nH}}{[\mathrm{JUND}]^{nH} + (IC50_{\mathrm{JUND}})^{nH}} - k_{degRNA}[\mathit{JUND}] \qquad (2)$$

$$\frac{d[\mathrm{TGFBR3}]}{dt} = k_{translation}[\mathit{TGFBR3}] - k_{degTGFBR3}[\mathrm{TGFBR3}] - k_{activation}[\mathit{TGFBR3}] \qquad (3)$$

$$\frac{d[\mathrm{JUND}]}{dt} = k_{translation}[\mathit{JUND}] - k_{degJUND}[\mathrm{JUND}] \qquad (4)$$

$$\frac{d[\mathrm{TGFBR3^*}]}{dt} = k_{activation}[\mathrm{TGFBR3}] - k_{degTGFBR3}[\mathrm{TGFBR3^*}] \qquad (5)$$

$$\frac{d[\mathrm{udsVenus}]}{dt} = k_{basaltxn}[\mathit{JUND}] - k_{degudsVenus}[\mathrm{udsVenus}] \qquad (6)$$

The negative feedbacks relating JUND to *TGFBR3* (Figure 3.2a), TGFBR3* to *TGFBR3* (Figure 3.2d), TGFBR3* to *JUND* (Figure 3.2c), and JUND to *JUND* (Figure 3.2c) were modeled as Hill functions. For simplicity, we did not assume any cooperativity in the negative feedbacks ($nH = 1$), and IC50 values were set to nonsaturating values ($IC50_{\mathrm{TGFBR3}} = 100$, $IC50_{\mathrm{JUND}} = 100$). The relative feedback strengths (f_1, f_2, and f_3) were adjusted manually to capture the experimentally observed dynamics (Figure 3.2f,g; $f_1 = 7$, $f_2 = 9$, $f_3 = 5$).

The degradation rates of TGFBR3, JUND, and udsVenus were estimated by treating cells with cycloheximide and quantifying protein loss by immunoblotting (Figure 3.3). These experiments yielded half-life estimates of 14 min for TGFBR3, 1.86 hr for JUND, and 15 min for udsVenus. Using the following relationship between degradation rate and half-life:

$$k_{degradation} = \frac{\ln(2)}{t_{1/2}} \qquad (7)$$

We arrived at the following degradation rate estimates: $k_{degTGFBR3} = 3.0\ \text{hr}^{-1}$, $k_{degJUND} = 0.37\ \text{hr}^{-1}$, and $k_{degudsVenus} = 2.8\ \text{hr}^{-1}$. Degradation of TGFBR3* was assumed to be equal to that of TGFBR3. As the half-lives of *TGFBR3* and *JUND* mRNA are comparable ($t_{1/2}$ = 2–4 hr) (147, 153), we assumed that $k_{degRNA} = 0.23\ \text{hr}^{-1}$ ($t_{1/2}$ = 3 hr). We obtained basal transcription ($k_{basaltxn}$ = 4 hr^{-1}) and translation rates ($k_{translation}$ = 100 $\text{mRNA}^{-1}\ \text{hr}^{-1}$) as representative values from proliferating mammalian cells (154). The basal activation rate of TGFBR3 was calculated using $k_{degTGFBR3}$ and the steady-state ratio of nuclear-cytoplasmic fluorescence of the RFP1-Smad2 reporter in unstimulated cells (assuming that the reporter is directly proportional to the relative activation of TGFBR3c):

$$k_{activation} = k_{degTGFBR3}\frac{[\text{TGFBR3*}]_{SS}}{[\text{TGFBR3}]_{SS}} = 3.0\left(\frac{1}{3}\right)\text{hr}^{-1} = 1.0\ \text{hr}^{-1} \qquad (8)$$

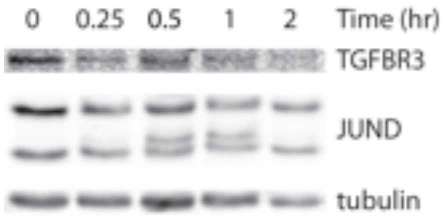


Figure 3.3 Half-life estimates for JUND and TGFBRIII.
Half-life estimates for endogenous TGFBR3 (t1/2 = 14 min) and JUND (t1/2 = 1.86 hr) following cycloheximide treatment. The JUND kinase JNK is activated upon protein synthesis inhibition, causing the upshift in JUND at 30–60 min.

The perturbations in Figure 3.2h were initiated by increasing $k_{activation}$ of TGFBR3, $k_{basaltxn}$ of *TGFBR3*, or $k_{basaltxn}$ of *JUND* to 50% higher than the standard value for 1 hr. The code for generating Figure 3.2h is available in the Chapter 3 Methods. The perturbations in Figure 3.2h were initiated by increasing $k_{activation}$ of TGFBR3, $k_{basaltxn}$ of *TGFBR3*, or $k_{basaltxn}$ of *JUND* to 50% higher than the standard value for 1 hr. The code for generating Figure 3.2h is available in the Chapter 3 Methods. The initial testing of the model showed that the coded circuit could generate anticorrelated oscillations that we observe experimentally.

We further probed the model by trying a variety of perturbations and we found that the system response fell into five categories (Figure 3.4): a) Undamped oscillations that remain in a limit cycle; b) Damped oscillations as in Figure 3.2h (left); c) No oscillations, characterized by a transient activation or repression event as in Figure 3.2h (right); d) Mixed oscillations, where one of the reporters oscillates but the other does not; e) Model error, where the steady-state activity of one reporter is near zero and the system no longer responds to the perturbation or returns infeasible values. Focusing on the damped oscillations that were noted upon TGFBR3 activation in the model (Figure 3.2), we performed a sensitivity analysis. For the six parameters that were not drawn from the literature or directly measured (f_1, f_2, f_3, nH, $IC50_{\mathrm{TGFBR3}}$, $IC50_{\mathrm{JUND}}$), we systematically perturbed the default parameter by tenfold in either direction and then assessed system behavior. This sensitivity analysis would indicate how fragile or robust the oscillatory network was to feedback parameters (148).

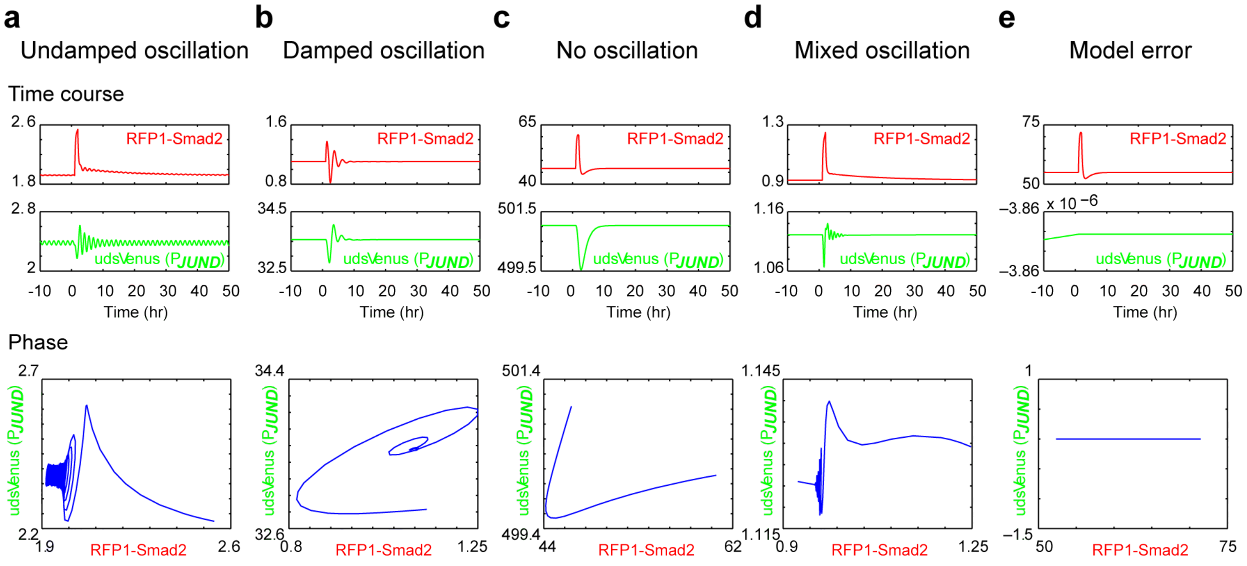


Figure 3.4 Categories of TGFBR3-JUND model response.

Time courses are shown in the upper plots and a phase-plane representation is shown in the lower plot.

For nearly all feedback parameters, we found that the system exhibited damped or undamped oscillations for a wide range of parameters (Figure 3.5). The one exception was for the relative strengths of the negative feedbacks from active TGFBR3 signaling and JUND expression to the expression of *TGFBR3* mRNA, where damped oscillations were observed over a somewhat narrow window (Figure 3.5e). If the three feedback terms were reasonably balanced, then oscillations were robust to any single change in feedback (Figure 3.5c-e). However, if the collective negative regulation on JUND was changed by concurrently increasing f_2 and f_3, then oscillations stopped. This emphasizes the need for tight coupling between the TGFBR3 and JUND branches of the circuit (Figure 3.2e), which may explain why not all matrix-attached basal breast epithelia oscillate during 3D culture, or why there is mosaicism in JUND or TGFBRIII protein levels amongst the matrix-attached basal breast epithelia.

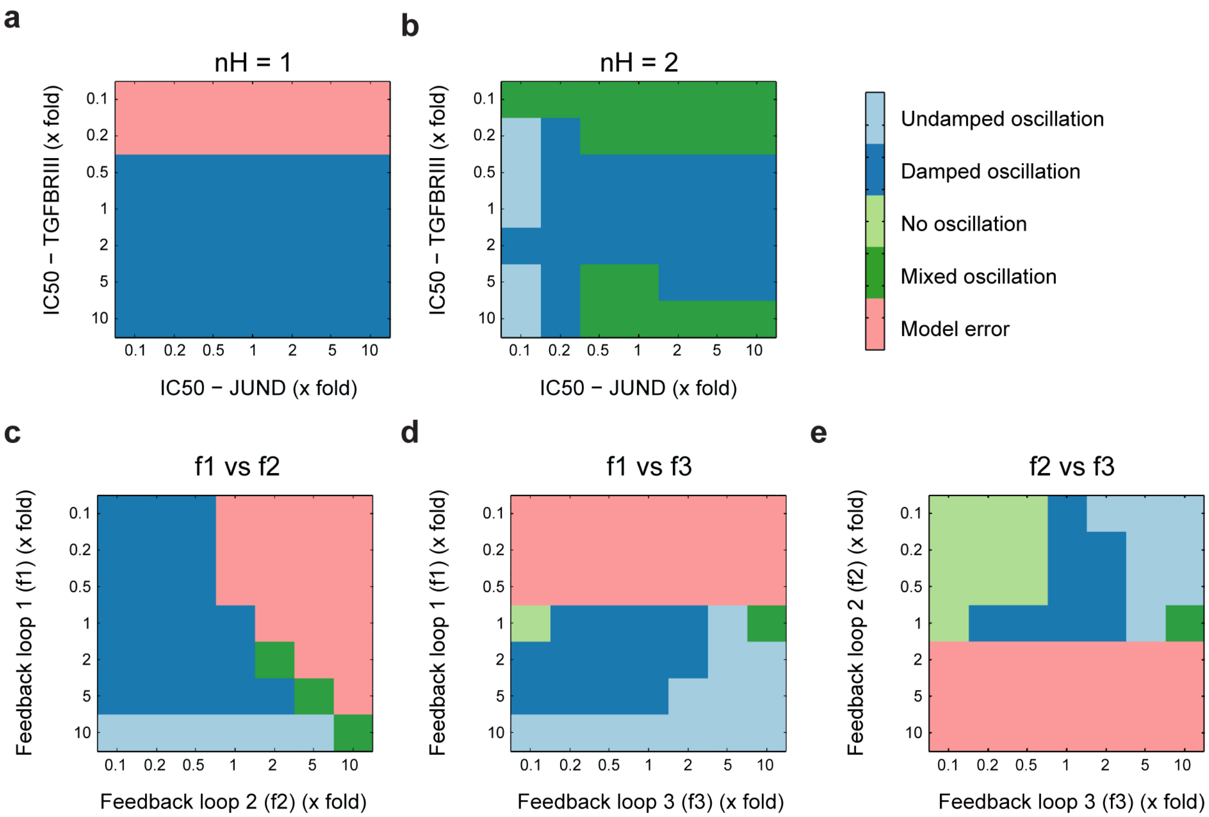


Figure 3.5 Pairwise sensitivity analysis of the TGFBR3-JUND model in response to transient activation of TGFBR3.

Model parameters were changed from 0.1x to 10x of their values listed in the Chapter text. See Figure 3.4 for examples of each category of model response.

3.3.3 Stabilization of anticorrelated JUND–KRT5 by TNC

Cell-to-cell mosaicism is observed clinically in basal-like breast cancer, where ~50% of cases show highly nonuniform expression of basal cytokeratins for the subtype (60, 71). Interestingly, one diagnostic cytokeratin, *KRT5*, lies within the *JUND* cluster, and *KRT5* is tightly coexpressed with *JUND* in ECM-attached cells (73). The JUND–TGFBR3 expression circuit might therefore be specifically engaged in basal-like carcinomas or premalignant lesions with basal-like features (155) ("basal-like DCIS").

To explore the relationship between JUND and TGFBR3 in this clinical context, we collected an independent cohort of premalignant basal-like DCIS lesions with heterogeneous KRT5 expression (71, 155). KRT5 is an important indicator of poor prognosis for basal-like carcinoma (59), and heterogeneous premalignancies would allow the cell-by-cell correlations of KRT5 to be examined with JUND and TGFBR3 while the tissue architecture was still intact. In normal breast tissue, we found that KRT5 and TGFBR3 were strongly expressed in the basal layer. KRT5 was predominantly localized to the ductal myoepithelia, whereas TGFBR3 was expressed mostly in the lobular myoepithelia. Conversely, JUND protein was very low in normal tissue but increased substantially in basal-like DCIS, where TGFBR3 was often undetectable. These results indicated a switch in TGFBR3–JUND–KRT5 regulation during premalignancy.

Next, we directly examined the coexpression of KRT5 and TGFBR3 or JUND in single cells by multicolor immunofluorescence. In the 59% of premalignant lesions where TGFBR3 could be detected, expression of TGFBR3 and KRT5 remained mutually exclusive (156) (Figure 3.6a). This single-cell anticorrelation was consistent with both our stochastic profiling of basal-like cultures. Conversely, in cases with KRT5-positive regions of primary DCIS (41% of total), there was a strong positive correlation between KRT5 and JUND among single cells (Figure 3.6b). The striking agreement between the clinical and *in vitro* studies suggests that basal-like ECM cultures may mimic the burst of proliferation and environmental stress experienced by early neoplasms (97, 99, 100).

High-grade intraductal carcinomas are frequently comprised of a primary DCIS region along with secondary regions of "clinging carcinoma" (CC) (157). CC forms when neoplastic cells disseminate intraluminally from the DCIS and cancerize peripheral breast lobules and ducts. When CC regions were carefully examined for JUND and KRT5, we discovered that the two proteins were anticorrelated (Figure 3.6c). JUND–KRT5 switching occurred without gross cytological changes in cases with both DCIS and CC. Tumor geography thus appeared to provide some sort of external control on the TGFBR3–JUND expression circuit and its coregulation with KRT5.

The dramatic reversal of JUND–KRT5 coexpression prompted us to reexamine their relationship *in vitro*. During 3D culture, JUND and KRT5 proteins were coordinately expressed among outer cells. For interior cells, however, the JUND–KRT5 coexpression pattern was anticorrelated. This transition could not have been anticipated by our initial profiling study, which focused exclusively on outer cells (73). Nonetheless, the finding provided an independent replication of the JUND–KRT5 switching observed in basal-like tumors (Figure 3.6). We were corroborated by a few exceptional cases of DCIS where cells had detached partly or entirely from the tumor margin and JUND–KRT5 coexpression was reversed.

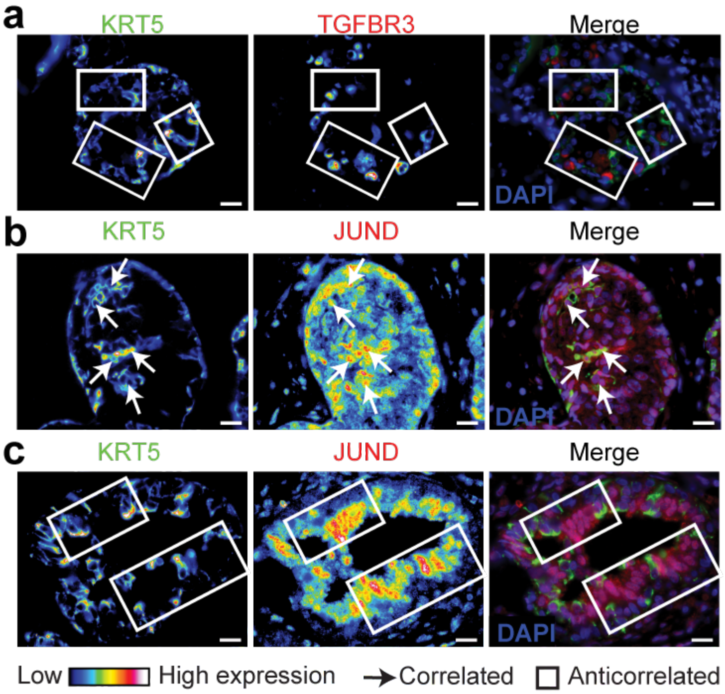


Figure 3.6 TGFBR3 and JUND expression reciprocally map to KRT5 in specific regions of heterogeneous basal-like premalignancies.

(**a**) Expression of TGFBR3 and KRT5 proteins is mutually exclusive in ER-negative premalignant lesions. (**b**) Expression of JUND and KRT5 proteins is correlated in ER-negative ductal carcinoma in situ (DCIS). (**c**) Expression of JUND and KRT5 is anticorrelated in peripheral regions of clinging carcinoma. Paraffin sections from basal-like premalignant lesions were stained for KRT5 (green) and TGFBR3 (red; **a**) or JUND (red; **b**, **c**) and imaged by widefield immunofluorescence. Nuclei were counterstained with DAPI (blue). Single-color fluorescence images are pseudocolored in the first two subpanels of (**a**), (**b**), and (**c**) to highlight quantitative differences in immunoreactivity. Correlated and anticorrelated regions of expression are indicated with arrows and rectangles respectively. Scale bars are 20 µm.

To identify the molecular basis for the JUND–KRT5 inversion, we considered the variegated microenvironments of ECM cultures and human tumors. The most-recognized difference within ECM cultures is the spatially segregated access to basement membrane (97, 143, 158). Outer cells contact the ECM-rich culture support and also secrete their own ECM molecules basolaterally (97, 159). Inner cells are deprived of both these ECM sources and thus should be starved for integrin engagement. Analogously, in regions of DCIS, the local tumor stroma is potently activated, providing ECM to the primary tumor (160). Cells in CC regions have left the primary site to colonize luminal, ECM-poor regions of the ductal tree and may behave like inner cells of the culture.

To simulate ECM deprivation, we placed cells in suspension culture. Before anoikis was evident, we observed clear changes in single-cell JUND–KRT5 expression that were highly stereotyped. For the first 8 hr, JUND–KRT5 were coexpressed as double-positive or double-negative cells. At 24 hr, the JUND^{+}–KRT5^{+} and JUND^{-}–KRT5^{-} subpopulations became more-clearly separated, when KRT5 increased with a filamentous pattern (KRT5^{F}) and anticorrelations started to appear. By 48 hr, JUND^{+}–KRT5^{F} cells had vanished, and a fourth "keratinized" state emerged with intense KRT5 staining and no JUND protein or nuclear DNA (KRT5^{K}). Live-cell imaging showed that progression to the JUND^{-}–KRT5^{K} state was rapidly executed, with keratinized skeletons eventually collapsing as cellular dust (Figure 3.7). These late cellular steps are highly reminiscent of cornification, a cell-death process typically associated with skin (161).

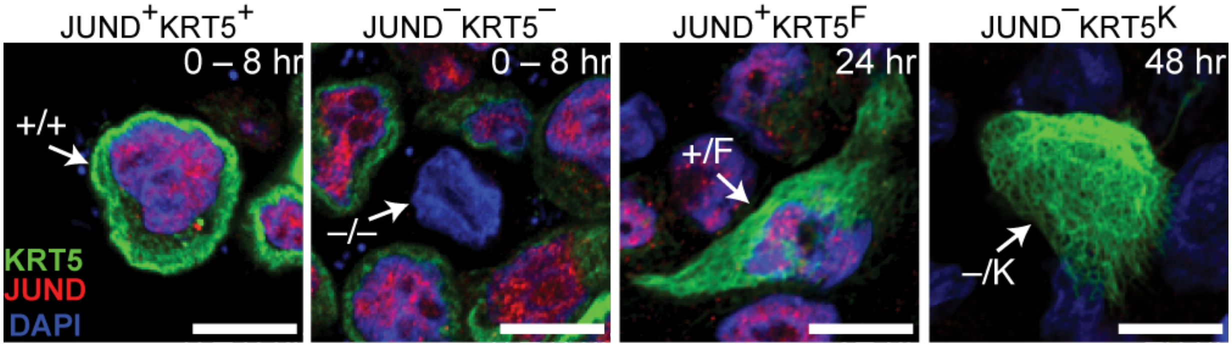


Figure 3.7 Keratinization in detached breast epithelial cells.

Loss of ECM-attachment induces keratinization and JUND–KRT5 anticorrelation. MCF10A-5E cells were placed in polyHEMA-coated plates with assay medium containing 5 ng/ml EGF. Cells were fixed at the indicated times, stained for KRT5 (green) and JUND (red), and analyzed by confocal immunofluorescence. Cells were counterstained with DAPI (blue) to label nuclei Single cells representing intermediate stages of keratinization are highlighted with arrows. Scale bar is 10 μm.

Endogenous JUND levels increased transiently before KRT5 upregulation, suggesting a role in the sequelae of ECM detachment. Although ectopic expression of JunD did not affect the induction of KRT5 protein, JunD-overexpressing cells largely remained in a double-positive state without overt keratinization ($P = 7.0 \times 10^{-5}$ and 3.0×10^{-6}, two-sided *t* test). Conversely, JUND knockdown accelerated keratinization and augmented it, with KRT5^{K} cells apparent as early as 8 hr ($P = 2.4 \times 10^{-4}$ and 8.7×10^{-4}, two-sided *t* test), even though KRT5 upregulation was unaffected. We conclude that JUND restrains detachment-induced keratinization but is independent of the upregulation of KRT5 itself.

Keratinization provided an appealing mechanism for the JUND–KRT5 anticorrelation observed in inner cells and in CC regions of basal-like breast cancer (Figure 3.6c). It also created a paradox—if detachment from ECM rapidly causes irreversible loss of JUND and upregulation of KRT5, why do JUND^{+}–KRT5^{-} cells exist at all? We carefully inspected the JUND–KRT5 staining pattern and noted that keratinized JUND^{-}–KRT5^{K} cells were often surrounded by JUND^{+}–KRT5^{-} cells, suggesting that JUND^{-}–KRT5^{K} cells could be exchanging juxtacrine signals with JUND^{+}–KRT5^{-} cells.

To identify candidate ECM ligands that could fulfill this role, we screened breast cancer immunohistochemistry from The Human Protein Atlas (162, 163). Among 71 ligands (164), only one was expressed heterogeneously in a cell-intrinsic manner within CC regions of breast carcinoma: the matricellular protein, TNC (165). TNC plays a critical role in early colonization of breast-cancer metastases to the lung (166). Sporadic TNC expression has also been noted in basal keratinocytes (167), suggesting a connection to epidermal keratinization. We hypothesized that TNC could stabilize JUND^{+}–KRT5^{-} cells if it were endogenously expressed *in vitro* and *in vivo*.

In ECM cultures, we found inner cells that strongly expressed TNC (Figure 3.8a). Interestingly, $JUND^{+}$–$KRT5^{-}$ cells appeared to extend lamellipodia around TNC^{+}–$JUND^{-}$–$KRT5^{K}$ skeletons (Figure 3.8a, inset), suggesting extensive adhesive contacts. In CC regions of basal-like premalignancies, TNC^{+}–$JUND^{-}$–$KRT5^{+}$ cells were similarly in direct apposition with cells that were $JUND^{+}$–$KRT5^{-}$ (Figure 3.8b). Upon ECM withdrawal, TNC was strongly upregulated in the $KRT5^{F}$ and $KRT5^{K}$ subpopulations (Figure 3.8c). Unlike other keratinization-related programs, TNC upregulation may be transcriptionally mediated. Only 2–6% of detached cells expressed TNC, but ~60% of keratinized cells were TNC^{+}. Importantly, when TNC was added to 2D cultures of basal-like breast epithelia, the single-cell JUND–KRT5 correlation reversed (Figure 3.8d), illustrating that TNC actively participates in anticorrelating JUND and KRT5 expression.

To determine whether TNC could explain the JUND–KRT5 mosaicism in ECM-poor microenvironments, we built a multi-cell agent-based model of CC (168). We coded for an arbitrary CC geometry, where individual cancer cells ("agents") spontaneously keratinize as a function of their JUND–KRT5 levels and the neighboring expression of TNC (see Chapter 3 Methods). The purpose of the agent-based model was to build a simplified representation of the mosaic KRT5–JUND expression patterns observed in ECM-poor regions of basal-like premalignancies. Agent-based models allow the arbitrary arrangement of "agents" (here, clinging carcinoma [CC] cells) to react subject to a user-defined rule set that is simulated in discrete time intervals (168). The process of keratinization happens within hours and is triggered before 24 hr of detachment (Figure 3.7). Therefore, we assumed that the effects of proliferation, death, and migration would be negligible over this time period. Notably, detachment-induced cell death (anoikis) of basal breast epithelia is not maximal until 48 hr or later (Ref. (158)) and epithelial proliferation is generally minimal without integrin engagement (169).

The key facets of the model rule set are:

- JUND and KRT5 compete to determine survival vs. keratinization
- Keratinization is irreversible after the nucleus has been lost (**Figure 3.7**)
- Keratinization is strongly associated with the expression of TNC
- Cells adjacent to TNC-expressing cells often have not keratinized (**Figure 3.8**a,b)

For the model, cells were seeded according to the characteristic geometry of a region of CC—multiple layers of neoplastic cells on the periphery of a hollow lumen. In addition to the geometry of Figure 3.8e,f, we also tested alternative geometries with differing thicknesses of cells (Figure 3.9). We consistently found strings of keratinized cells on the luminal face of the clinging region, as well as local homogeneities of JUND expression. These additional simulations indicate that our conclusions are not limited to a specific geometry of CC. Source code for the NetLogo script and tissue geometries can be found in the Chapter 3 Methods.

Without TNC, we found that virtually all cells keratinized (Figure 3.8e), consistent with the irreversibility of keratinization in the model. By contrast, including TNC caused a stable mosaic of cells that were JUND^{-}–KRT5^{K} or JUND^{+}–KRT5^{-} (Figure 3.8f). Importantly, this model made two predictions that were subsequently verified in clinical specimens. First, keratinization should be extensive among cells immediately adjacent to the lumen because of fewer opportunities to be stabilized by adjacent TNC-positive cells (Figure 3.8f, solid). Retrospectively, we identified many stretches of keratinized cells along CC lumina (Figure 3.8g). Second, the model predicted multi-cellular clusters that were locally homogeneous for JUND (Figure 3.8f, dashed). The reason here is that JUND increases up until keratinization occurs, and

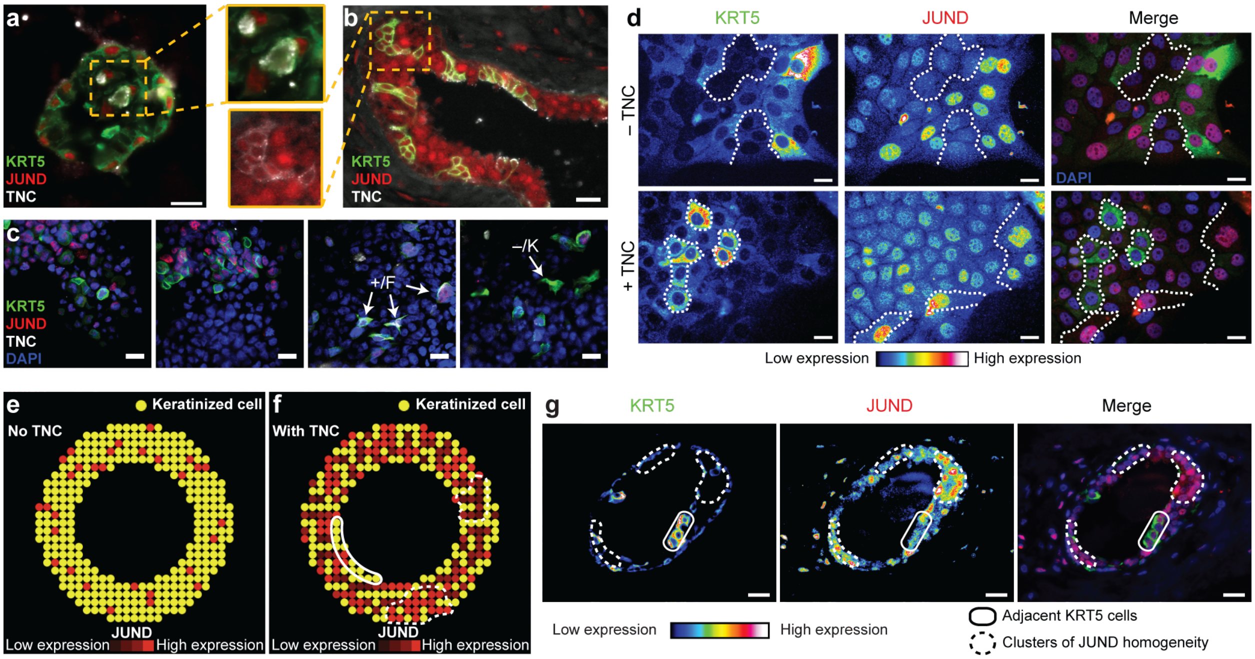


Figure 3.8 JUND-KRT5 mosaicism in ECM-poor environments is stabilized by TNC.
(**a** and **b**) The JUND–KRT5 anticorrelation state reflects a microenvironment that lacks basement membrane but contains TNC. Day 10 frozen sections of MCF10-5E acini (**a**) and paraffin sections from premalignant basal-like neoplasms (**b**) were stained for KRT5 (green), JUND (red), and TNC (white) and imaged by widefield immunofluorescence. (**c**) TNC protein expression is upregulated during detachment. MCF10A-5E cells were placed in suspension for the indicated times. Cells were fixed and stained for KRT5 (green), JUND (red), and TNC (white) and analyzed by confocal immunofluorescence. Cells were counterstained with DAPI (blue) to label nuclei. (**d**) The JUND–KRT5 correlation is reversed *in vitro* by exogenous TNC. MCF10A-5E cells were grown on coverslips in assay medium(98) + 5 ng/ml EGF in the presence or absence of 5 µg/ml TNC for 8 days. The cells were stained with antibodies against KRT5 (green) and JUND (red) and imaged by widefield immunofluorescence. Nuclei were counterstained with DAPI (blue). In the first two panels, single-color fluorescence images are pseudocolored to highlight quantitative differences in immunoreactivity. Dashed lines separate regions that stain strongly or weakly for KRT5 expression. (**e** and **f**) An agent-based model requires a TNC-like molecule to stabilize JUND–KRT5 expression patterns. Solid lines highlight strings of keratinized cells adjacent to the lumen (yellow). Dashed lines highlight clusters of locally homogeneous JUND expression (red). (**g**) Paraffin sections from early basal-like carcinomas were stained for KRT5 (green) and JUND (red) and imaged by widefield immunofluorescence. Nuclei were counterstained with DAPI (blue). In first two panels, single-color fluorescence images are pseudocolored to highlight quantitative differences in immunoreactivity. Strings of keratinized cells (solid) and clusters of local JUND homogeneity (dashed) are highlighted. For (**a**–**d**) and (**g**), scale bar is 20 µm. For simulation code, see Chapter 3 Methods.

TNC-positive cells “corral” the multi-cellular clusters at different times during the model simulation. A similar mechanism may operate in CC, because we uncovered several multi-cell clusters with roughly equal JUND expression, even though lesions were heterogeneous overall (Figure 3.6c and Figure 3.8g). We conclude that keratinization—triggered by detachment-induced RPS6 dephosphorylation and modulated by TNC—is responsible for the single-cell anticorrelation of JUND–KRT5 in basal-like CC.

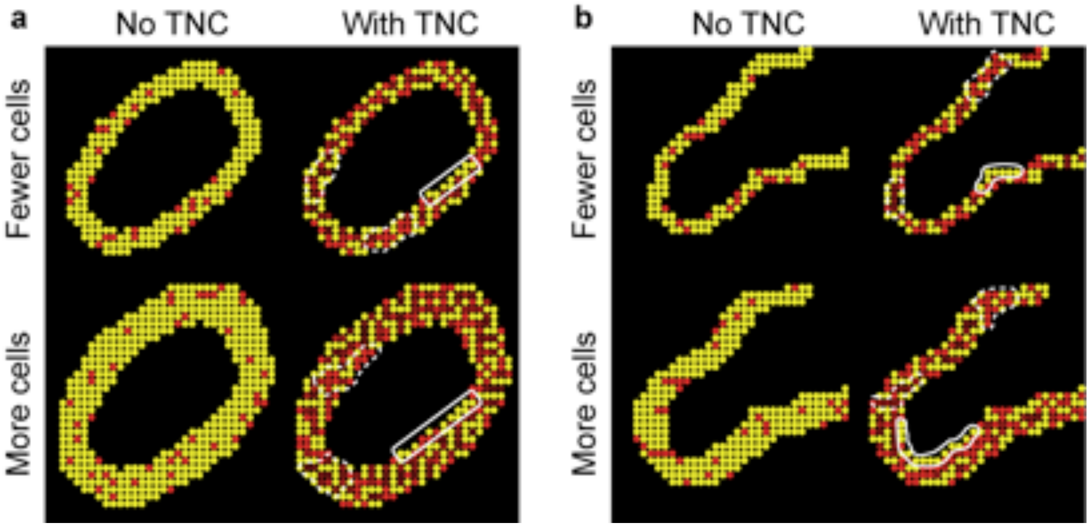


Figure 3.9 Agent-based model predictions do not depend on the specific geometry of clinging carcinoma. (**a**) Elliptical and (**b**) open-diagonal geometries were seeded with various thicknesses of carcinoma cells and simulated as described in the manuscript. Keratinized cells (yellow) and JUND levels (red) are shown with and without the simulated role of TNC. Stretches of luminally positioned, keratinized cells are highlighted in solid boxes. Local homogeneities of JUND expression are highlighted in dashed boxes.

3.4 Discussion

By profiling expression heterogeneities in a relevant ECM context, we have uncovered a major signaling circuit within basal-like breast epithelia. Cells in contact with basement membrane undergo transient oscillations between two molecular states defined by their *TGFBR3–JUND* expression. Perturbation of either state profoundly disrupts normal acinar morphogenesis. Proper dynamic regulation of the circuit must be critical for establishing and stabilizing the identity of ECM-attached cells. By extension, proliferating neoplasias may reengage the TGFBR3–JUND circuit in search of a cell fate amidst a heterogeneous ECM microenvironment.

Our study began with a transcriptional dichotomy between two single-cell expression states, but the overall circuit extends beyond transcription. Circuit activation in ECM-attached cells likely occurs via posttranslational signaling from TGFβ-family receptors. Interestingly, TGFβ ligands bind ECM and exist as latent complexes that become disinhibited by mechanical force (170). Considering that breast epithelia are known to be mechanoresponsive (45), the earliest trigger for circuit oscillations may be changes in local cell-ECM mechanics.

The ECM-dependent relationship between KRT5 and TGFBR3–JUND is reflected both in basal-like cultures and preinvasive basal-like neoplasias. Outer ECM-attached cells of cultured acini and primary DCIS show correlated expression of JUND–KRT5. By contrast, the inner ECM-deprived cells of a 3D acinus may mimic facets of preinvasive dissemination that partly explain the macroscopic heterogeneity of clinical specimens. Detached epithelial cells stochastically execute a keratinization program (171), which delays anoikis by creating a TNC mosaic within 3D cultures and in CC. Breast-cancer patients with TNC-positive tumor cells frequently have lymph-node metastases and very-poor prognosis (172). Our data build on recent animal models (166) by suggesting that juxtacrine TNC may be critical for secondary orthotopic colonization within the duct.

The most-recognized driver of late-stage tumor heterogeneity is genomic instability, but how heterogeneous tumors evolve from premalignancy has been more enigmatic. Our work here places renewed emphasis on the microenvironment and the dynamic asynchronicity of the constituent cells. Reversible lineage switching has been described in several contexts (52, 173), suggesting together with our results that breast cancer may be far more dynamic than previously appreciated. Lastly, we demonstrate that computational modeling, of different frameworks, can reduce the complexity of heterogeneity and reveal important insight into the role heterogeneity plays in the context of interest.

4 Chapter 4 – Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles

4.1 Foreword

Stochastic profiling uses repeated random gene expression measurements from a small (e.g., ten) number of cells to reveal heterogeneously regulated transcripts. By using more than one cell, one avoids the increase in measurement noise that occurs with low starting material. Conversely, using a smaller number of cells avoids averaging out infrequent regulatory states. Despite these advantages, stochastic profiling is limited to giving a qualitative assessment on whether a given gene or gene program is heterogeneously regulated (see Section 1.4). Here, we present a computational model that infers quantitative parameters of heterogeneous gene regulation, extending the utility of stochastic sampling. With this approach, we can now obtain single-cell information without using single-cell techniques (see Section 1.3). This work was done in collaboration with Christiane Fuchs, Andreas Roller, and Fabian Theis of the Helmholtz Center, Munich. Additionally, this work was published in *The Proceedings of the National Academy of Sciences* (*Proc Natl Acad Sci*, 111, E626-35).

4.2 Introduction

Cell-to-cell differences in transcriptional or posttranslational regulation can give rise to heterogeneous phenotypes within a population (22, 52, 106, 130, 131, 133, 174). There are several elegant techniques for monitoring regulatory states in single cells after a network of marker and effector genes has been identified (20, 80, 116, 175-177). However, the options are much more limited when seeking to discover novel states without a predefined network. At the transcript level, global methods have been developed to profile single cells by oligonucleotide microarrays (178, 179) or RNA-seq (13, 19, 89, 90). But generally, such approaches overlook the considerable technical variation in RNA extraction (72) and reverse-transcription (95) when applied to the limited starting material of single cells. Single-cell profiles also retain the biological noisiness (180) associated with each cell's isolation and handling. These confounding sources of variation cannot be separated from reproducible heterogeneities in regulation unless many (>50) cells are individually profiled (175). Therefore, challenges remain for single-cell methods to discover regulatory heterogeneities in a reliable, unbiased, and efficient way.

An attractive alternative to single-cell methods is to analyze sets of population-averaged data and define regulatory signatures for discrete subpopulations. Existing approaches for transcriptomic data are able to deconvolve mixed cellular states computationally, but they require hundreds of coexpressed markers (181) or calibration with purified cell populations (182, 183). Usually, the size or identity of regulatory states is not defined beforehand and their discovery is what motivates the study (20, 175, 184). Certain states may also lack well-defined surface markers that would allow purification. It thus remains unclear whether computational inference with multiple cell averages can track quantitative characteristics of regulatory states not previously thought to exist.

As a hybrid between single-cell and mixture-based approaches, we previously developed a technique that applies probability theory to transcriptome-wide measurements (73). The method begins with random collections of up to 10 cells isolated in situ where cell-to-cell regulatory heterogeneities could possibly reside. Each of these "stochastic samples" is then profiled for overall mRNA expression by using a heavily customized cDNA amplification procedure together with oligonucleotide microarrays (72, 73). The process of random sampling is repeated 15–20 times to build a distribution of 10-cell averages. Transcripts with stark cell-to-cell variations can be distinguished statistically because of binomial fluctuations in single-cell expression that convolve their 10-cell averages. Last, candidate heterogeneities are clustered on a gene-by-gene basis according to the patterns of their sampling fluctuations to indicate putative regulatory states in single cells (73).

Stochastic-profiling experiments are quantitative and highly reproducible as a result of the 10-fold increase in starting material compared to a single cell (72). However, a recognized drawback of the approach is that explicit information about single cells is "lost" in the 10-cell averages. Here, we report that one can recover this information computationally and reconstruct the single-cell distribution of regulatory states with remarkable accuracy. Our method combines maximum-likelihood estimation with mixture models that are grounded in known mechanisms of transcriptional regulation. This approach of maximum-likelihood inference quantifies the single-cell characteristics of each regulatory state, including the probability that a cell will reside in one state or the other. Our predictions are validated with independent gene-specific observations in single cells, and we demonstrate for one very-rare state (~2–3% of the population) that it is important for normal morphogenesis of breast epithelial cells in 3D culture. Last, we show that when sampling is limited to fewer than 20 observations, the parameterization of regulatory states is substantially more accurate when given 10-cell data compared to one-cell data. Maximum-likelihood inference now enables stochastic profiling to bridge the gap between -omic datasets and single-cell information.

4.3 Results

4.3.1 Probability models for heterogeneous transcriptional regulation

To make reliable single-cell inferences, it was critical to start with simple probabilistic models of gene expression that were biologically accurate. Our method considers genes that exhibit two distinct regulatory states in a population of cells (19, 72, 73). Within each state, the cell-to-cell variation of expression was originally described by a lognormal distribution according to measurements of high-copy transcripts in single mammalian cells (185, 186). We tested whether there was a mechanistic foundation for using two lognormal subpopulations by examining a standard model of regulated gene expression (3, 187) (Figure 4.1). In this model, transcript levels per cell are determined by the kinetics of polymerase binding-unbinding, transcriptional elongation, and mRNA degradation. The relative magnitudes of the kinetic rate parameters together govern the steady-state distribution of transcripts in the population (188), allowing different regulatory states to be simulated.

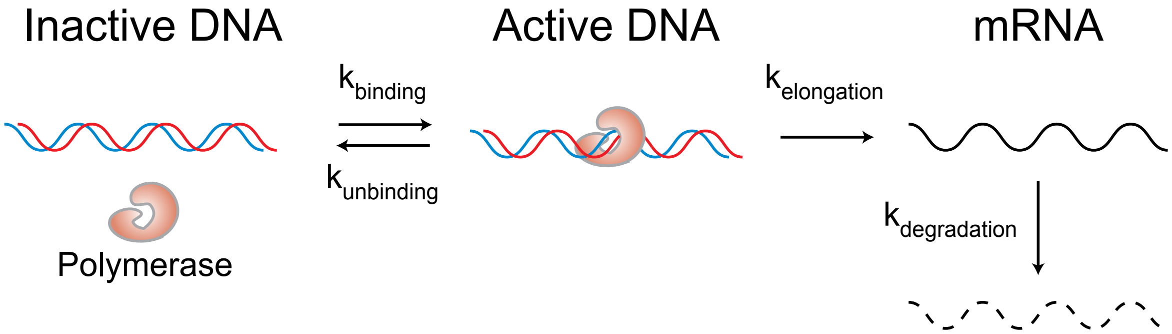


Figure 4.1 Kinetic model of gene activation and mRNA expression.

DNA activity is regulated by the rate of polymerase binding ($k_{binding}$) and the rate of polymerase unbinding ($k_{unbinding}$). Polymerase-bound DNA then transcribes mRNA copies at a defined elongation rate ($k_{elongation}$). The half-life of the mRNA species is determined by its degradation rate ($k_{degradation}$ = log(2)/t1/2). The mRNA distribution is calculated at the steady state of this model (3, 187).

For parameter sets where the probability of observing zero transcripts per cell was near zero, we found that the lognormal distribution was a suitable approximation of basal expression (Figure 4.2*A*, blue). Parameter sets yielding median expression levels as low as 20 copies per cell showed only minor skewness in quantile-quantile (QQ) comparisons with a lognormal distribution (Figure 4.2*A*, blue inset). Starting with this basal distribution, we simulated a second cellular regulatory state by increasing the rate of polymerase binding, decreasing the rate of mRNA degradation, or both (Figure 4.2*A*, orange). The apparent rate of polymerase binding increases upon recruitment by transcription factors that are expressed or activated heterogeneously within a population of cells (130, 131). Conversely, mRNA stabilization occurs posttranscriptionally through dedicated signal-transduction pathways activated by environmental stresses and proinflammatory stimuli (189). We found that either mechanism of gene upregulation led to right-shifted distributions that were lognormal (Figure 4.2*A*, orange inset). These simulations indicated that lognormal random variables were appropriate for the regulated expression of mid- to high-abundance transcripts.

One drawback of the lognormal distribution is that it has no support at zero copies (190), making it poor for capturing low-abundance genes that are completely silenced in some cells. To identify an alternative in this circumstance, we reconfigured the parameters of the model and defined a steady-state population where most cells would contain zero transcripts (Figure 4.2*B*, blue). As noted before (188), this regulatory state was best captured by an exponential distribution (Figure 4.2*B*, blue inset). Importantly, we found that when the kinetic parameters of

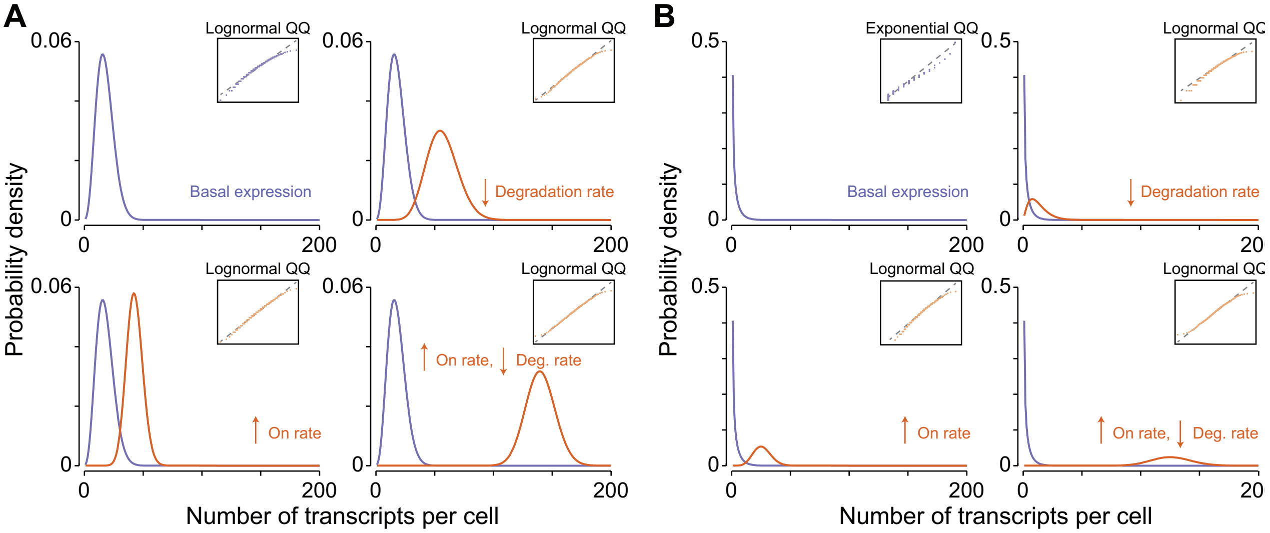


Figure 4.2 Simple probablity models capture regulated changes in gene expression.
(*A* and *B*) Probability densities for the number of transcripts per cell were calculated using a kinetic model (3, 187) whose parameters led to basal regulatory states (blue) with either nonzero copies per cell in *A* or with zero copies per cell in *B*. The basal-regulatory states were compared to a lognormal distribution in *A* or an exponential distribution in *B* through a quantile-quantile (QQ) plot (blue insets). A second, induced regulatory state (orange) was created by increasing the polymerase binding rate (lower left), decreasing the transcript degradation rate (upper right), or both (lower right) in the model (Figure 4.1). All induced regulatory states were compared to a lognormal distribution through a QQ plot (orange insets).

a basal exponential state were modified to create a second right-shifted state (Figure 4.2*B*, orange), the resulting distributions were lognormal (Figure 4.2*B*, orange inset). Together, we conclude that the basic mechanisms of gene expression lead to steady-state distributions described by probability models that are relatively simple.

4.3.2 Deconvolution of random 10-cell averages by maximum-likelihood inference

Our results from the gene-expression model suggested that single-cell regulatory heterogeneities could be depicted as a mixture of two lognormal states or as a mixture of an exponential state and a lognormal state (Figure 4.2). Either mixture gives rise to a probability distribution characterized by four key parameters. The lognormal-lognormal (LN–LN) mixture requires the log-mean expression of the two regulatory states (μ_1, μ_2), the log-standard deviation for biological noise (σ), and the expression frequency (F) describing the probability that cells will occupy the higher regulatory state (Step 1; Figure 4.3*A*). (For simulations, the two lognormal states are assumed to share a common σ, but in practice we test whether inferences are improved when each lognormal state is allowed its own noise parameter; see below.) Thus, an LN–LN gene that is expressed at an ~8 fold higher level in 20% of the population with a coefficient of variation (CV) of ~50% is captured by $\mu_1 - \mu_2 = 2$, $F = 20\%$, and $\sigma = 0.48$.

The exponential-lognormal (EXP–LN) mixture also requires σ and F, along with a single log-mean for the high lognormal state (μ) and a rate parameter for the low exponential state (λ) (Step 1; Figure 4.4). The rate parameter relates to how quickly the lower distribution decays above zero copies per cell. For example, a rate parameter of $\lambda = 1$ creates a distribution that has ~37% overlap with that of a high lognormal state of $\mu = 0.5$ and $\sigma = 0.225$ when $F = 50\%$, whereas $\lambda = 3$ causes only a ~6.3% overlap. We modeled two distinct regulatory states by restricting the simulations to rate parameters that caused negligible overlap with the high lognormal state ($\lambda > 3$). Together, the different mixture models enabled us to simulate stochastic-profiling data by summing the expression of 10 cells randomly sampled from the appropriate two-state distribution (Step 2; Figure 4.3*A* and Figure 4.4).

To infer the most-likely parameters from a collection of random 10-cell samples, we derived maximum-likelihood estimators for the LN–LN and EXP–LN mixtures (see Chapter 4 Methods). Maximum-likelihood estimation requires a defined probability density function (pdf). The stochastic-sampling pdf is the convolution of 10 binomial choices drawn from the two underlying distributions in the mixture (Step 3; Figure 4.3*A* and Figure 4.4). The pdf has an ≤11-modal shape where each mode corresponds to choosing 0 to 10 cells from the high regulatory state. The most-likely parameter combination was calculated by maximizing the likelihood function (see Chapter 4 Methods), yielding parameters with interval estimates that best explained the data (Step 4; Figure 4.3*A* and Figure 4.4). By performing this maximum-likelihood estimation, we could "invert" stochastic profiling data to infer single-cell characteristics from 10-cell samples.

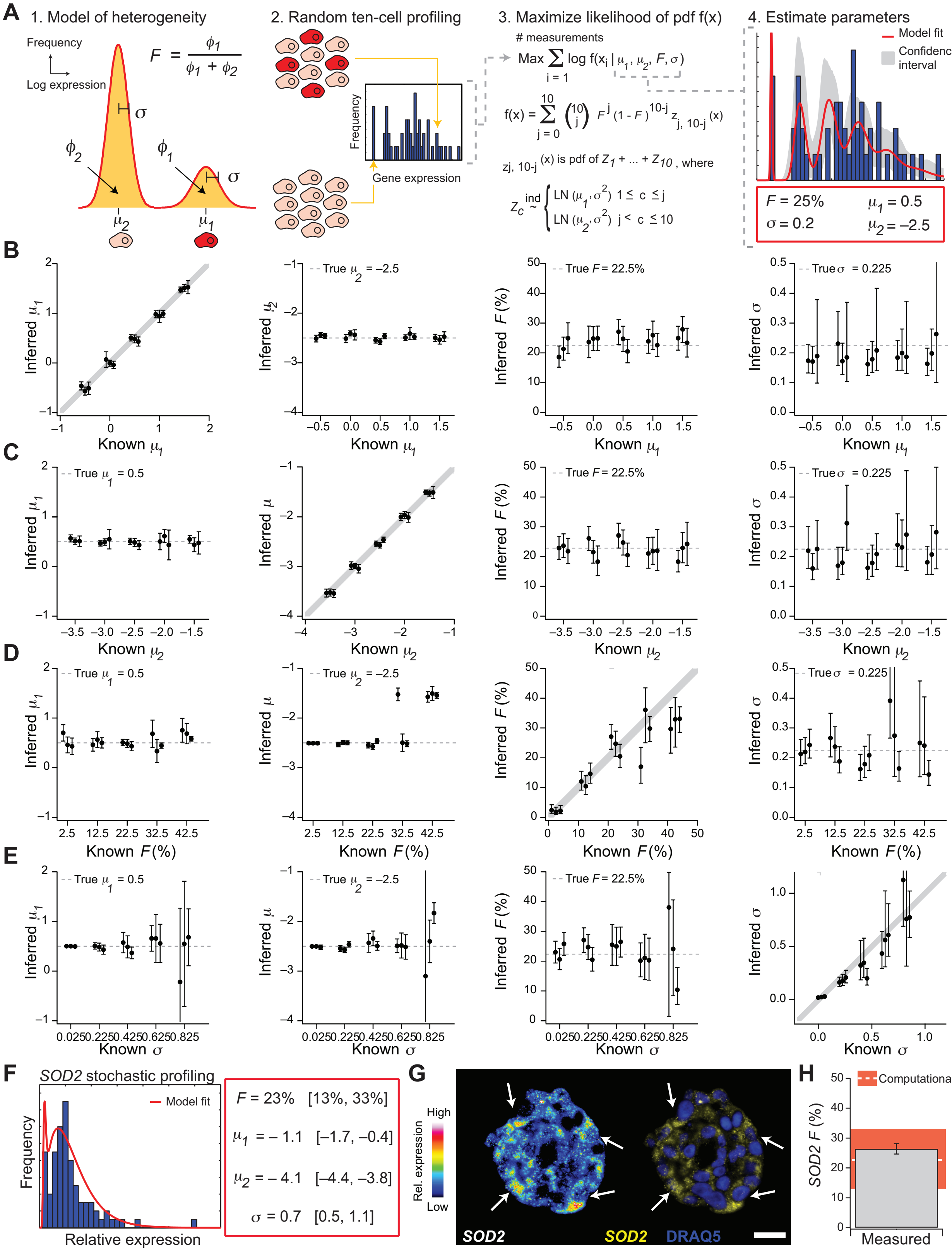

A
1. Model of heterogeneity
Frequency
Log expression
$F = \frac{\phi_1}{\phi_1 + \phi_2}$
σ
ϕ_2
ϕ_1
σ
μ_2
μ_1
2. Random ten-cell profiling
Frequency
Gene expression
3. Maximize likelihood of pdf f(x)
# measurements
$\text{Max} \sum_{i=1} \log f(x_i \mid \mu_1, \mu_2, F, \sigma)$
$f(x) = \sum_{j=0}^{10} \binom{10}{j} F^j (1-F)^{10-j} z_{j,10-j}(x)$
$z_{j,10-j}(x)$ is pdf of $Z_1 + \dots + Z_{10}$, where
$Z_c \overset{ind}{\sim} \begin{cases} LN(\mu_1, \sigma^2) & 1 \le c \le j \\ LN(\mu_2, \sigma^2) & j < c \le 10 \end{cases}$
4. Estimate parameters
Model fit
Confidence interval
F = 25%
$\mu_1 = 0.5$
σ = 0.2
$\mu_2 = -2.5$
B
Inferred μ_1
Known μ_1
Inferred μ_2
True μ_2 = –2.5
Inferred F (%)
True F = 22.5%
Inferred σ
True σ = 0.225
C
True μ_1 = 0.5
Known μ_2
Inferred μ
D
Known F (%)
2.5 12.5 22.5 32.5 42.5
E
Known σ
0.025 0.225 0.425 0.625 0.825
F
SOD2 stochastic profiling
Frequency
Model fit
Relative expression
F = 23% [13%, 33%]
μ_1 = – 1.1 [–1.7, –0.4]
μ_2 = – 4.1 [–4.4, –3.8]
σ = 0.7 [0.5, 1.1]
G
High
Rel. expression
Low
SOD2
SOD2 DRAQ5
H
Computational F
SOD2 F (%)
Measured

Figure 4.3 Inferring cellular subpopulations by maximum-likelihood inference of stochastic ten-cell samples from an LN–LN mixture of regulatory states.

(*A*) The maximum-likelihood approach involves four steps: 1) A model of heterogeneous gene regulation is posed, where single cells are assumed to express genes at a low or high level with a common coefficient of variation for both subpopulations. The weight of each subpopulation is defined by the integrated single-cell expression distribution of the subpopulation (f_1 and f_2). The four parameters of the model are the log-mean expression for each subpopulation (μ_1, μ_2), the proportion of cells in the high subpopulation (F), and the common log-standard deviation of expression (σ). 2) Random 10-cell samples are collected to build a distribution of measurements for inference by the model. 3) Based on the model in Step 1, a likelihood function is derived (see Chapter 4 Methods). 4) The likelihood function is then maximized by searching through the four parameters of the model to identify those that are most likely given the experimental observations. Additionally, we obtain measures of confidence for each estimated parameter (gray). (*B–E*) Accurate prediction of single-cell parameters from simulated ten-cell samples. Ten-cell expression data were simulated using different values of (*B*) μ_1, (*C*) μ_2, (*D*) F, and (*E*) σ (gray solid) and then estimated by maximum likelihood. For each group of simulations, the remaining three model parameters were kept fixed at (*C–E*) $\mu_1 = 0.5$, (*B* and *D–E*) $\mu_2 = -2.5$, (*B–C* and *E*) $F = 22.5\%$, and (*B–D*) $\sigma = 0.225$ (gray dashed). Solid gray line shows the one-to-one mapping of inferred-to-known parameter value. Off-diagonal plots are categorical plots of the fixed parameter estimates for a given value of the perturbed parameter. Graphs show the parameter estimates together with 95% maximum-likelihood confidence intervals from three independent sets of 50 ten-cell samples. (*F–H*) Prediction and validation of expression frequency for the heterogeneous transcript, *SOD2*, during breast-epithelial acinar morphogenesis. (*F*) Distribution of 81 ten-cell qPCR measurements of *SOD2* in outer ECM-attached epithelial cells and estimated subpopulation distribution (red line). Maximum-likelihood parameters (red box) are shown with 95% CI in brackets. (*G*) Representative RNA FISH image of endogenous *SOD2* expression. A pseudocolored image (left) is shown alongside a two-color image with DRAQ5 counterstain to visualize nuclei (right). Arrows indicate ECM-attached cells with high *SOD2* expression. Scale bar = 20 µm. (*H*) Percentage of cells showing high expression of *SOD2* by RNA FISH (gray bar) compared with the maximum-likelihood estimate of *F* (white dashed). RNA FISH data are shown as the mean percentage ± 95% CI of ECM-attached cells showing high expression of *SOD2*. Maximum-likelihood predictions are shown as the parameter point estimate (white) ± 95% CI (red).

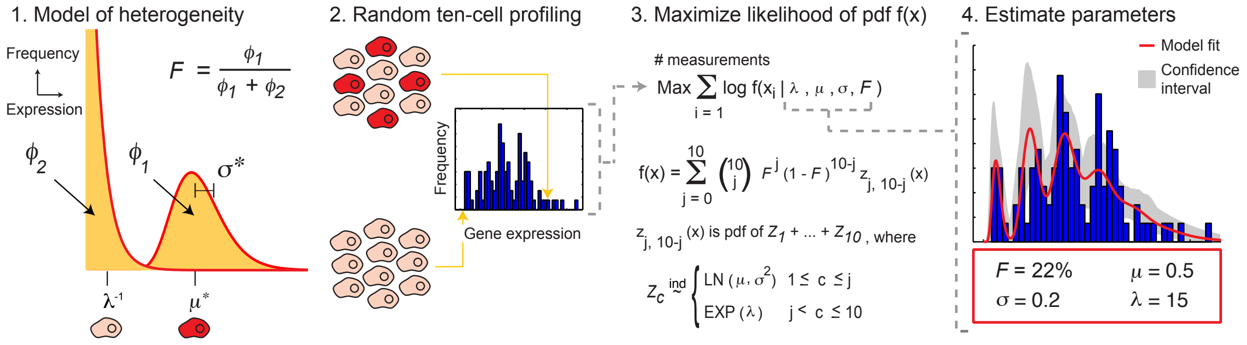

1. Model of heterogeneity
Frequency
Expression
$F = \frac{\phi_1}{\phi_1 + \phi_2}$
ϕ_2
ϕ_1
σ^*
λ^{-1}
μ^*
2. Random ten-cell profiling
Frequency
Gene expression
3. Maximize likelihood of pdf f(x)
$\text{Max} \sum_{i=1}^{\#\text{ measurements}} \log f(x_i \mid \lambda, \mu, \sigma, F)$
$f(x) = \sum_{j=0}^{10} \binom{10}{j} F^j (1-F)^{10-j} z_{j,10-j}(x)$
$z_{j,10-j}(x)$ is pdf of $Z_1 + \ldots + Z_{10}$, where
$Z_c \overset{\text{ind}}{\sim} \begin{cases} LN(\mu, \sigma^2) & 1 \leq c \leq j \\ EXP(\lambda) & j < c \leq 10 \end{cases}$
4. Estimate parameters
Model fit
Confidence interval
F = 22%
μ = 0.5
σ = 0.2
λ = 15

Figure 4.4 Inferring mixtures of a near-silen, exponential regulatory state and lognormal regulatory state.

The inverse-modeling approach involves four steps similar to Figure 4.3*A*: 1) A model of heterogeneous gene regulation is posed, where single cells are assumed to express genes at an absent-to-low or high level. The weight of each subpopulation is defined by the integrated single-cell expression distribution of the subpopulation (ϕ_1 and ϕ_2). The four parameters of the model are the rate parameter of the basal subpopulation (λ, where λ^{-1} is the mean expression), the mean of the second subpopulation ($\mu^* = \exp\left(\mu + \frac{\sigma^2}{2}\right)$), the standard deviation of the second subpopulation ($\sigma^* = \sqrt{\left[\exp\left(\sigma^2\right) - 1\right] \cdot \exp\left(\sigma^2 + 2\mu\right)}$), and the proportion of cells in the high subpopulation (F). 2) Random 10-cell samples are collected to build a distribution of measurements for inference by the model. 3) Based on the model in Step 1, a likelihood function is derived (Chapter 4 Methods). 4) The likelihood function is then maximized by searching through the four parameters of the model to identify those that are most likely given the experimental observations. Additionally, we obtain measures of confidence for each estimated parameter (gray).

4.3.3 Theoretical and experimental validation of maximum-likelihood inference

We evaluated the performance of our approach by using computational simulations of 10-cell samples with known distribution parameters. First, it was important to identify the minimum number of random samples needed to ensure accurate parameter estimation. Given hundreds to thousands of samples, we found that robust and accurate estimates were obtained for all model parameters irrespective of the mixture type (Figure 4.5*A*, *B*). Conversely, with very few samples (~20 or fewer), the convolved distributions were incompletely populated and our resulting estimates were highly uncertain and sometimes inaccurate for the LN–LN and EXP–LN mixtures. The transition between the two regimes occurred at 50–100 samples, which we defined as the approximate number of data points required for effective maximum-likelihood inference of single transcripts.

We next used simulations to identify the parameter ranges where maximum-likelihood inference makes accurate estimates of each regulatory state. Starting with the LN–LN mixture, we perturbed μ_1, μ_2, σ, or F individually while keeping the other three parameters fixed and simulated 50 random 10-cell samples. For a wide range of subpopulation log-means (μ_1, μ_2), maximum-likelihood inference accurately inferred model parameters with negligible bias (Figure 4.3*B*, *C*). We also observed good performance when altering the expression frequency (F). Accuracy declined near F = 50%, when the two subpopulations offset one another and disguise as a distribution with large σ (Figure 4.3*D*). Nevertheless, the estimation procedure still accurately and confidently captured ~70% of the total parameter space (F = 0–35% over the range of 0–50%). For the log-standard deviation (σ), performance declined only when this parameter was extremely high (Figure 4.3*E*). Parameter estimates were accurate until σ reached ~0.8, corresponding to a ~95% CV that is higher than nearly all genes examined thus far (191, 192). None of the mixture parameters could be reliably inferred from higher-order moments of the 10-cell distributions, although low F or high σ correlated with a slight increase in skewness (Figure 4.6). These results indicated that maximum-likelihood inference could extract parameters that were otherwise inaccessible by descriptive statistics.

We repeated the simulations for the EXP–LN mixture and arrived at very similar conclusions. As long as λ and μ were large enough to prevent overlap of the two regulatory states, we found that parameter estimates were accurate, although the variance of inferred σ was somewhat higher than in the LN–LN mixture (Figure 4.7). Together, these simulations suggested that maximum-likelihood inference is able to deconvolve a wide range of regulatory heterogeneities from 10-cell samples.

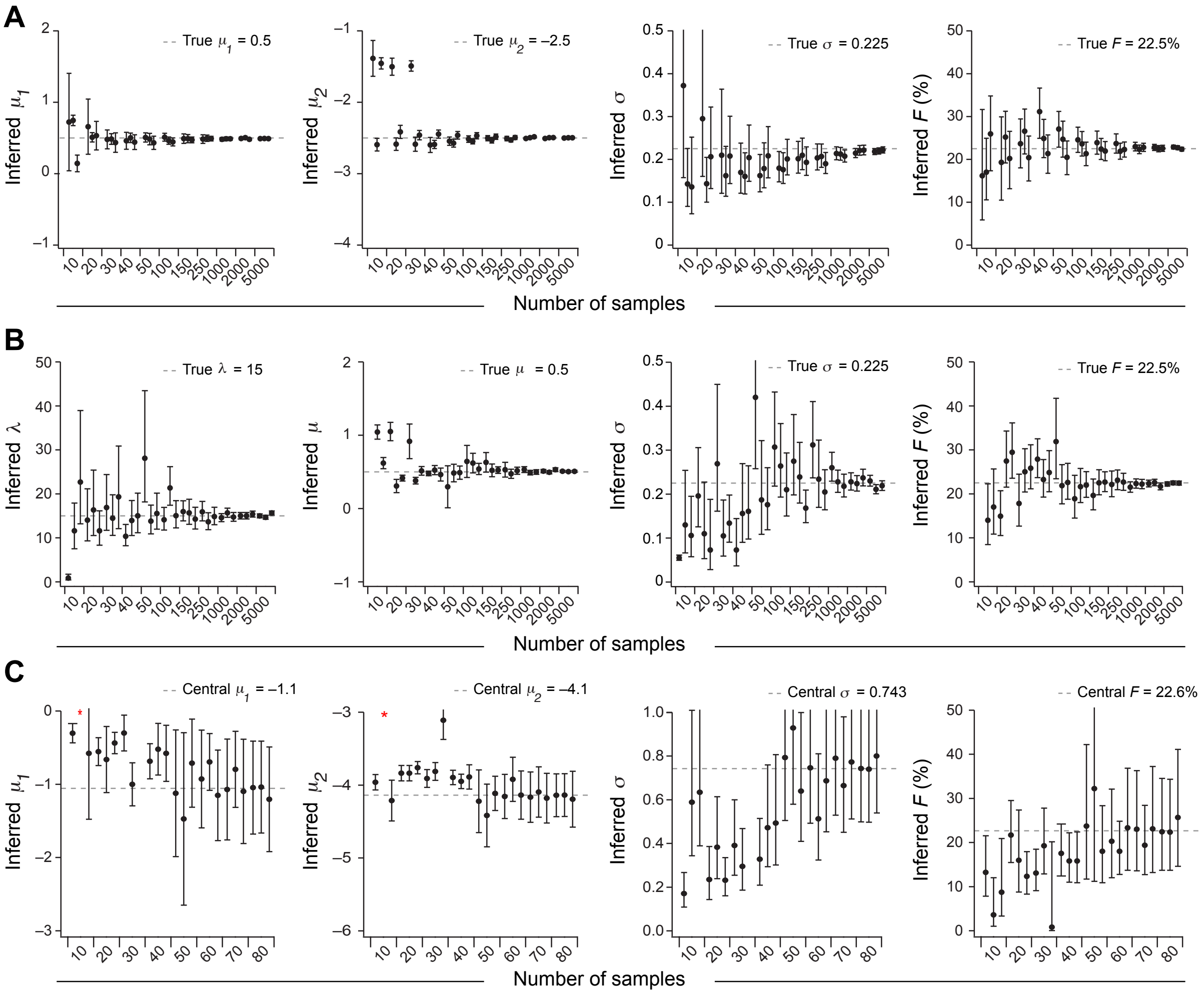


Figure 4.5 Parameter accuracy and confidence stabilizes with 50–100 random 10-cell samples. Ten-cell expression data were computational simulated at fixed parameters: μ_1 = 0.5, μ_2 = –2.5, F = 22.5%, and σ = 0.225, gray dashed for the LN–LN mixture (*A*); λ = 15, μ_1 = 0.5, F = 22.5%, and σ = 0.225, gray dashed for the EXP–LN mixture (*B*), and the number of 10-cell samples varied from 10 to 5000. (*C*) Experimental 10-cell measurements of *SOD2* from Figure 4.3*F* were resampled over the indicated range. Gray dashed lined shows the central maximum-likelihood estimate with the full dataset. Red asterisks indicate estimations that fell outside the indicated parameter range. Note the decrease in parameter accuracy and confidence when the sample size is less than 50 samples. Graphs show the central parameter estimate ± 95% CI from three independent sets of the indicated number of 10-cell samples.

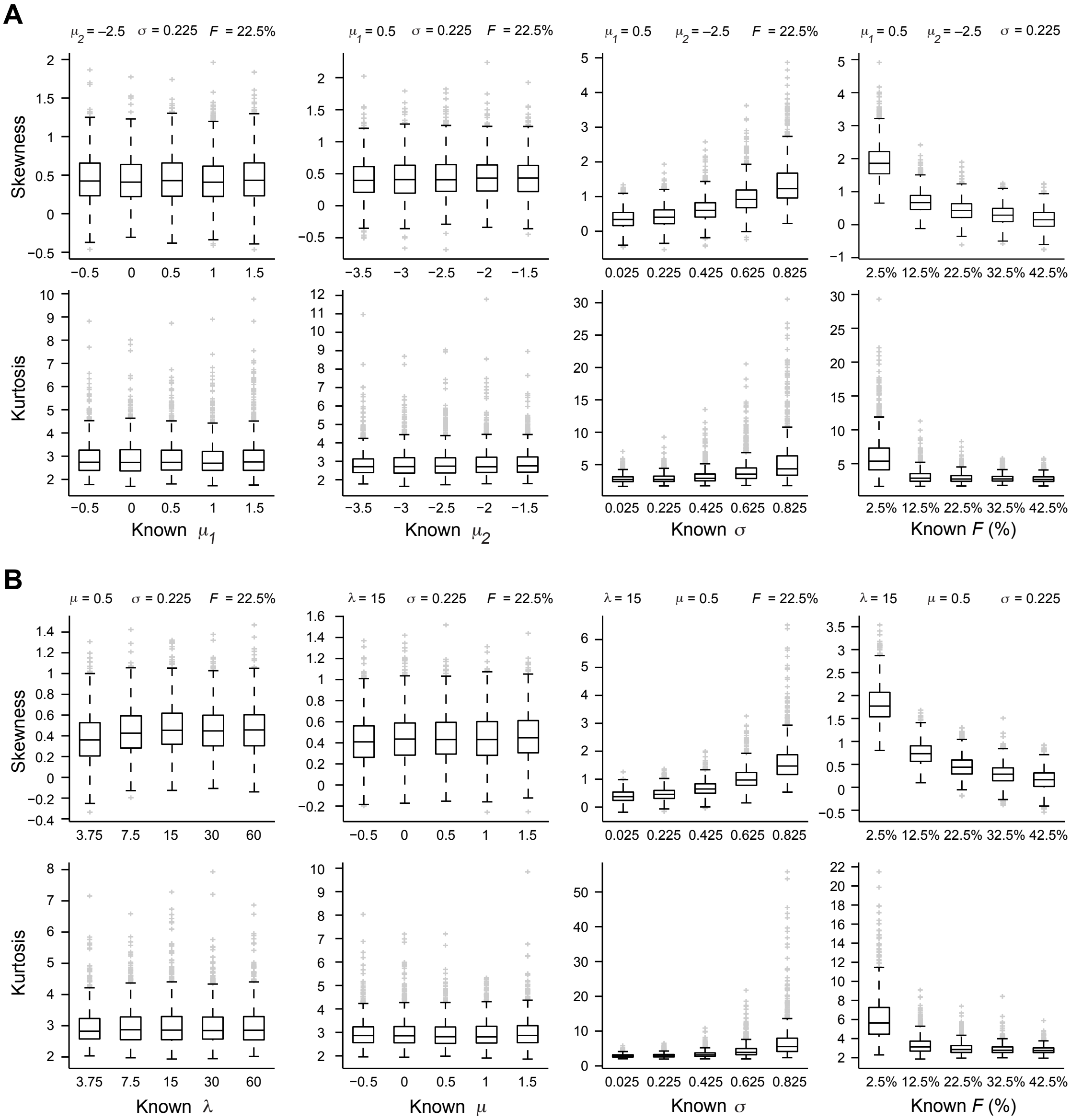


Figure 4.6 Skewness and kurtosis are not strongly predictive of subpopulation parameters. (*A,B*) Stochastic profiles were computationally simulated (see Chapter 4 Methods) with three fixed parameters (top) and one varying parameter (x-axis) for either the LN–LN mixture with 50 samples (*A*) or the EXP–LN mixture with 100 samples (*B*). The skewness and kurtosis were calculated for each simulated distribution. Box plot shows the distribution of 1000 independent computational simulations. Qualitatively similar results were obtained for the EXP–LN mixture with 50 samples.

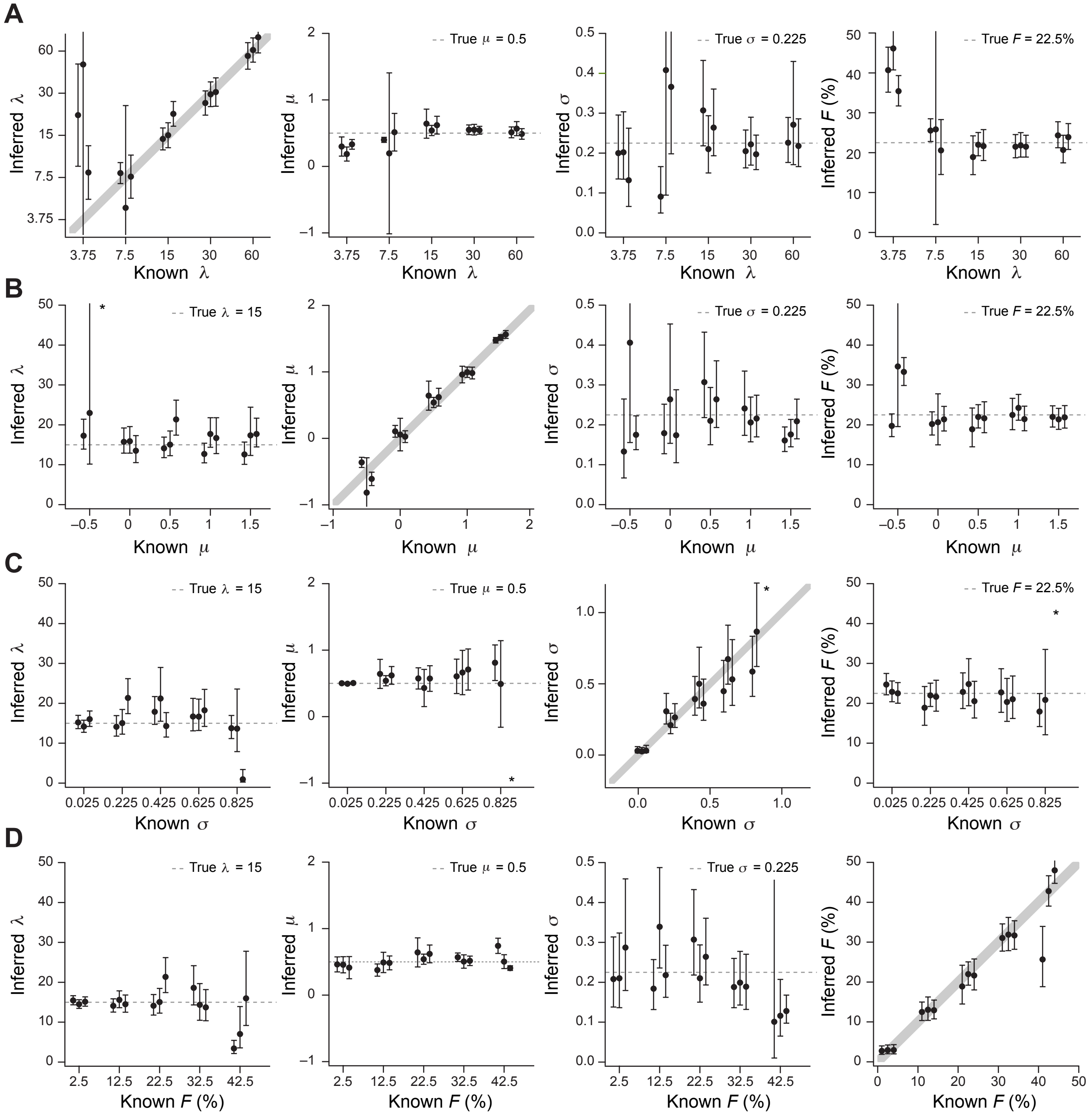


Figure 4.7 Accurate prediction of single-cell parameters from simulated ten-cell samples of an EXP–LN mixture of regulatory states.

(*A–D*) Ten-cell expression data were simulated using different values of λ (*A*), μ (*B*), σ (*C*), and F (*D*) (gray solid) and then estimated by maximum likelihood. For each group of simulations, the remaining three model parameters were kept fixed at $\lambda = 15$ (*B–D*), $\mu = 0.5$ (*A, C–D*), $\sigma = 0.225$ (*A–B, D*), and $F = 22.5\%$ (*A-C*) (gray dashed). Solid gray line shows the one-to-one mapping of inferred-to-known parameter value. Off-diagonal plots are categorical plots of the fixed parameter estimates for a given value of the perturbed parameter. Graphs show the parameter estimates together with 95% maximum-likelihood confidence intervals from three independent sets of 100 ten-cell samples. Asterisks indicate inferences that fell outside of the axis range. Qualitatively similar results were obtained for three independent sets of 50 ten-cell samples.

To examine the accuracy of maximum-likelihood inference with real 10-cell samples, we focused on expression of the detoxifying enzyme, *SOD2*, during breast-epithelial acinar morphogenesis. We used a culture model in which immortalized human breast epithelial cells are grown as single-cell clones in reconstituted basement-membrane extracellular matrix (ECM) to form 3D organotypic spheroids (97, 98). Earlier stochastic-profiling studies of developing spheroids had suggested that there were two *SOD2* regulatory states among the ECM-attached cells (73, 100). To apply maximum-likelihood inference, we deeply sampled *SOD2* expression by qPCR in 81 random samples of 10 ECM-attached cells (left; Figure 4.3*F*). Using these data, we maximized the likelihood of the LN–LN and EXP–LN estimators, as well as that of a relaxed LN–LN estimator, which allowed each regulatory state to have its own log-standard deviation (σ_1 and σ_2). The three inferences were compared by using the Bayesian Information Criterion (BIC) score to calculate the quality of the fit relative to the number of inferred parameters (Table 1). The best overall inference was the mixture model that parameterized two distinct regulatory states with the lowest BIC score.

For the 10-cell measurements of *SOD2*, we found that the LN–LN mixture was slightly preferred over the EXP–LN mixture (right; Figure 4.3*F* and Table 1). The inability to discriminate clearly between these two models was likely caused by the basal regulatory state, which could be described as an exponential distribution ($\lambda = 46$) or a lognormal distribution with a very small log-mean ($\mu_2 = -4.1$) given the sampling data. Regardless, the two models predicted similar *SOD2* expression frequencies among ECM-attached cells: 23% (13–33%) for the LN–LN mixture vs. 19% (12–27%) for the EXP–LN mixture. To determine the accuracy of this shared prediction, we directly measured *F* in 3D spheroids by RNA fluorescence in situ hybridization (RNA FISH) (Figure 4.3*G*). Scoring individual cells with high *SOD2* fluorescence intensity, we calculated an expression frequency of ~26%. This measurement closely agreed with the inferred parameter of the LN–LN mixture (the better scoring model; Figure 4.3*H* and Table 1) and lay within the estimated confidence interval of the EXP–LN mixture. By resampling the 10-cell *SOD2* data, we found that at-least 50 observations were required to arrive at an accurate result (Figure 4.5*C*), confirming our earlier estimates using simulated data (Figure 4.5*A*, *B*). The *SOD2* parameterization suggested that maximum-likelihood inference could correctly extract single-cell information from 10-cell sampling data.

4.3.4 Maximum-likelihood inference of coordinated stochastic transcriptional profiles

Programs of gene expression are often controlled by common upstream factors that enforce the regulatory state. We reasoned that coordinated single-cell gene programs would be the product of an overarching regulatory heterogeneity characterized by a shared *F*. If true, then it should be possible to estimate the expression frequency more confidently and with fewer samples by extending maximum-likelihood inference to gene clusters with coordinated 10-cell fluctuations.

We extended the approach as follows (Figure 4.8*A*). First, each gene within the cluster was assigned its own μ_1 and μ_2 (or μ and λ for the EXP–LN mixture) to account for gene-to-gene differences in expression level and detection sensitivity. Next, we assumed that the genes within a cluster share a common *F* and σ (or *F*, σ_1, and σ_2 in the relaxed LN–LN mixture), implying a shared mechanism of regulation (100, 193). Therefore, each mixture model of a cluster of *g* genes involved $2g + 2$ or $2g + 3$ parameters. Even for small gene programs ($g \leq 10$), this parameter search space was too large for non-convex optimization methods to maximize the global

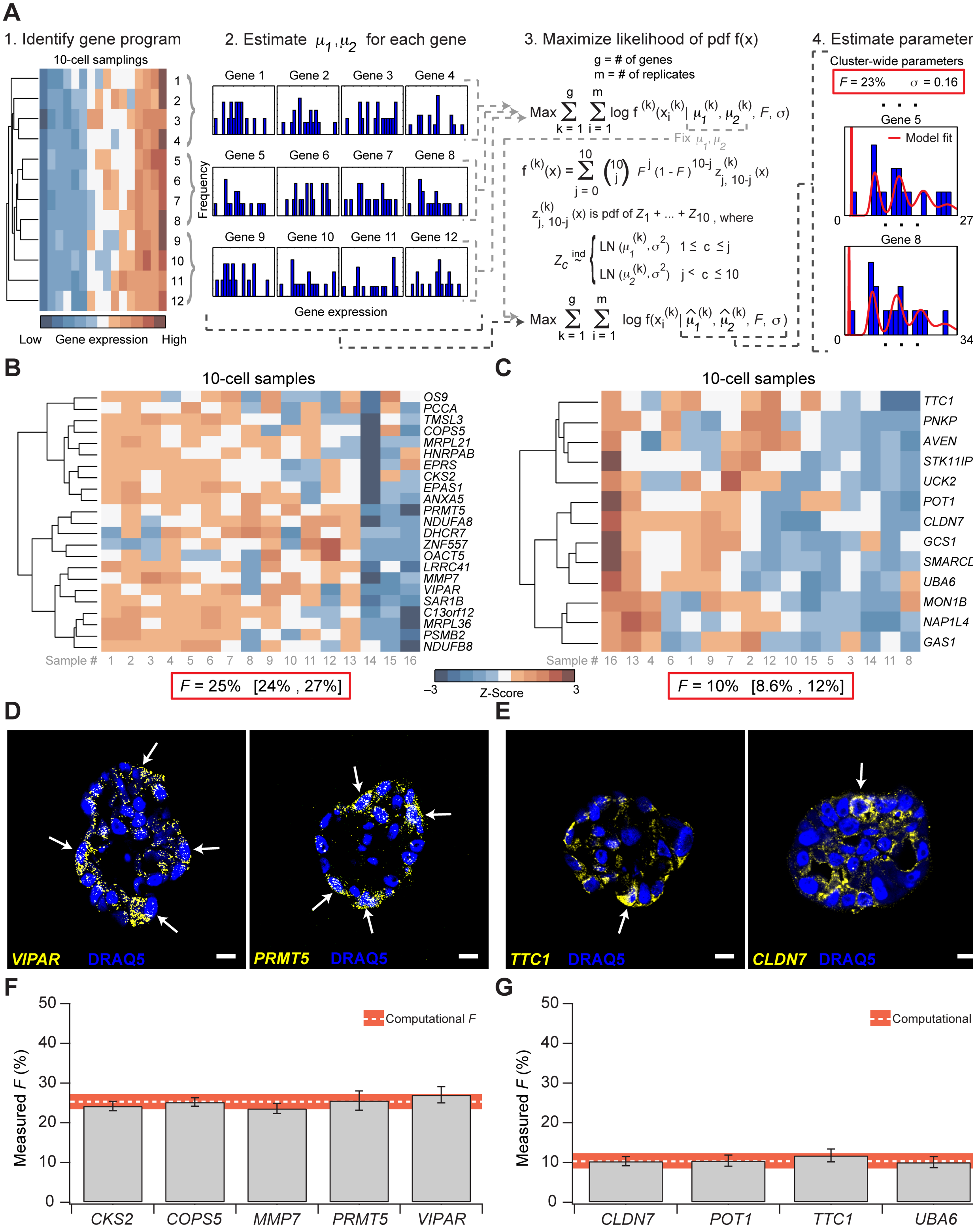

A
1. Identify gene program
10-cell samplings
1
2
3
4
5
6
7
8
9
10
11
12
Low
Gene expression
High
2. Estimate μ_1, μ_2 for each gene
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
Gene 7
Gene 8
Gene 9
Gene 10
Gene 11
Gene 12
Frequency
Gene expression
3. Maximize likelihood of pdf f(x)
g = # of genes
m = # of replicates
$\mathrm{Max} \sum_{k=1}^{g} \sum_{i=1}^{m} \log f^{(k)}(x_i^{(k)} | \mu_1^{(k)}, \mu_2^{(k)}, F, \sigma)$
Fix μ_1, μ_2
$f^{(k)}(x) = \sum_{j=0}^{10} \binom{10}{j} F^j (1-F)^{10-j} z_{j,\,10-j}^{(k)}(x)$
$z_{j,\,10-j}^{(k)}(x)$ is pdf of $Z_1 + \dots + Z_{10}$, where
$Z_c \overset{ind}{\sim} \begin{cases} LN(\mu_1^{(k)}, \sigma^2) & 1 \le c \le j \\ LN(\mu_2^{(k)}, \sigma^2) & j < c \le 10 \end{cases}$
$\mathrm{Max} \sum_{k=1}^{g} \sum_{i=1}^{m} \log f(x_i^{(k)} | \hat{\mu}_1^{(k)}, \hat{\mu}_2^{(k)}, F, \sigma)$
4. Estimate parameter
Cluster-wide parameters
F = 23%   σ = 0.16
Gene 5
Model fit
0
27
Gene 8
0
34
B
10-cell samples
OS9
PCCA
TMSL3
COPS5
MRPL21
HNRPAB
EPRS
CKS2
EPAS1
ANXA5
PRMT5
NDUFA8
DHCR7
ZNF557
OACT5
LRRC41
MMP7
VIPAR
SAR1B
C13orf12
MRPL36
PSMB2
NDUFB8
Sample # 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
F = 25%   [24% , 27%]
C
10-cell samples
TTC1
PNKP
AVEN
STK11IP
UCK2
POT1
CLDN7
GCS1
SMARCD
UBA6
MON1B
NAP1L4
GAS1
Sample # 16 13 4 6 1 9 7 2 12 10 15 5 3 14 11 8
F = 10%   [8.6% , 12%]
–3
Z-Score
3
D
VIPAR
DRAQ5
PRMT5
DRAQ5
E
TTC1
DRAQ5
CLDN7
DRAQ5
F
Measured F (%)
50
40
30
20
10
0
Computational F
CKS2
COPS5
MMP7
PRMT5
VIPAR
G
Measured F (%)
50
40
30
20
10
0
Computational
CLDN7
POT1
TTC1
UBA6

Figure 4.8 Maximum-likelihood inference accurately estimates subpopulation frequencies from 10-cell gene-expression clusters.

(*A*) The maximum-likelihood approach was modified for gene clusters with coordinated 10-cell sampling fluctuations as follows: 1) Global gene measurements are grouped and assumed to share a common *F* and *σ*. 2) An expression cluster of interest is divided into four-gene subsets for the first round of parameter estimation of μ_1 and μ_2 for each gene in the subset. 3) A maximum-likelihood estimator is derived based on an expanded version of the model in Figure 4.3*A*, where each gene *k* in a group of genes, *1, ..., g*, has its own $\mu_1^{(k)}$ and $\mu_2^{(k)}$. The likelihood function is maximized to infer $\mu_1^{(k)}$ and $\mu_2^{(k)}$ locally. 4) The likelihood function is then re-maximized for the entire dataset keeping the log-mean estimates ($\hat{\mu}_1^{(k)}$ and $\hat{\mu}_2^{(k)}$) fixed to provide clusterwide estimates of *F* and *σ*. Note that each gene has a different range of gene expression to reflect differences in overall expression levels, which are captured in the model predictions as well. (*B–G*) Prediction and validation of expression frequency for heterogeneously expressed gene programs during breast-epithelial acinar morphogenesis. (*B* and *C*) Heatmap of clustered 10-cell transcriptional profiles (73). Gray labels indicate the 10-cell sample numbers. Maximum-likelihood estimate of expression frequency (red box) is shown with 95% CI in brackets for each cluster. Note that the two gene clusters are predicted to have substantially different frequencies of high expression based on their 10-cell sampling fluctuations. (*D* and *E*) Representative RNA FISH images of transcripts from (*D*) the infrequent cluster and (*E*) the rare cluster. Images are shown with DRAQ5 counterstain to visualize nuclei. Arrows show ECM-attached cells with high expression. Scale bar = 10 μm. (*F* and *G*) Percentage of cells showing high expression by RNA FISH (gray bar) of a subset of genes in each cluster compared with the maximum-likelihood estimate of *F* (white dashed). RNA FISH data are shown as the mean percentage ± 95% CI of ECM-attached cells showing high expression. Maximum-likelihood predictions are shown as the parameter point estimate (white) ± 95% CI (red).

likelihood function quickly (see Chapter 4 Methods). To increase the speed and efficiency of estimation, the cluster was broken down into smaller four-gene subgroups to infer log-means for each gene in the subgroup together with local estimates of *F*, σ, and λ (Steps 1–2; Figure 4.8*A*). After log-means were locally estimated, the remaining parameters were globally inferred by re-maximizing the likelihood function for the entire gene cluster while retaining the local gene-specific estimates of μ_1 and μ_2 (LN–LN mixture) or μ (EXP–LN mixture) (Steps 3–4; Figure 4.8*A*). As before, selection of the LN–LN, relaxed LN–LN, and EXP–LN mixture model was made according to the lowest BIC score (Table 1). This revised formulation of maximum-likelihood inference enabled accurate and confident estimates of the expression frequency while requiring only ~1/3 of the sample size (Figure 4.9).

We tested our extension of maximum-likelihood inference by extracting from an earlier study two coexpression clusters that were completely uncharacterized (73) (Figure 4.10). These clusters contained one- to two-dozen genes with strongly coordinated expression fluctuations across 16 samples of 10 ECM-attached cells, but the patterns of fluctuation for each cluster were markedly different (Figure 4.8*B,C*). Accordingly, when we inferred the parameters for the two clusters, the model predicted two very different expression frequencies. The first “infrequent” gene cluster was predicted to be upregulated in ~25% of the ECM-attached population (Figure 4.8*B*). The LN–LN mixture model was preferred over the EXP–LN or relaxed LN–LN mixtures (Table 1), although all three models converged upon similar values for *F*. By contrast, the expression frequency of the “rare” second cluster was predicted to be ~10% by the LN–LN mixture (Figure 4.8*C*), which was the best scoring model of the three (Table 1). Our parameterization of the two clusters emphasizes the mosaicked regulatory states that evolve even in a very simple model of tissue architecture (73, 97, 135).

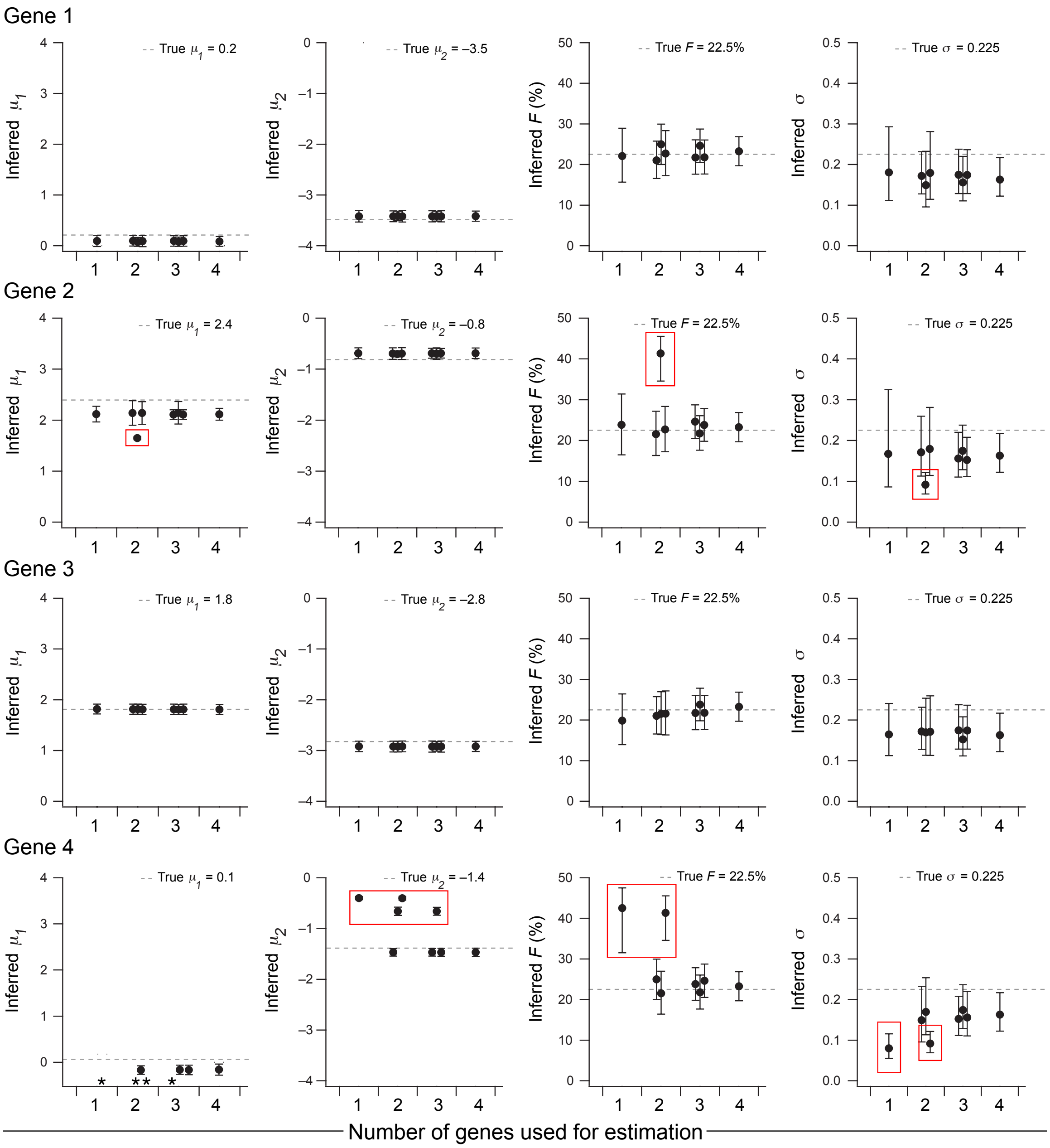


Figure 4.9 Parameter estimation with limited samples is improved by evaluating coexpression clusters.

Groups of stochastic 10-cell samples were computationally simulated with known parameters as in Figure 4.3 The estimation procedure was performed on a single gene, all pairwise combinations ($_{[4-1]}C_{[2-1]}$ = 3 combinations), all triplicate combinations ($_{[4-1]}C_{[3-1]}$ = 3 combinations), and all four genes together. Estimations with increasing number of genes use the previous iteration as a seed guess for optimization. Note that parameter predictions stabilize when 3–4 genes are considered simultaneously. Red boxes highlight poor estimations that improve as more genes are considered. Red asterisks indicate estimations that fell outside the indicated parameter range. Genes 9–12 correspond to genes 9–12 in the heatmap in Figure 4.8*A*.

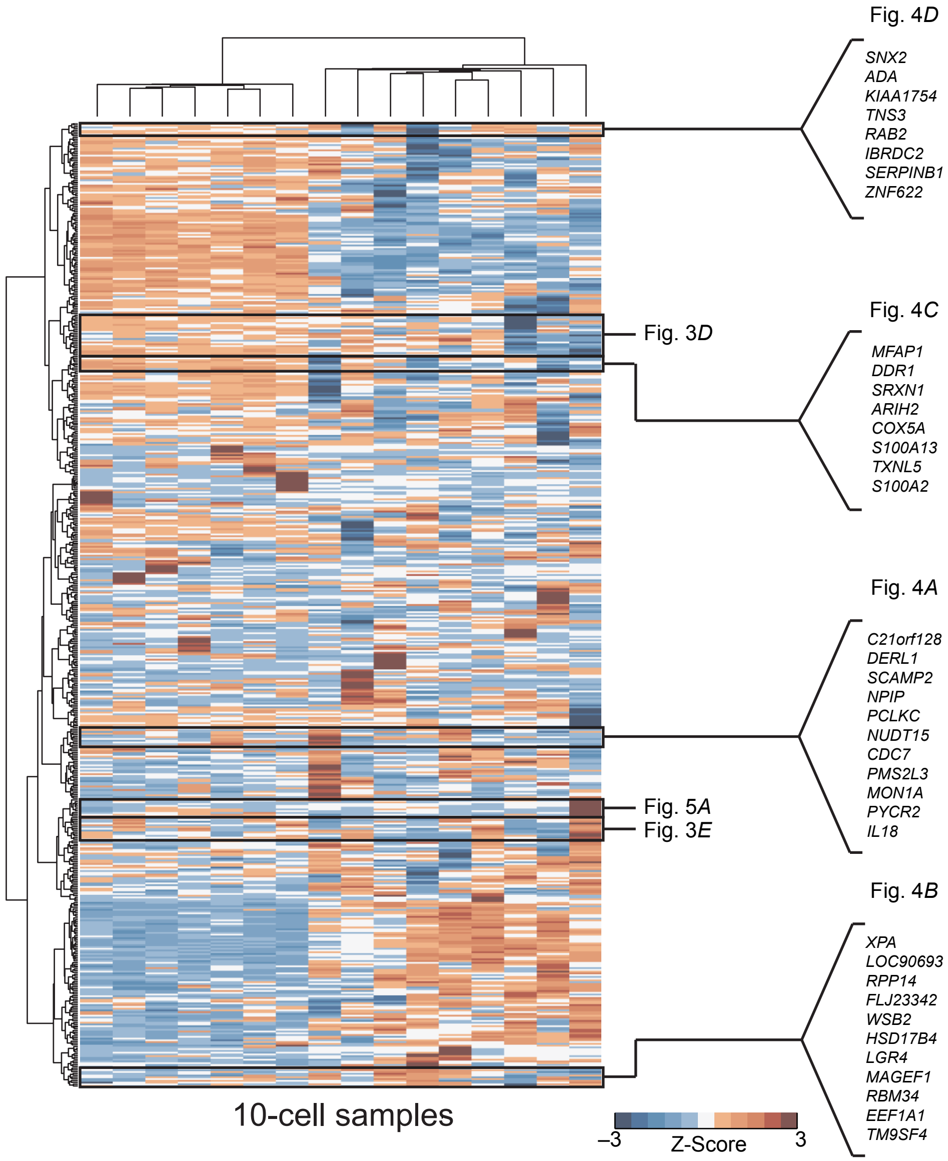


Figure 4.10 Heterogeneously coregulated transcriptional programs parameterized by maximum-likelihood inference.

Heatmap of clustered 10-cell transcriptional profiles identified by stochastic profiling (73). Profiles were used for the inferences in the indicated subpanels in the main figures. Individual genes are listed for the coregulated clusters used in Figure 4.13**.**

Table 1 Bayes information criterion scores for different maximum-likelihood inferences.

	Mixture model		
	LN–LN	Relaxed LN–LN	EXP–LN
SOD2	183	165[a]	184
Infrequent cluster	2286	2296	2373
Rare cluster	440	454	580
Very-rare cluster	–17	–37[b]	14
Fig. 4*A*	308[c]	311	515
Fig. 4*B*	780	1087	875
Fig. 4*C*	898	972	1090
Fig. 4*D*	910	922	967

Minimum scores indicating best fit are underlined. LN, lognormal; EXP, exponential

[a] Relaxed LN–LN was excluded here because the two log-means associated with this inference were less than twofold different from one another. A constrained optimization of this model (see Chapter 4 Methods) yielded F = 24.3% with BIC = 180.

[b] Constrained optimization of this model yielded F = 2.3% with BIC = –35.

[c] Constrained optimization of this model yielded F = 5.1% with BIC = 309.

To test if the predicted values of *F* were accurate within the coexpressed clusters, we designed and validated riboprobes for 4–5 genes in each cluster and quantified their frequency of high expression by RNA FISH (Figure 4.11*A, B*). We found that transcripts in the infrequent expression cluster were strongly expressed in 3–5 ECM-attached cells per acinus cross-section (Figure 4.8*D* and *F* and Figure 4.12*A*), yielding an average expression frequency of ~25%. Conversely, genes in the rare expression cluster (Figure 4.8*E, G* and Figure 4.12) were strongly expressed in 1–2 ECM-attached cells per acinus cross-section, consistent with an expression frequency of ~10%. The expression frequencies of both clusters closely agreed with the inferred *F* parameters, suggesting that our extended inference approach was effective and accurate.

We evaluated the estimates of expression frequency more broadly by selecting four additional clusters from the same dataset for parameterization (Figure 4.10) (73). The clusters showed distinct fluctuation patterns and consequently led to *F* inferences that ranged from less than 5% to greater than 25% (Figure 4.13*A–D*, upper). We validated riboprobes for representative gene in each cluster and scored the expression frequency (Figure 4.13*A–D*, lower and Figure 4.11*C*). Together with the earlier clusters, we observed a strong correlation between the expression frequency inferred computationally and the manual counts obtained by RNA FISH (R = 0.89, Figure 4.13*E*). The accuracy of the manual counts was further confirmed by correlation with an expression-frequency index derived from digital image analysis of segmented acini (Figure 4.14 and Chapter 4 Methods). Taken together, these data indicate that maximum-likelihood inference accurately infers single-cell expression frequencies from cluster-wide patterns of 10-cell fluctuations.

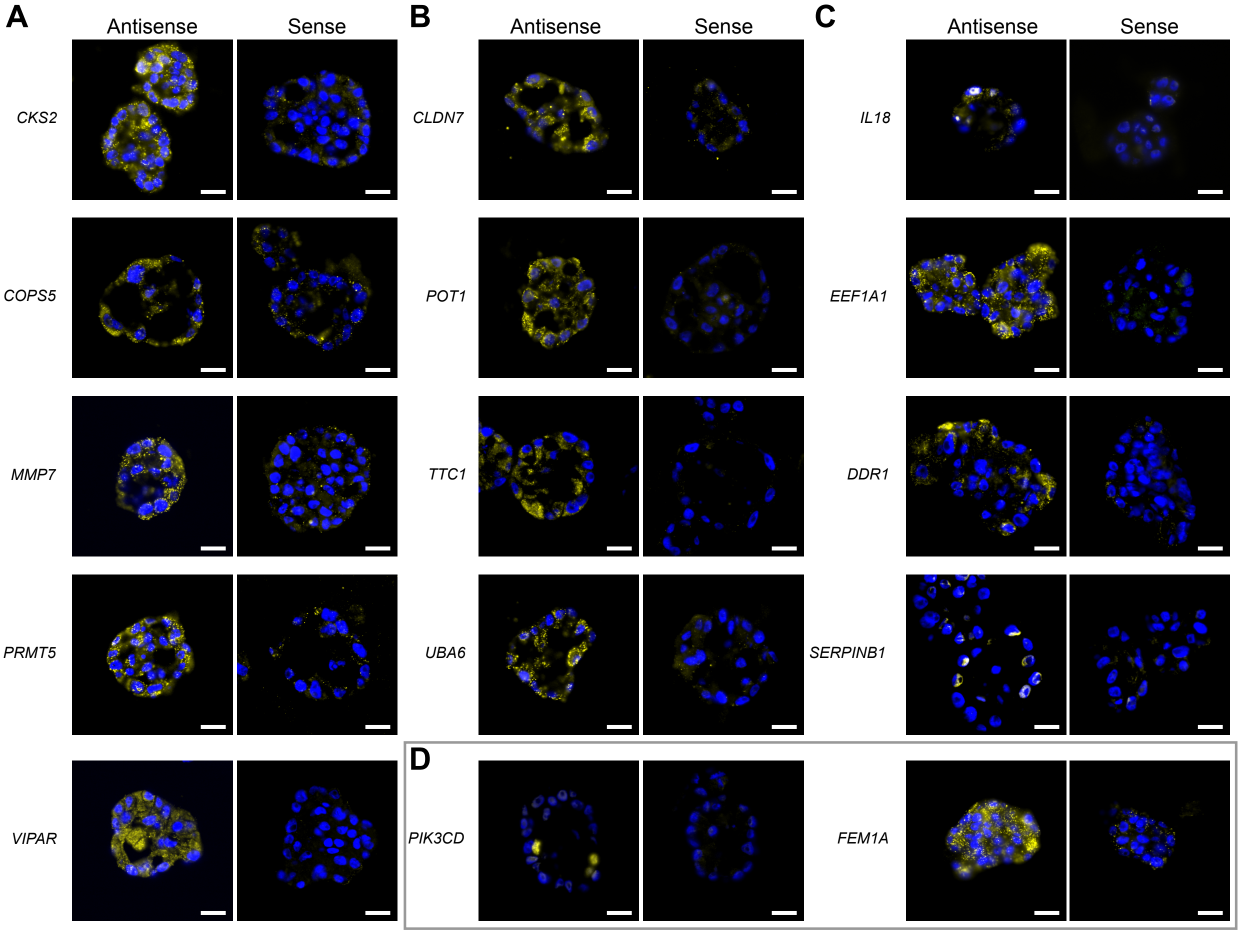


Figure 4.11 Fluorescence intensities for RNA FISH are specific to endogenous transcripts. Riboprobes against each of the indicated genes were hybridized to MCF10A-5E day 10 frozen sections of 3D acini and imaged by widefield immunofluorescence (see Methods). For each gene, the antisense and sense control are shown in identically hybridized samples. The exposure time and image scaling is matched for each pair of antisense-sense samples to compare specific (antisense) and background (sense) signal strength. Probe validations are for genes in the infrequent cluster (*A*), the rare cluster (*B*), assorted clusters in Figure 4.13 (*C*), and the very-rare cluster (*D*). Scale bar is 20 μm.

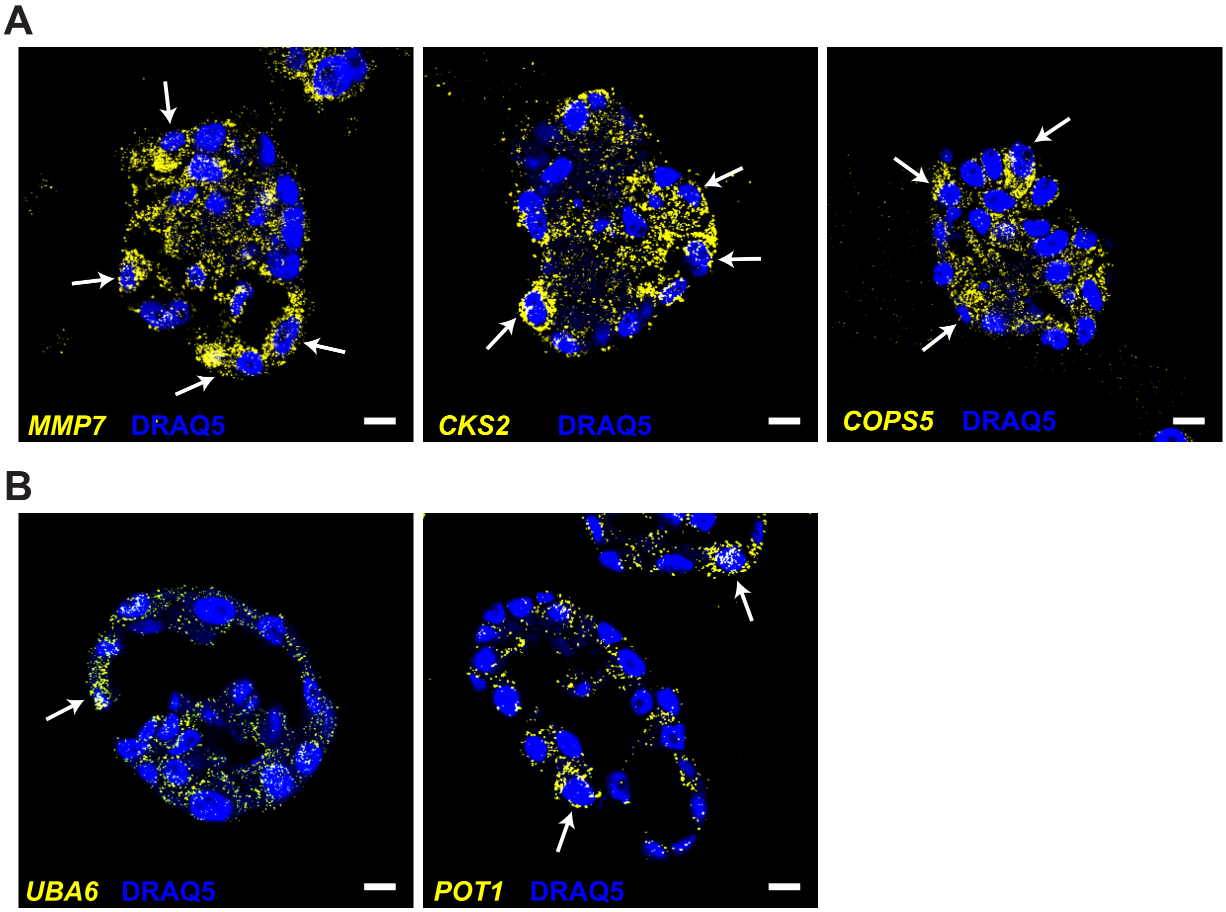


Figure 4.12 RNA FISH validation of gene expression frequencies.
Representative RNA FISH images of genes in the infrequent cluster (*A*) and the rare cluster (*B*) at day 10 of MCF10A-5E 3D morphogenesis. Images are shown alongside DRAQ5 counterstain to visualize nuclei. Arrows show ECM-attached cells with high expression. Scale bar = 10 µm.

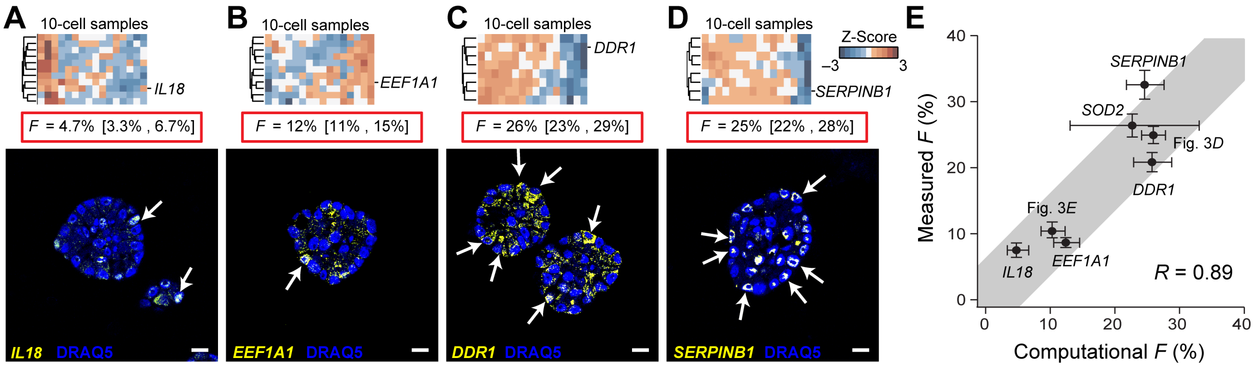


Figure 4.13 Widespread parameterization of single-cell expression frequency by maximum-likelihood inference.

(*A*–*D*, upper) Clusters of 10-cell expression fluctuations among ECM-attached cells (73). A complete list of transcripts in each cluster is shown in Figure 4.10. Maximum-likelihood estimate of expression frequency (red box) is shown with 95% CI in brackets for each cluster. (*A*–*D*, lower) RNA FISH images of a representative transcript in each cluster. Images are shown with DRAQ5 counterstain to visualize nuclei. Arrows show ECM-attached cells with high expression. Scale bar = 10 µm. (*E*) Percentage of cells scored for high expression by RNA FISH compared with the maximum-likelihood estimate of *F*. RNA FISH data are shown as the mean percentage ± 95% CI of ECM-attached cells showing high expression. Maximum-likelihood predictions are shown as the parameter point estimate ± 95% CI. The gray bar shows a one-to-one correspondence with 5% measurement tolerance. Estimates for *SOD2* are reprinted from Figure 4.3*H*. Estimates for Figure 4.8, *D* and *E* were calculated by pooling all scored transcripts within each cluster. Pearson correlation (*R*) between measured and inferred expression frequencies is shown.

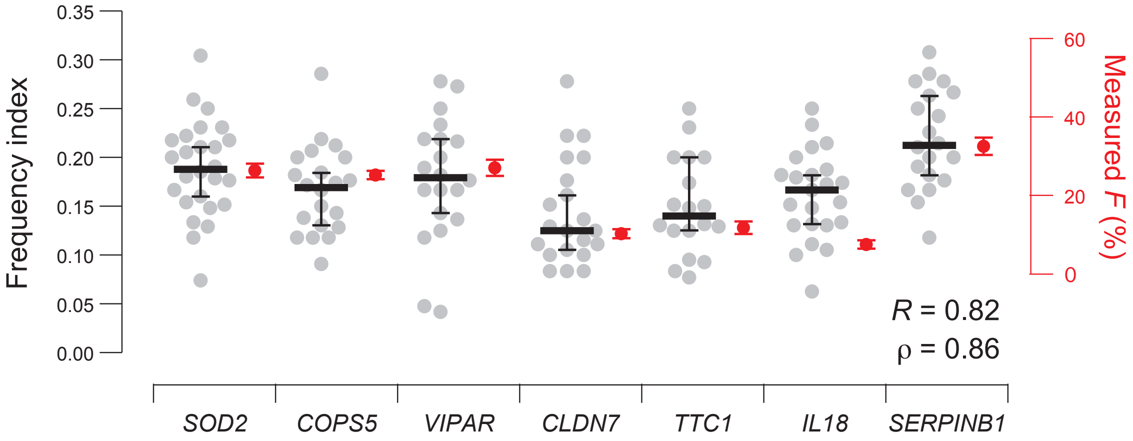


Figure 4.14 Manually scored expression frequencies correlate with a digitally quantified frequency index.

A minimum of 18 image fields of view (10+ cells per field of view) were manually segmented and analyzed digitally (see Chapter 4 Methods). The median frequency index (left axis) with 95% nonparametric confidence interval is displayed in black with individual replicates shown in gray. The manually scored frequencies (right axis) are reprinted in red from Figure 4.3*H*, Figure 4.8*G*, Figure 4.8*H*, Figure 4.13*A*, and Figure 4.13*D*. Pearson (*R*) and Spearman (ρ) correlations are shown.

4.3.5 Identification of a peculiar, very-rare transcriptional regulatory state

Maximum-likelihood inference provides critical information about the state distribution and expression frequency of any gene cluster identified by stochastic profiling to be heterogeneously regulated. As a proof-of-concept application, we screened gene clusters from the 3D profiling data (73) to identify unusual regulatory states that warranted follow-up study. One cluster was notable among those surveyed because the predicted expression frequency of the high regulatory state was very rare (F = 2.3%). The "very-rare" cluster was also distinguished by its strong concordance with the relaxed LN–LN mixture compared to the alternative mixture models (Table 1). Moreover, the log-mean of the low regulatory state was extremely low ($\mu_2 \sim -3.3$), suggesting that the cluster was at or below detection in the population. Within this coexpression cluster, we recognized several genes that were strongly associated with breast cancer, including the breast cancer susceptibility gene *BRIP1* (alternatively called *FANCJ* or *BACH1* (194)), the breast cancer associated gene *IRF2* (195), and the zinc-finger gene *HIVEP2*, which is frequently downregulated or mutated in breast cancer (41, 196) (Figure 4.15*A*). We speculated that genes within the cluster were tightly regulated so that they could be activated in a restricted cellular context where their expression was critical.

Among the genes in the very-rare cluster, we were most intrigued by the phosphoinositide-3-kinase (PI3K) isoform *PIK3CD* (alternatively called *p110δ*). 3D breast epithelial cultures abundantly express two other PI3K isoforms, *PIK3CA* and *PIK3CB* (Figure 4.15*B* and Figure 4.16), and it is generally thought that any PI3K isoform can support proliferation and survival (197). Nevertheless, we found that the low-copy expression of *PIK3CD* was transcriptionally upregulated with delayed kinetics compared to the other PI3K isoforms (Figure 4.15*B*), suggesting a unique regulatory mechanism. When *PIK3CD* abundance was visualized in single cells by RNA FISH, we observed a striking pattern. Most cells lacked *PIK3CD* or expressed it at very-low levels; however, we consistently identified a sporadic subpopulation of cells (roughly 1–2 cells every other acinus cross-section) with high *PIK3CD* expression (Figure 4.15*C* and Figure 4.11*D*). The overall frequency of cells in the *PIK3CD*$^{\mathrm{hi}}$ state was somewhat higher than the maximum-likelihood inferences of *F* for the cluster, but the inferred frequency agreed with the very-rare expression of two other members of the cluster, *FEM1A* and IRF2 (Figure 4.17). Together, these observations pointed to an acute (and likely transient) regulatory event triggering the selective induction of cluster genes in single ECM-attached cells.

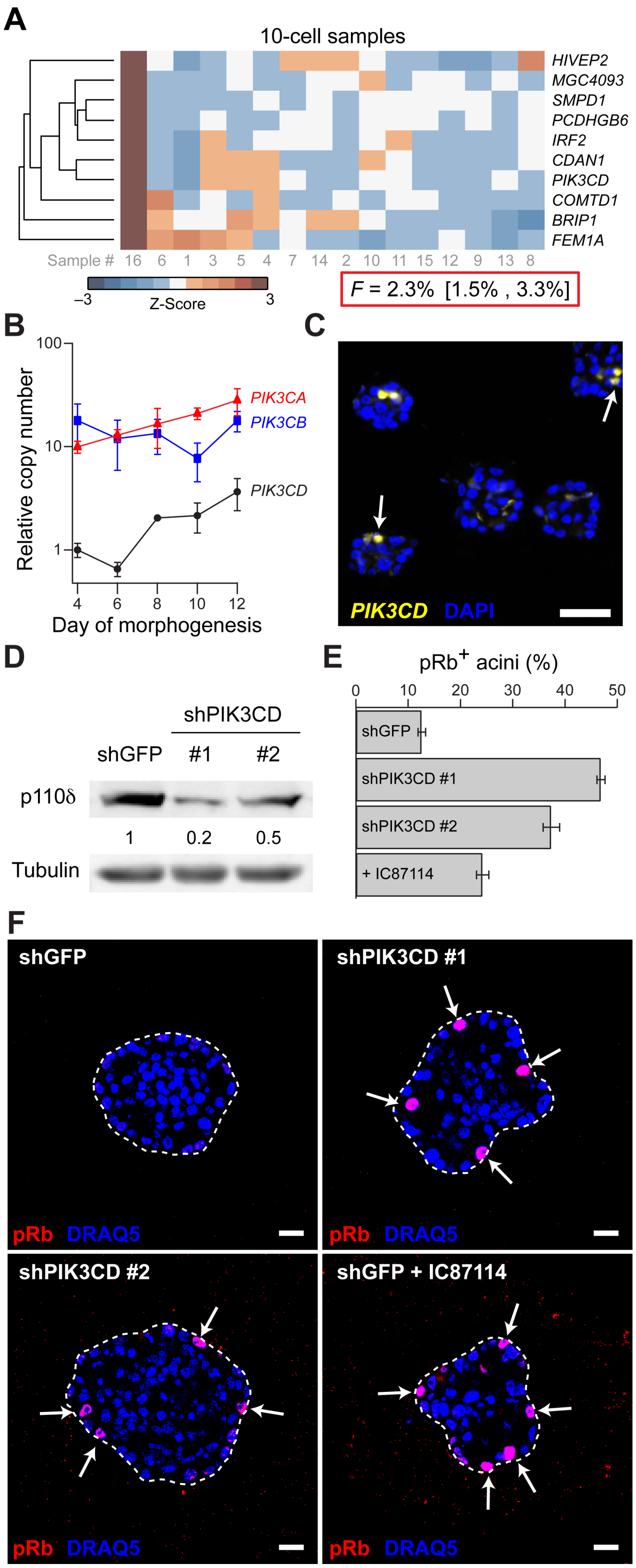

A
10-cell samples
HIVEP2
MGC4093
SMPD1
PCDHGB6
IRF2
CDAN1
PIK3CD
COMTD1
BRIP1
FEM1A
Sample # 16 6 1 3 5 4 7 14 2 10 11 15 12 9 13 8
–3
Z-Score
3
F = 2.3% [1.5% , 3.3%]
B
100
10
1
Relative copy number
4 6 8 10 12
Day of morphogenesis
PIK3CA
PIK3CB
PIK3CD
C
PIK3CD DAPI
D
shPIK3CD
shGFP #1 #2
p110δ
1 0.2 0.5
Tubulin
E
pRb+ acini (%)
0 10 20 30 40 50
shGFP
shPIK3CD #1
shPIK3CD #2
+ IC87114
F
shGFP
shPIK3CD #1
shPIK3CD #2
shGFP + IC87114
pRb DRAQ5
pRb DRAQ5
pRb DRAQ5
pRb DRAQ5

Figure 4.15 A unique, very-rare regulatory state is marked by *PIK3CD*, which is important for normal suppression of proliferation during breast epithelial acinar morphogenesis.

(*A*) Heatmap of clustered 10-cell transcriptional profiles (73). Gray labels indicate the 10-cell sample numbers. Maximum-likelihood estimate of expression frequency (red box) is shown with 95% CI in brackets. (*B*) *PIK3CD* expression is upregulated during 3D morphogenesis. Relative *PIK3CA* (red), *PIK3CB* (blue), and *PIK3CD* (black) expression was measured by qPCR at various time points during 3D morphogenesis. Data are shown as mean expression ± s.e.m. normalized to the day 4 expression of *PIK3CD* of three independent experiments. *PIK3CG* was not expressed in MCF10A-5E cells (Figure 4.16). (*C*) Representative RNA FISH image of *PIK3CD* expression is shown with DRAQ5 counterstain to visualize nuclei. Arrow shows ECM-attached cells with high expression of *PIK3CD*. Scale bar = 20 µm. (*D*) Knockdown of p110δ by shRNA. MCF10A-5E cells were infected with either shGFP (lane 1) or with one of two shRNA sequences targeting p110δ (lanes 2 and 3). Lysates were analyzed by immunoblotting with tubulin used as the loading control. Densitometry of p110δ abundance is shown relative to the shGFP control. (*E, F*) Disruption of normal *PIK3CD* regulation elicits a hyperproliferative phenotype in 3D culture. shGFP, shPIK3CD #1, and shPIK3CD #2 cells or shGFP cells + 20 µM p110δ inhibitor IC87114 were fixed at day 15 of 3D morphogenesis, stained for pRb (red), and analyzed by confocal immunofluorescence. Cells were counterstained with DRAQ5 (blue) to label nuclei. Arrows in *F* highlight pRb-positive cells. Scale bar = 20 µm. Quantification of proliferating acini in each condition is shown in *E* as the mean ± s.e.m. of eight independent experiments.

We next asked whether *PIK3CD* was specifically important for normal acinar morphogenesis. To eliminate the very-rare *PIK3CD*^{hi} subpopulation, we perturbed p110δ by two independent methods: RNA interference and the p110δ-specific small-molecule inhibitor, IC87114 (ref. (198); Figure 4.15*D* and Figure 4.18). When shPIK3CD cells were placed in 3D culture, we found that acini were larger and distorted, suggesting a defect in proliferation arrest. Using phosphorylated Rb (pRb) as a proliferative marker, we observed that shPIK3CD acini were still cycling after fifteen days of 3D culture when shGFP control acini had quiesced (Figure 4.15*E, F*). Furthermore, when control cells were cultured with IC87114, we observed sustained proliferation that phenocopied *PIK3CD* knockdown (Figure 4.15*E, F*). These data together indicate that p110δ activity stemming from the very-rare *PIK3CD*^{hi} regulatory state is critical for normal proliferation arrest of breast epithelia in 3D culture. More broadly, our results with the very-rare cluster illustrate how maximum-likelihood inference can be used to hone in on gene programs with an expression frequency or regulatory pattern of interest.

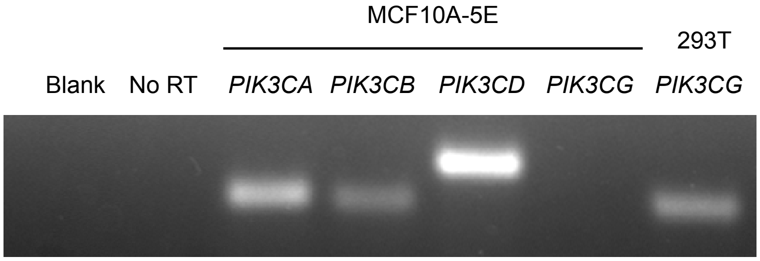


Figure 4.16 *PIK3CG* is not expressed in MCF10A-5E cells. qPCR amplicons for the indicated primer sets were run on an agarose gel.

qPCR amplicons for the indicated primer sets were run on an agarose gel. *PIK3CA, PIK3CB,* and *PIK3CD* were detected in MCF10A-5E cells, while *PIK3CG* was not. *PIK3CG* in 293T cells was used as a positive control for *PIK3CG* amplification. Representative blank and no RT samples are shown in the left two lanes.

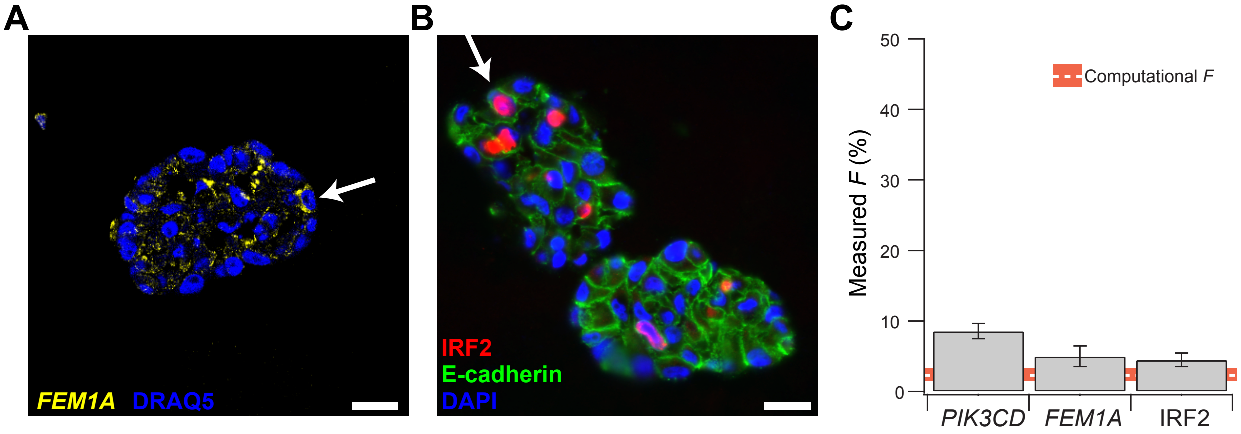


Figure 4.17 Extended validation in ECM-attached cells of a very-rare regulatory state. (*A*) Representative image of *FEM1A* (yellow) shown together with DRAQ5 counterstain (blue) to visualize nuclei. The arrow shows an ECM-attached cell with expression of *FEM1A*. (*B*) Representative image of IRF2 protein (red), whose transcript was predicted to have very-rare expression. Image is shown together with E-cadherin (green) to visualize cell membranes and DAPI counterstain (blue) to visualize nuclei. The arrow shows an ECM-attached cell with expression of IRF2. (*C*) Percentage of cells showing high expression of *PIK3CD* transcript, *FEM1A* transcript, or IRF2 protein (gray bar) compared with the maximum-likelihood estimate of *F* (white dashed) for the very-rare expressed cluster. Data are shown as the mean percentage ± 95% CI of ECM-attached cells showing high expression. Inverse-modeling predictions are shown as the parameter point estimate (white) ± 95% CI (red). Scale bar is 20 μm.

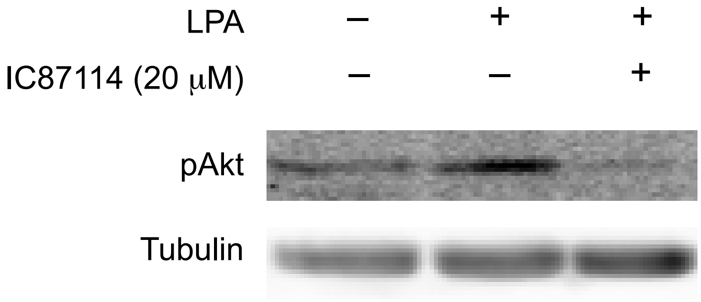


Figure 4.18 IC87114 inhibits Akt phosphorylation stimulated by p110δ.

MCF10A-5E cells were serum starved overnight and stimulated with the p110δ agonist lysophosphatidic acid (LPA, 10 μM for 5 min) in the presence or absence of 20 μM IC87114 (198).

4.3.6 Comparison with alternative deconvolution methods

We compared the performance of maximum-likelihood inference to other computational approaches for deconvolving mixed populations (199-201). The alternative methods invoked different mathematical formalisms—Bayesian statistics (199), nonnegative matrix factorization (200), and principal component analysis (201)—and none had previously been applied to transcriptional profiles of small samples. Using the sampling fluctuations within the infrequent, rare, and very-rare clusters, we attempted inferences of expression frequency and found that all were substantially less accurate than maximum-likelihood inference (Table 2). The comparison illustrates that our method is uniquely effective at parameterizing transcriptional regulatory states within cell populations.

Table 2 Expression frequency inferences from alternative deconvolution methods

	Stochastic-profiling cluster		
Method	Infrequent	Rare	Very Rare
Erkkila *et al*. (48)[a]	20%	~0%[b]	~0%[b]
Repsilber *et al*. (49)	22%	60%	25%
Tolliver *et al*. (50)	18, 40, 23, 19%[c]	59, 7.2, 11, 22%	30, 21, 19, 30%
Maximum-likelihood inference	25% [24%, 27%][d]	10% [8.6%, 12%]	2.3% [1.5%, 3.3%]
RNA FISH	25% [24%, 26%][d]	10% [9.4%, 12%]	5.6% [4.7%, 7.3%]

[a] Bayesian priors were set to 25%, 10%, and 5% for the infrequent, rare, and very-rare clusters, respectively.

[b] The estimated frequency was 2 × 10^{-12}%.

[c] A minimum of four subpopulations must be estimated with this deconvolution method.

[d] Bracket denotes 95% confidence interval.

4.3.7 Direct comparison of single-cell and ten-cell sampling strategies

Maximum-likelihood inference reconstructs the single-cell expression distribution without the need to measure single cells. Ignoring the technical challenges of global single-cell methods (72, 73, 89, 95), it should also be theoretically possible to recreate the complete expression distribution by measuring many individual cells. However, it was not clear whether single-cell profiling would be as effective as stochastic profiling when reconstructing from a limited number of one- or 10-cell samples. We anticipated that low expression frequencies would be particularly difficult for single-cell methods to characterize because of uncertainty in reliably capturing the rare regulatory state.

To compare single-cell profiling with stochastic profiling, we repeatedly simulated one- or 10-cell measurements of gene clusters with similar characteristics to those previously examined (Figure 4.8*F* and *G*, 5*A*, and Chapter 4 Methods). The three 12-gene clusters varied in their expression fraction—infrequent (*F* = 25%), rare (*F* = 10%), and very rare (*F* = 5%)—and the very-rare cluster was simulated as an LN–LN mixture or an EXP–LN mixture. When the number of observations was limited to 16 (as in the actual data), we found that maximum-likelihood inference provided superior estimates of *F* when using 10-cell groups (Table 3). Inferences from simulated observations of 16 single cells showed substantially higher mean squared error (MSE) for all gene clusters when compared to 16 10-cell observations. The larger MSE of one-cell inferences was caused by increases in both the bias and variance of the estimate, whose magnitudes depended on the cluster characteristics and mixture model. These computational simulations provide an upper bound on performance, because experimental error from actual single-cell experiments (19, 89) should blur the data much more. By collecting a greater total number of cells when observations are limited, maximum-likelihood inference of stochastic 10-cell profiles provides a more-accurate picture of the single-cell distribution than single-cell profiles.

Table 3 Expression frequency inferences from repeated observations of one vs. ten cells

True *F*	Mixture	Cells	Maximum-likelihood estimate of *F* MSE × 10^{-2}	 Bias × 10^{-2}	 Variance × 10^{-2}
25%	LN–LN	1	4.32	–20.76	0.01
		10	0.30	–3.40	0.19
10%	LN–LN	1	2.35	–2.83	2.27
		10	0.19	1.37	0.17
5%	LN–LN	1	19.73	29.09	11.27
		10	0.16	1.73	0.13
5%	EXP–LN	1	57.18	75.1	0.79
		10	4.50	0.80	1.77

MSE, bias, and variance were calculated across 100 simulations of 16 observations. *F* is defined from 0–100 × 10^{-2}. MSE = bias^2 + variance. MSE, mean squared error; LN, lognormal; EXP, exponential.

4.4 Discussion

Maximum-likelihood inference of mixed regulatory states enables accurate single-cell expression characteristics to be gleaned from 10-cell measurements. For individual genes, the model requires a large number of samples to obtain precise estimates, and its advantage over explicit single-cell methods is debatable. However, by extending the approach to coregulated gene clusters (73, 100, 193), we can infer expression frequencies much more robustly than single-cell methods when the extent of sampling is limited. In fact, after identifying heterogeneously regulated genes at the transcriptome-wide level by stochastic profiling (72, 73), global inferences are achievable with the same number of random 10-cell samples. Maximum-likelihood inference can thus be immediately incorporated into stochastic profiling studies that seek a further understanding of single-cell regulation (72, 100).

Multiple studies have demonstrated that heterogeneous phenotypes are primed by earlier regulatory non-uniformities in gene expression (22, 52, 106, 130, 131, 133, 174). But to date, these discoveries have relied on either predefined intracellular circuits or a mix of screening and serendipity. By combining stochastic profiling with maximum-likelihood inference, one can now examine the single-cell transcriptome for expression frequencies or other regulatory patterns that correlate with a downstream phenotype of interest. Such programs are most likely to contain one or more triggers of the heterogeneous phenotype. For example, our follow-on work with *PIK3CD* suggests that it may enforce a quiescent phenotype in a subpopulation of cells that would otherwise enter the cell cycle.

One day, it may be possible to measure the genome, transcriptome, and proteome accurately and cheaply in single cells. While progress is being made toward this goal (37, 89, 90, 175, 191), in the meantime it is valuable to develop alternative methods with less-stringent sample requirements. Our study shows that a surprising amount of quantitative single-cell information can be deconvolved mathematically from measurements with 10-fold more starting material. The average cell might indeed be a myth (74), but that does not mean that small-sample averages of cells cannot point to the truth.

Part II: Experimental analyses of heterogeneity

In this Part of the Dissertation, we discuss the experimental follow up to identify the biological function of growth-differentiation factor 11 (GDF11) heterogeneity. In Chapter 5, we begin with simple experiments to identify a function of GDF11, which we continue to explore down to the molecular mediators of the function. Additionally, we use xenograft experiments to extend these in vitro observations. Last, we use clinical specimens obtained from the University of Virginia hospital to show relevance of our in vitro and in vivo findings to the human disease. Chapter 5 serves as a case study on how to study a regulatory heterogeneity and identify biology importance to human disease and yield potential translational impact.

5 Chapter 5 – Growth-differentiation factor 11 (GDF11) tumor suppression is sequestered by basal-like breast cancer cells

5.1 Foreword

The previous Chapters focused on development of computational methods to uncover new importance to cellular heterogeneity. The methods make predictions or analyses about molecular or cellular phenotype, but these observations must be validated experimentally for the results to hold any biological traction. In this Chapter, we closely study the heterogeneous regulation of a TGFβ-family member ligand, growth differentiation factor 11 (GDF11). GDF11 was discovered in the original stochastic profiling dataset, but its function in breast epithelial biology was unknown. The tools and rationale used in this Chapter demonstrate how molecular heterogeneities in our 3D spheroid model can be propagated to understand human basal-like breast cancers. Excitingly, we uncover a peculiar mechanism of tumor suppressor silencing that is based on cell biology, not on genomic or epigenetic mechanisms.

5.2 Introduction

Breast cancer is a heterogeneous disease (57). Efforts cataloguing patient-to-patient variability have identified predominant subtypes of the disease (27). Of these subtypes, basal-like and claudin-low are molecularly a completely different disease than hormone- and receptor-positive breast cancers (61, 202). These subtypes tend to be negative for the three therapeutically relevant receptors. This triple-negative status gives the basal-like and claudin-low tumors a particularly poor prognosis (63).

Basal-like and claudin-low tumors also exhibit significant intratumoral variation between individual cells. Intratumor heterogeneity has been linked to poor prognosis, suggesting a role for heterogeneity in the progression of the disease (58). Uncovering the role of heterogeneity in tumor progression is limited by two factors: 1) the identification of heterogeneously regulated molecular markers (58) and 2) a scarcity of premalignant, triple-negative, basal-like ductal carcinoma in situ (DCIS) clinical specimens (203). We have developed a suite of tools that allows us to globally identify heterogeneously regulated genes, which we applied to breast epithelial morphogenesis (70, 73, 96). Breast epithelial spheroid morphogenesis shares similar molecular states that occur during tumor progression from lobular epithelial tissue to carcinomas (53).

We previously uncovered a set of heterogeneously regulated TGFβ-related genes in basal-like breast epithelial spheroids. This program of genes was highlighted by growth-differentiation factor 11 (*GDF11*), transforming growth factor beta receptor 3 (*TGFBR3*), and transforming growth factor beta induced (*TGFBI*). While roles of *TGFBR3* and *TGFBI* have been studies in the context of cancer (53, 142, 156, 204, 205), no role has been identified for *GDF11*. Previously implicated in development and regenerative medicine (206, 207), here, we discover a role for GDF11 as a tumor suppressor of basal-like breast cancer. GDF11 treatment improved normal breast epithelial morphogenesis and reduced invasion in a panel of basal-like breast cancer (BLBC) cell lines. We identify SMAD4 and ID2 as critical mediators for GDF11 to improve spheroid morphology, and demonstrate that the GDF11 pathway is inactive in a subset of triple-negative breast cancers (TNBC). GDF11 treatment inhibits intraductal growth of claudin-low breast cancer cells in vivo, suggesting a role for GDF11 in breast cancer progression. In human specimens of normal, ductal carcinoma in situ (DCIS), and TNBC, we discover that GDF11 displays different patterns of GDF11 immunoreactivity, which we link to a secretion defect of GDF11. Basal-like breast cancer cells may inactive tumor suppressors by sequestering them in advantageous cellular compartments.

5.3 Results

5.3.1 GDF11 is a heterogeneously regulated transcript important for 3D morphogenesis of immortalized breast epithelial cell lines

The stochastic profiling method identifies transcripts whose 10-cell sampling fluctuations fall far outside the range for normal biological variation; some transcripts could covary with a heterogeneous regulatory state but get overlooked by the analysis because of the stringent false-discovery rate required to screen the entire transcriptome (72, 73). To determine whether *GDF11*, *TGFBR3*, and *TGFBI* were indicative of a broader TGFβ-related transcriptional program, we searched the profiling dataset for the strongest covariates and identified thirteen genes strongly implicated in TGFβ-family signaling (Figure 5.1A). Notably, this gene set included the activin type-IIB receptor (*ACVR2B*) that, together with *GDF11*, comprised the only TGFβ-family ligand-receptor pair in the panel (Figure 5.1A) (208). We confirmed GDF11 expression heterogeneity at the transcript level by RNA FISH and at the protein level by immunofluorescence with two separate monoclonal antibodies. *GDF11* transcripts were abundant in ~20% of matrix-attached cells (Figure 5.1B), and both antibodies labeled with Golgi nonuniformly within MCF10A-5E 3D cultures (Figure 5.1C-D). Although the cleaner GDF11 antibody (“EPR”) also recognizes GDF8/MSTN (209), GDF11 is significantly more abundant in

51 of 53 breast cancer cell lines, as well as 97 or 97 basal-like and 121 of 122 triple-negative breast cancers in The Cancer Genome Atlas, suggesting specificity for GDF11 in breast tissue (Figure 5.1E) (27, 210, 211). Our observations raised the possibility of GDF11 triggered signaling heterogeneities during reorganization and stress adaptation of proliferating 3D breast-epithelial cultures.

In MCF10A-5E cells, knockdown of *TGFBR3* results in large, budding structures that are distinct from the proliferation-arrested spheroids normally observed during 3D culture (53). As a high-affinity coreceptor for many TGFβ-family ligands (140), we hypothesized that the shTGFBR3 phenotype was caused by a diminished sensitivity to endogenous GDF11, which was locally and heterogeneously produced. To abrogate the need for TGFBRIII, we treated the shTGFBR3 spheroids with saturating levels of GDF11. Remarkably, exogenous GDF11 completely blocked the larger, budding spheroids from forming (Figure 5.1E). The improved circularity in the control condition upon GDF11 treatment suggested that GDF11 treatment could also improve normal breast spheroid morphology. We treated MCF10A-5E and normal murine mammary gland (NMuMG (212, 213)) spheroids with exogenous GDF11 (Figure 5.1G-H). We first observed that GDF11 treatment reduced the size variation (Figure 5.1F) and increased lobular morphology in MCF10A-5E spheroids (Figure 5.1G). Additionally, the size variation of MCF10A-5E spheroids was increased upon knockdown of GDF11 with two shRNA hairpins (Figure 5.2). Structurally, the cells in the MCF10A-5E treated spheroids show a columnar, epithelial phenotype (Figure 5.1H). A significant increase in rounding was observed in the NMuMG cell line (Figure 5.1I), and structurally, individual cells were packed close together after GDF11 treatment (Figure 5.1J). These data suggest that GDF11 is an important molecule in breast epithelial morphogenesis, particularly in driving compacted, lobular architecture.

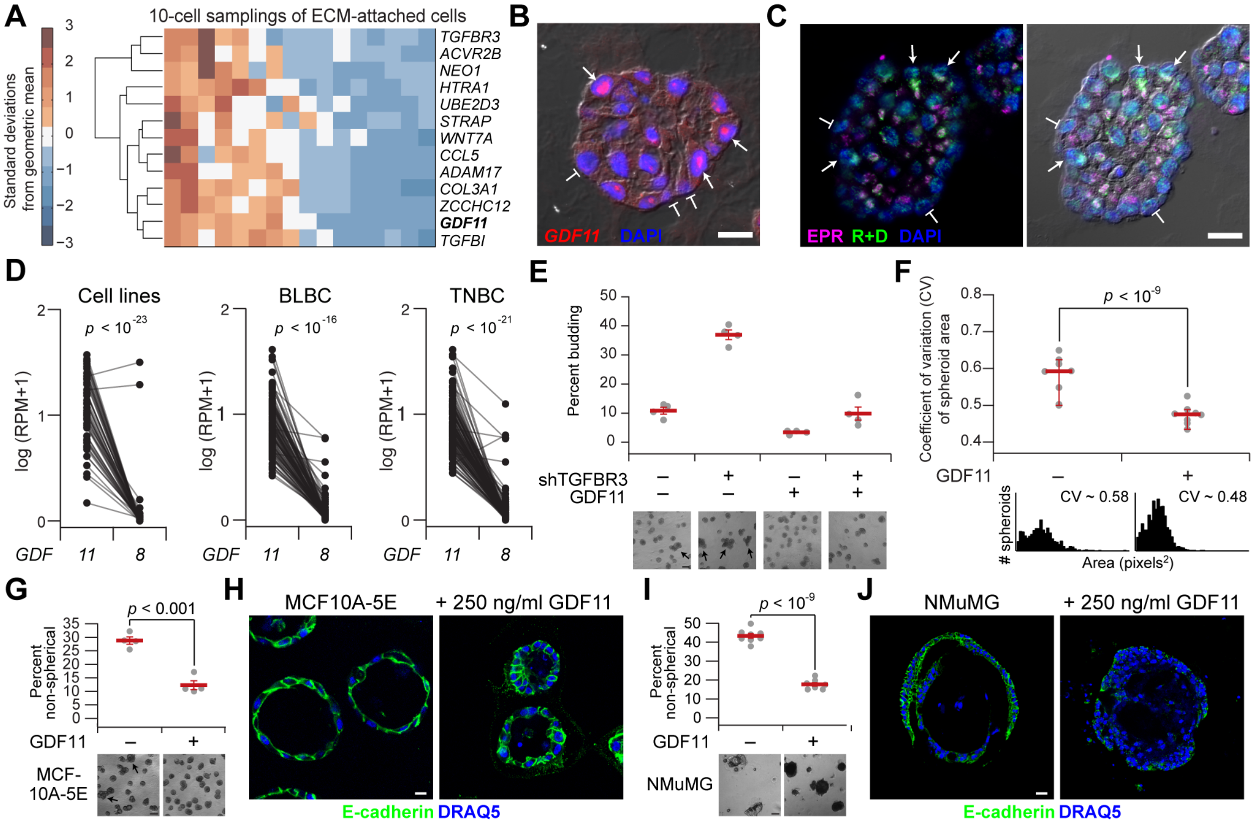


Figure 5.1 GDF11 is a heterogenously regulated transcript important to 3D morphogensis of immortalized breast epithelial cells.

(A) Stochastic profiling of ECM-attached MCF10A-5E cells in 3D culture. Transcripts that were highly correlated to *GDF11*, *TGFBI*, and *TGFBR3* were hierarchically clustered. (B) *GDF11* expression is heterogeneous amongst matrix-attached breast epithelia as measured by RNA FISH. (C) GDF11 expression is heterogeneous amongst matrix-attached breast epithelia as measured by two-color immunofluorescence with two monoclonal antibodies raised against GDF11. (D) RNA-seq data shows that *GDF11* is more abundant than *GDF8/MSTN* in breast cancer cell lines, human basal-like breast cancers, and human triple-negative breast cancers. (E) GDF11 reduces the percentage of budding structures induced by knockdown of *TGFBR3*. (F) Knockdown of GDF11 increases the variation in spheroid area. Lower plot shows the histogram of spheroid area and the median coefficient of variation. (G, I) GDF11 treatment promotes rounding in MCF10A-5E (G) and NMuMG (I) spheroids. (H) GDF11 causes a columnar arrangement in MCF10A-5E spheroids. (J) GDF11 treatment causes more condensed packing of cells in NMuMG spheroids. Data is mean ± s.e.m. for (E), (G), and (I). Data is median ± 95% confidence interval for (F). Scale bar is 200 µm for (E) , (G), (I), and 20 µm for (B), (C), (H), and (J).

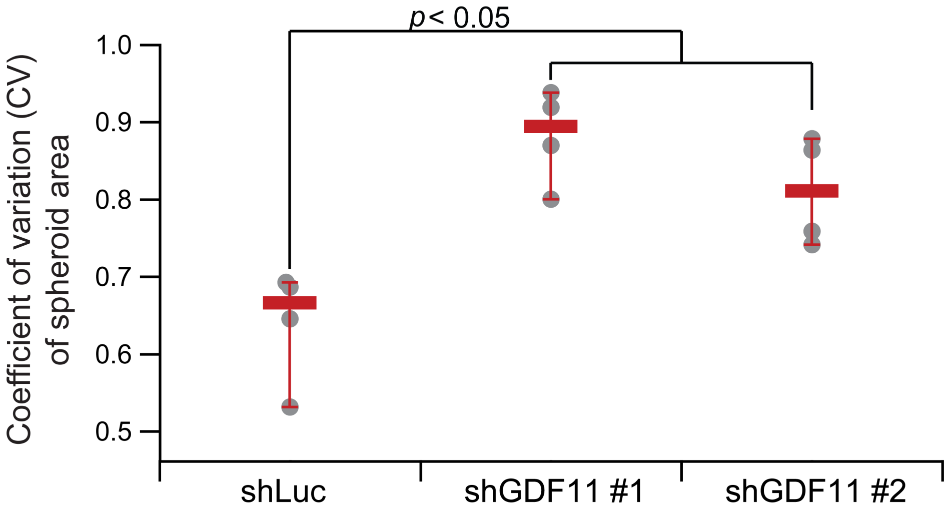


Figure 5.2 GDF11 knockdown increases size variation in MCF10A-5E spheroids.
Coefficient of variation (CV) in spheroid size was measured in quadruplicate for two GDF11 hairpins and hairpin control. Data is median ± 95% confidence interval.

5.3.2 GDF11 uniquely promotes compact tissue architecture of BLBC spheroids

We explored the possibility that GDF11 treatment could induce rounding and improve spheroid morphologies in transformed cells. We first tested this hypothesis by transforming a cell line we knew to be responsive to GDF11 (Figure 5.1E-H). We engineered a transformed version of our MCF10A-5E cell line by ectopically expressing oncogenic RasV12 (214) . In the absence of treatment, the RasV12 spheroids were highly invasive, even at an early time point of morphogenesis (Figure 5.3A). Treatment with GDF11 significantly increased the frequency of rounded spheroids. We next sought to test GDF11 treatment in basal-like breast cancer cell lines (BLBC). We assembled a panel of cell lines containing basal-like and claudin-low cell lines, spanning the two categories of BLBCs and 50% of the TNBC categories (65). Across many cell lines that could transduce GDF11, we observed a recurring pattern where GDF11 promoted a rounded, non-invasive phenotype (Figure 5.3B and Figure 5.4). For example, GDF11 treatment significantly reduces the invasion in the claudin-low cell line MDA-MB-231 (Figure 5.3B). Analogously, GDF11 treatment significantly reduced sheet-like protrusions in the basal-like cell line HCC1937 (Figure 5.4A). These data suggest that GDF11 reduces invasion and improves roundness in BLBC spheroids.

To test if the improvement in rounding and decrease in invasion induced by GDF11 was not due to a generic TGFβ response, we tested GDF11 against the closest and biologically relevant TGFβ−family member, TGFβ1 (215). We first measured extent of pathway activation by stimulating cells with saturating concentrations of GDF11 or TGFβ1. We used a canonical downstream target of TGFβR1, SMAD2, as a measure of pathway activity (215). Across two phosphosites, GDF11 was a weaker agonist of the pathway (Figure 5.3C-D). To exclude if GDF11 functioned as a TGFβ1 hypomorph, we tested lower doses of TGFβ1 to find a concentration of ligand that equally activated the pathway (Figure 5.3E). Importantly, we could not recapitulate the rounding phenotype with lower concentrations of TGFβ1, which instead induced a dissociated, EMT-like phenotype (Figure 5.3F) (216). The specificity of the rounding phenotype in response to GDF11 was also observed in the Hs578T cell line as well (Figure 5.4B). These data demonstrate that GDF11 is unique in its ability to promote lobular phenotypes in BLBC spheroids.

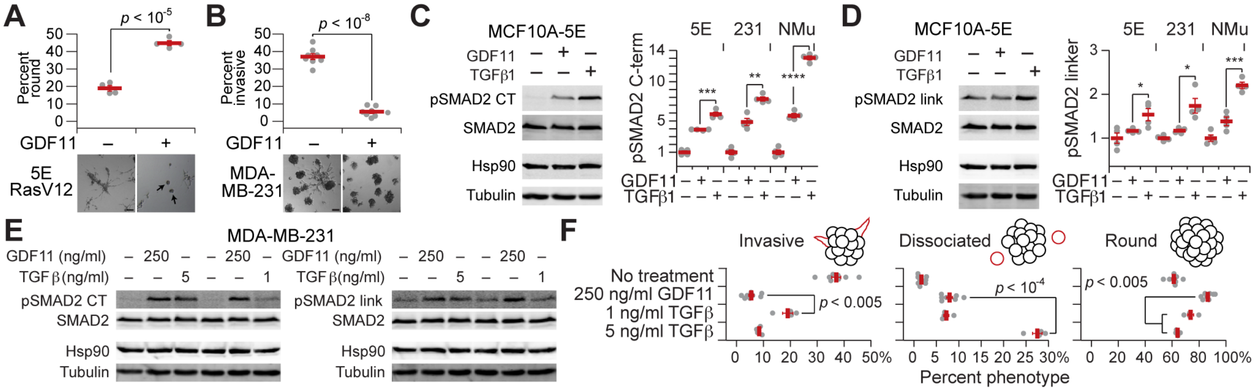


Figure 5.3 GDF11 uniquely promotes compact tissue architecture of BLBC spheroids.

(A) GDF11 improves rounding in RasV12 transformed MCF10A-5E spheroids. (B) GDF11 blocks stellate invasion of MDA-MB-231 claudin-low breast carcinoma cells. (C,D) GDF11 is a weaker agonist of Smad2 phosphorylation on C-terminal (CT) sites and linker (link) sites. *p < 0.05. **p < 0.01. ***p < 0.001. ****p < 10-6. (E,F) Lower concentrations of TGFβ1 cannot mimic the normalization phenotype of GDF11 in MDA-MB-231 cells. Scale bar is 200 μm.

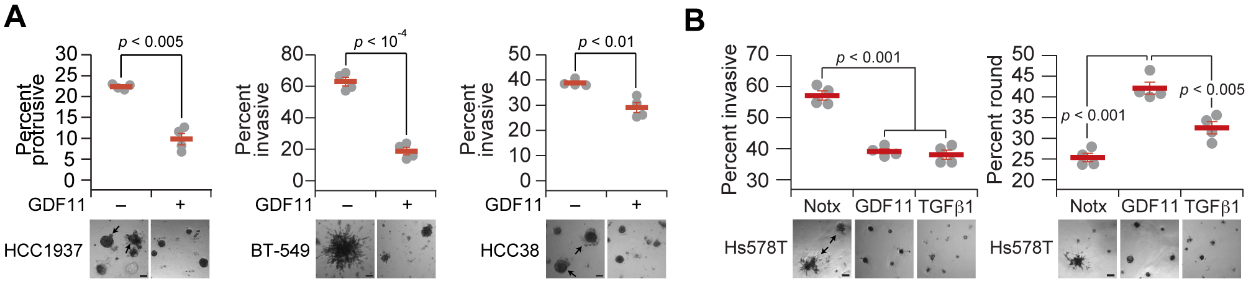


Figure 5.4 GDF11 promotes compact tissue architecture in a panel of BLBC spheroids.
(A) GDF11 blocks sheet-like protrusion of HCC1937 basal-like breast carcinoma cells, stellate invasion of BT-549 claudin-low breast carcinoma cells, and sheet-like invasion of HCC38 claudin-low breast carcinoma cells. GDF11, but not TGFβ1, improves lobule-like organization of Hs578T spheroids. Scale bar is 200 μm.

5.3.3 GDF11 phenotype requires SMAD4 and ID2

The perplexing difference in phenotypes induced by GDF11 and TGFβ1, despite equal pathway activation, warranted further examination. To test whether transcriptional changes were required for the GDF11-induced phenotype, we knocked down a central node of the canonical TGFβ-family transcriptional pathway, SMAD4 (Figure 5.5A) (217). In MDA-MB-231 cells, knockdown of SMAD4 caused an increase in invasion, suggesting TGFβ-family responsiveness is key in mitigating invasion in the parental cell line. Importantly, treatment of shSMAD4 spheroids with GDF11 resulted in no change in invasion or rounding (Figure 5.5B). We tested the gain of SMAD4 function by ectopically expressing SMAD4 in the MDA-MB-468 basal-like cell line that has a genomic deletion of *SMAD4* (Figure 5.5C) (218). In the absence of SMAD4 expression, GDF11 elicited no change in spheroid phenotype. However, upon expression of SMAD4, GDF11 treatment induced rounded spheroids (Figure 5.5D). Together, these data suggest that SMAD4 is a critical transcriptional mediator of the GDF11-induced phenotype.

We next sought to identify downstream transcriptional targets of SMAD4 that could further mediate GDF11 function. We transcriptionally profiled two non-transformed breast epithelial cell lines (MCF10A-5E and NMuMG) with normally functioning TGFβ pathways, stimulated with either GDF11 or TGFβ1 (212, 213, 219). Of the transcripts that were differentially by GDF11, we identified a single transcriptional regulator that was upregulated by GDF11, *inhibitor of DNA binding 2* (*ID2*) (Figure 5.5E). Inhibitor of DNA binding (ID) family members have been implicated in TGFβ signaling (220), but not in the context of GDF11. Additionally, ID2 has been associated with decreased invasion and improved prognosis in breast cancer cells, corroborating a potential role for ID2 in mediating phenotypes induced by GDF11 (221). We used a loss of function approach to test if ID2 was required for GDF11 to promote compacted, lobular architecture. Knockdown of Id2 in NMuMG spheroids prevented GDF11 from blocking abnormal morphogenesis (Figure 5.5F). Similarly, knockdown of ID2 in MDA-MB-231 spheroids prevented GDF11 from blocking invasion (Figure 5.5G), suggesting ID2 is required for the phenotype changes induced by GDF11.

To examine if ID2 was relevant to breast cancers where GDF11 signaling could be active, we mined the breast cancer TCGA data (27). There were not enough claudin-low tumors for a statistical analysis, but among basal-like breast cancers, we found that *GDF11*, *ACVR2B*, *TGFBR3*, and *ID2* showed substantially more interpatient variability than canonical pathway components *TGFBR1*, *SMAD2/3*, and *SMAD4* (Figure 5.6A). When examining tumors in the upper 40th percentile of GDF11 expression, we observed a significant increase in tumors with higher expression of the GDF11-receptors, *ACVR2B* and *TGFBR3*, suggesting this pathway could be active in a subset of BLBCs (Figure 5.6B). These data support SMAD-dependent transcriptional regulation of *ID2* as a key effector of GDF11 in 3D breast epithelial cultures and in human BLBCs.

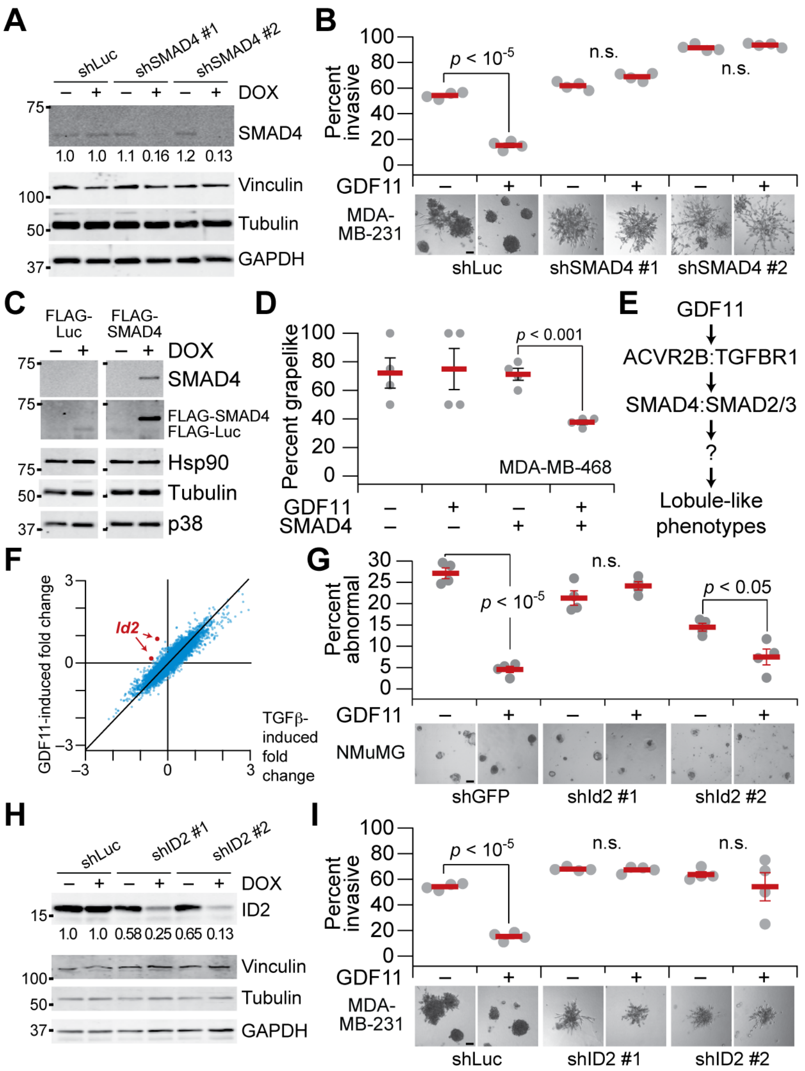


Figure 5.5 GDF11-specific phenotype are mediated by SMAD4 and ID2.

(A) Inducible knockdown of SMAD4 in MDA-MB-231 claudin-low breast cancer cells. (B) Knockdown of SMAD4 prevents GDF11 from reducing invasion in MDA-MB-231 spheroids. (C) Inducible overexpression of SMAD4 in the *SMAD4* null MDA-MB-468 basal-like breast cancer line. (D) SMAD4 expression allows GDF11 to improve rounding in MDA-MB-468 spheroids. (E) Identified signaling axis for GDF11-induced lobule-like phenotypes. (F) Transcriptional profiling of NMuMG cells stimulated with GDF11 or TGFβ1. (G) Knockdown of Id2 prevents GDF11 from reducing abnormal NMuMG morphogenesis. (H) Inducible knockdown of ID2 in MDA-MB-231 claudin-low breast cancer cells. (I) Knockdown of ID2 prevents GDF11 from reducing invasion in MDA-MB-231 spheroids. Scale bar is 200 μm.

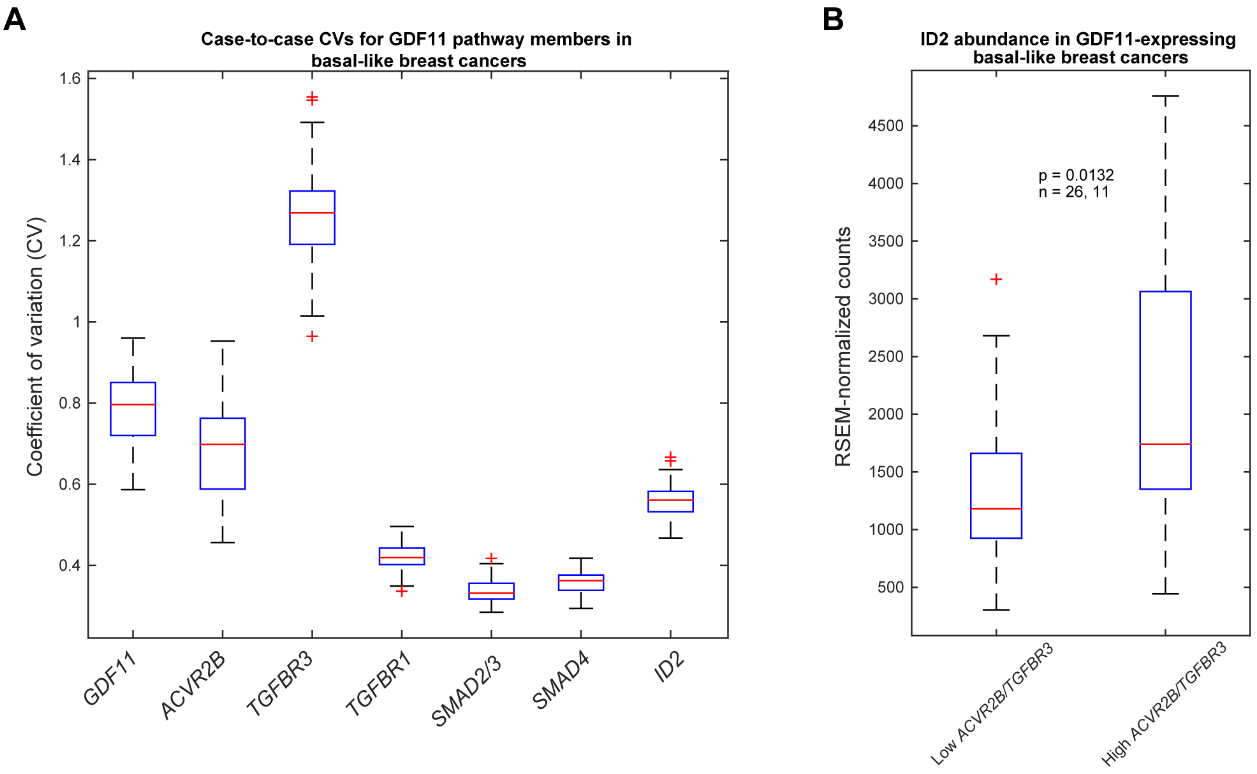


Figure 5.6 GDF11 signaling axis is present in human basal-like breast cancers.

(A) TCGA breast cancer sequencing data was mined for expression of GDF11 signaling components. The coefficient of variation amongst patients for each transcript shows high variability in *GDF11*, *ACVR2B*, *TGFBR3*, and *ID2*. (B) *ID2* levels are higher in basal-like breast cancers with expression of GDF11 receptors *ACVR2B* and *TGFBR3*.

5.3.4 GDF11 effects in vivo

The TCGA meta-analysis suggested that GDF11 signaling was active in a subset of BLBCs; however, ~10% of BLBCs have less than one copy of *GDF11* per cell, implying that in these samples *GDF11* is absent in over half of the tumor cells (27). To test if loss of GDF11 altered BLBC phenotype, we knocked down GDF11 in a transformed derivative of the MCF10A, MCF10ADCIS.COM (121), and a basal-like breast cancer cell line, HCC1937, which could secrete GDF11 at high levels (Figure 5.10E). Delayed knockdown of GDF11 caused a ruptured-like appearance in both cell lines, which was not observed in a non-transformed cell line (MCF10A-5E) (Figure 5.7A). These data suggest that loss of GDF11 caused a cellular state that is stressful for transformed cells.

To test if this cellular stress had a role in tumorigenesis, we injected luciferase labeled MCF10ADCIS.COM cells expressing an inducible hairpin against control or *GDF11* into the murine mammary gland intraductally. Intraductal injection xenografts are a stringent model of tumorigenesis, requiring tumor cells to survive in suspension until they colonize in the duct, proliferate to fill the duct forming an early in situ lesion, and progressing to invasive carcinoma only after the cells have invaded through the basement membrane surrounding the murine epithelium (222). Additionally, the progression of cell behavior in the xenograft model closely mimics the cellular state transitions present in our 3D spheroid model (53). We allowed the MCF10ADCIS.COM tumors to grow for two weeks and then induced the hairpin. In a pilot study, we observed a significant increase in bioluminescence in GDF11 knockdown tumors with one of our two validated GDF11 hairpins.

To rigorously confirm the results of the pilot study, we created a preregistered study on the Open Science Framework (https://osf.io/8pv7d/) (223). Pregistration creates a binding documentation of the data collection, data analysis and processing, and statistical analysis of the data before it is collected. The goal of this process is to increase openness in science research, promoting reproducibility (224). All pilot data collected informs the design of the confirmatory study, but is not publishable. In the confirmatory study, we injected MCF10ADCIS.COM cells expressing one of two GDF11 hairpins into one mammary gland. The contralateral mammary gland was injected with MCF10ADCIS.COM cells expressing a control hairpin sequence. After two weeks of hairpin expression, we observed only a modest increase in bioluminescence (Figure 5.7B). The effect size was smaller than in our pilot study (224), suggesting that loss of GDF11 does not significantly alter tumor growth.

Based on pathway expression, GDF11 was suggested to be bioactive in a subset of human BLBCs, raising the question of its role in tumorigenesis, particularly in early premalignancies where cells must colonize the duct (53). To explore this question, luciferase labeled MDA-MB-231 cells were injected into the mammary gland. As we could not overexpress bioactive GDF11 in this cell line (Figure 5.10C,E), we tested overexpression of GDF11 by co-inoculating the mammary gland with GDF11 or BSA as control. In a pilot study, we observed that GDF11 significantly reduced the growth rate of intraductal tumors. However, the mouse-to-mouse variation exceeded the variation in bioluminescence, precluding a direct comparison of tumor bioluminescence. In a preregistered study (https://osf.io/kjypa/), we proposed a paired design with one gland receiving GDF11 co-inoculation and the contralateral gland receiving BSA as control, alleviating any overall mouse-to-mouse variation. The BSA tumor bioluminescence decreased until approximately two weeks post injection and then began to increase for the duration of the experiment. This suggests the intraductal environment is stressful and difficult to survive for most BLBCs and also causes selection for robust tumor clones. In contrast, the GDF11 treated tumors had a minimal rebound from the selection process, with many tumors failing to grow (Figure 5.7C). This suggests that GDF11 inhibits the ability of claudin-low breast cancer cells to seed tumors.

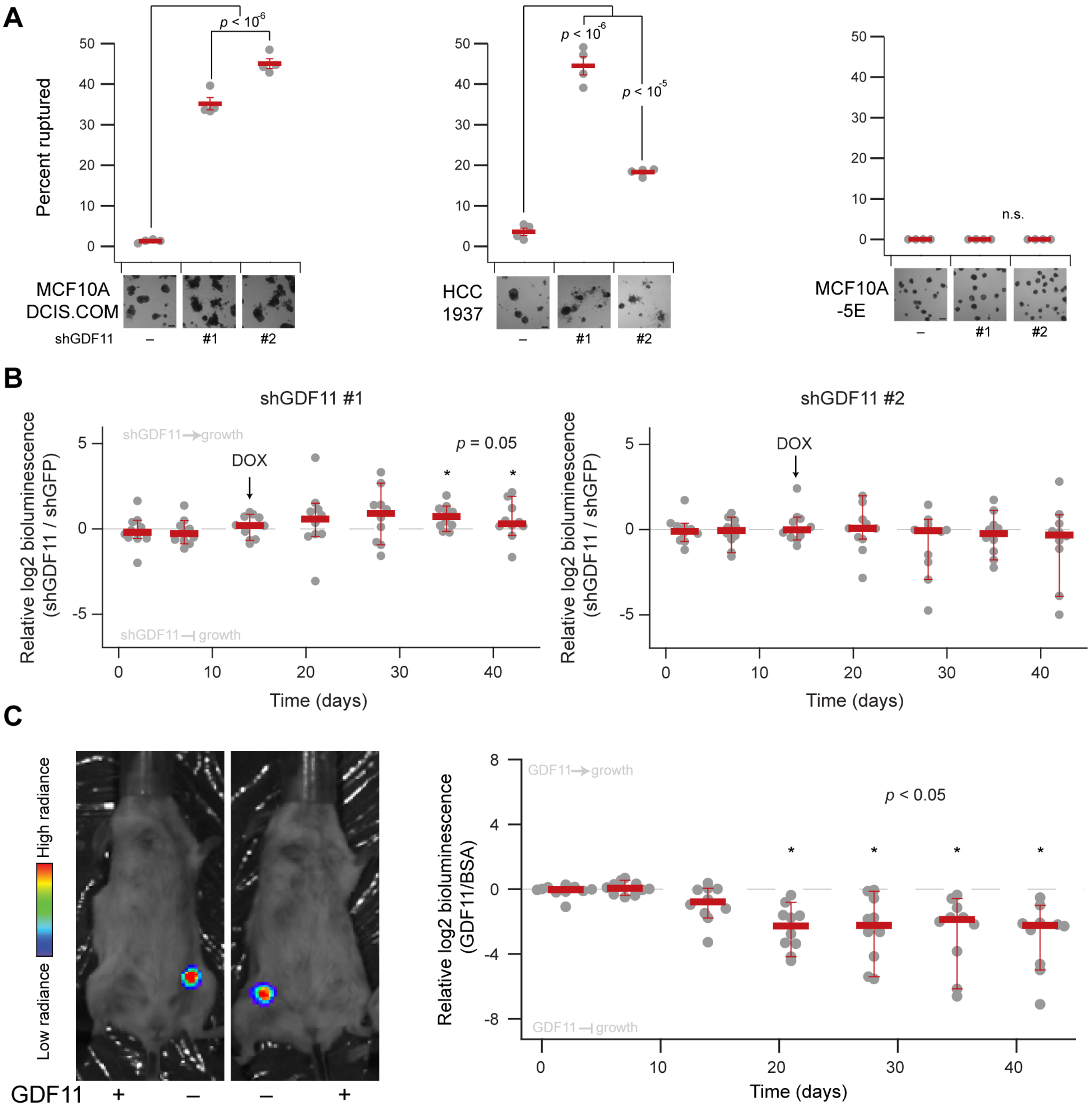


Figure 5.7 Loss and gain of GDF11 function during tumorigenesis.

(A) The transformed cells MCF10ADCIS.COM (left) and basal-like breast carcinoma cell line HCC1937 (center) rupture from knockdown of GDF11. This phenotype is not seen in a non-transformed breast epithelial cell line MCF10A-5E (right). (B) Intraductal injection of luciferase labeled MCF10ADCIS.COM cells. Knockdown of GDF11 was induced at day 14 and bioluminescence was recorded weekly. (C) Co-inoculation of 100 µg/mL GDF11 in intraductal injections of luciferase labeled MDA-MB-231 cells blocks tumor seeding and growth. Scale bar is 200 µm.

5.3.5 GDF11 expression in human breast and basal-like breast cancers

GDF11 transcript abundance was highly variable among clinical breast cancer cases and at the single-cell level in 3D breast epithelial cultures (Figure 5.1C). These two observations could be reconciled if there were intratumor heterogeneity of GDF1 bioactivity in advanced tumors. Using the EPR antibody clone as the cleanest reagent for detection of GDF11 in tissue (Figure 5.1C), we optimized immunohistochemical conditions for formalin fixed paraffin embedded (FFPE) samples and examined dozens of cases of reduction mammoplasty ("normal"), premalignant ductal carcinoma in situ lesions with basal-like features ("basal-like DCIS"), and advanced breast cancers deemed triple-negative by routine clinical pathology ("TNBC"). In normal tissue, we observed positive EPR staining that was often stronger in the lobules compared to the duct (Figure 5.8A), consistent with the lobule-like phenotypes induced by recombinant GDF11 in 3D culture (Figure 5.1G,I). Moreover, when TNBC was compared to adjacent normal tissue, there was a significant enhancement of GDF11 abundance in the tumor (Figure 5.8B). Contrary to our expectation based on animal experiments, the clinical data indicated that GDF11 protein abundance increases in breast cancer cells during progression.

While performing the immunohistochemical studies, we serendipitously discovered a second monoclonal GDF8/11 antibody ("1E6") marketed commercially for ELISA. Under the same antigen-retrieval conditions, this antibody yielded a specific staining pattern that was distinct from the EPR antibody. GDF11 staining with 1E6 was elevated in lobules compared to ducts of normal breast tissue as with the EPR clone. However, rather than localizing to the Golgi, 1E6 immunoreactivity was more diffusely localized to intracellular vesicles and the pericellular space (Figure 5.8C). The discrepancy could be explained if the 1E6 antibody were selective for a more-mature proteoform of GDF11 (225). Both antibodies were raised against similar C-terminal fragments of GDF11, and we observed correlated signal between the two antibodies by two-color immunoblotting and immunofluorescence (Figure 5.9C,D). By contrast, 1E6 did not clearly detect proGDF11 unless its prodomain was deglycosylated, indicating that glycosylation of the GDF11 precursor masks the 1E6 epitope. We conclude that the 1E6 antibody achieves selectivity for mature GDF11 once the prodomain has been cleaved by the proprotein convertase PCSK5 (226), an isoform of which is sorted to secretory granules (227).

When staining of the basal-like DCIS and TNBC cohorts was repeated with the 1E6 antibody, we observed a markedly different trend of GDF11 immunoreactivity. In contrast to EPR, there was no clear difference in overall 1E6 staining between tumor and normal tissue. Instead, TNBC cases were distinguished by their intratumor heterogeneity, evidenced by sporadic foci of intense 1E6 immunoreactivity that diffused radially from single cells (Figure 5.8D,H). TNBC specimens that lacked foci were often characterized by pockets of tumor cells engorged with 1E6-reactive vesicles, further indicating a heterogeneous misregulation in GDF11 processing or release (Figure 5.8G). 1E6 foci were almost never found in normal tissue (Figure 5.8E), but a TNBC-like pattern was noted in ~40% of basal-like DCIS cases, raising the possibility that a shift in GDF11 regulation precedes the premalignant-to-malignant transition (Figure 5.8F). Despite an overall increase in GDF11 protein, the 1E6 staining results suggest that TNBC cells are largely unable to convert it into a bioactive autocrine or paracrine ligand.

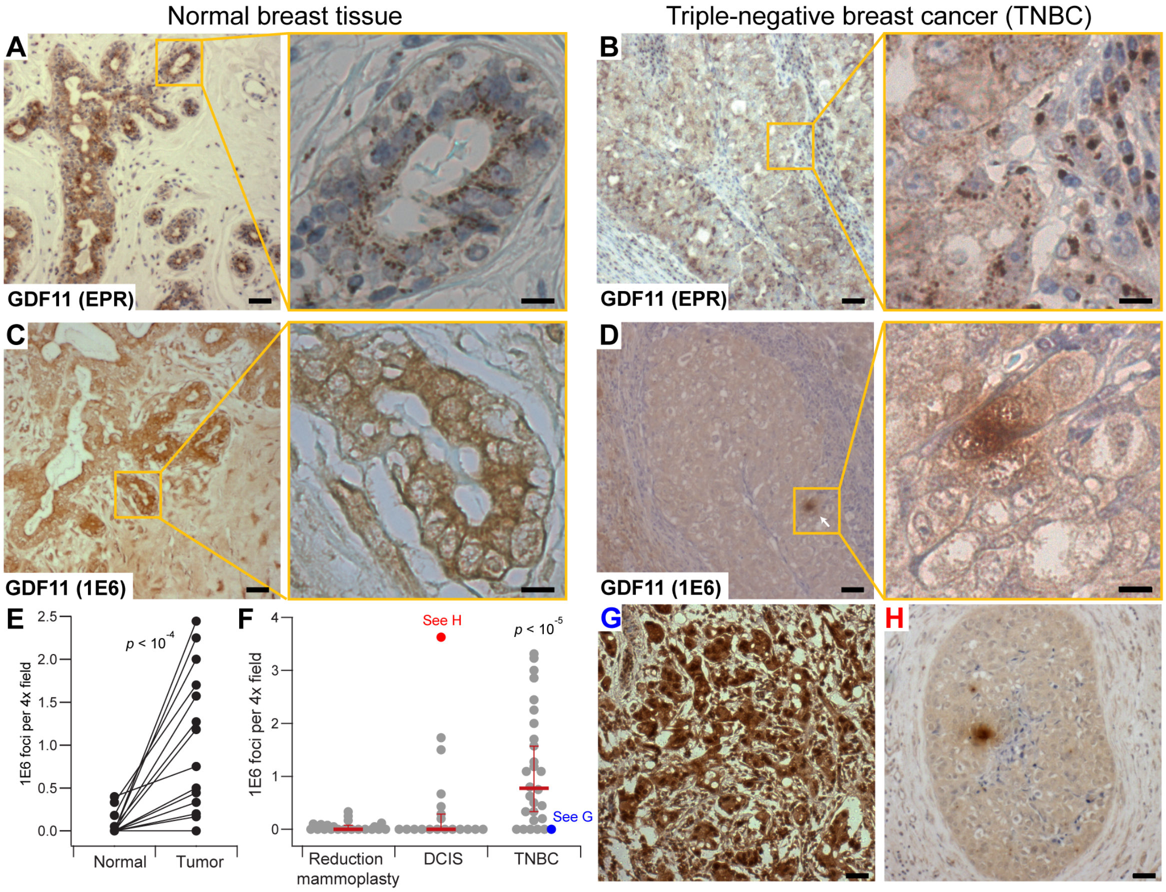


Figure 5.8 Distinct immunolocalization of GDF11 in triple-negative breast cancer according to EPR and 1E6 monoclonal antibodies.

(A,C) Coordinate localization of (A) EPR and (C) 1E6 immunoreactivity on the apical side of lobular epithelia in normal breast. (B,D) Discrepant localization of (E) EPR and (F) 1E6 immunoreactivity in triple-negative breast cancer. (E, F) Quantification of 1E6 immunoreactive foci in matched normal-tumor pairs (E) and broadly across reduction mammoplasties, DCIS, and TNBC (F). (G) Example of TNBC tumor with no 1E6 immunoreactive foci, but high intracellular immunoreactivity. (H) Example of DCIS lesion with 1E6 immunoreactive foci. Scale bar is 200 μm (A,B,C,D left), 80 μm (G,H), 20 μm (A,B,E,F right).

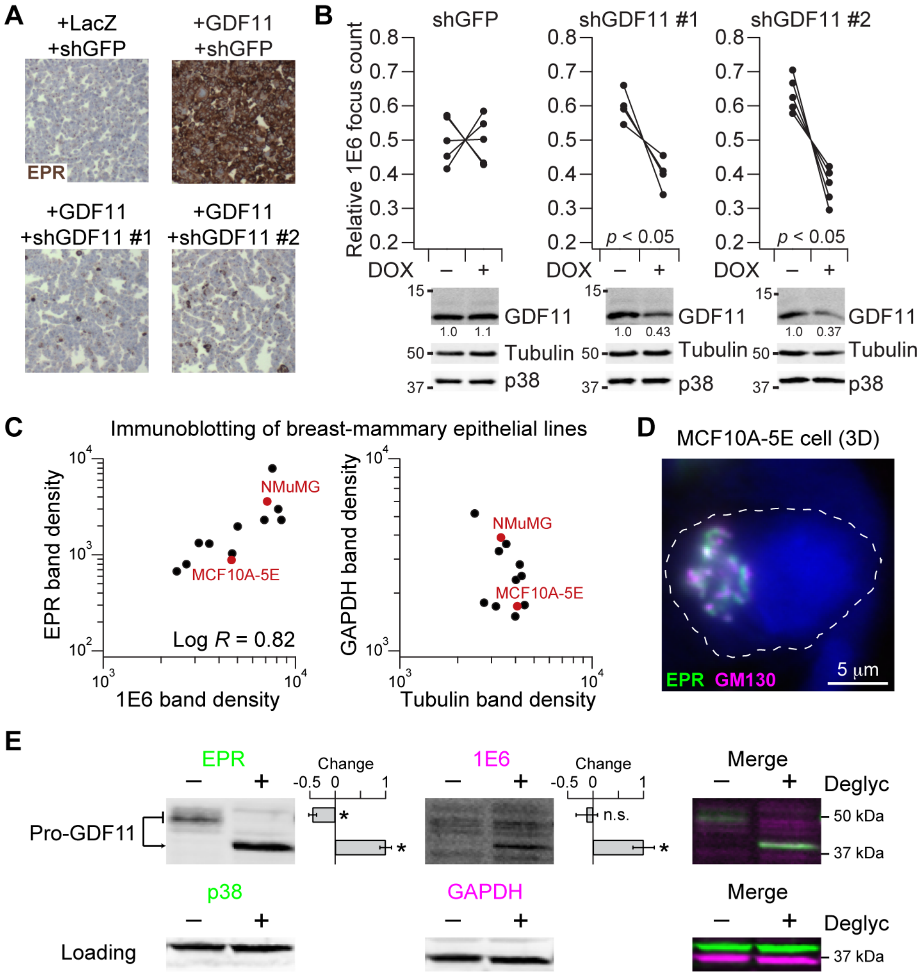


Figure 5.9 Validation of EPR and 1E6 monoclonal antibodies.

(A) Immunohistochemical staining of 293T cells overexpressing GDF11 or LacZ control with or without GDF11 knockdown. Scale bar is 200 µm. (B) Inducible shRNA knockdown of 1E6 focus count and endogenous GDF11 in MDA-MB-231 cell pellets. (C) EPR and 1E6 immunoblot intensities from 40–55 kDa are highly correlated among breast-mammary cancer lines (black) and immortalized cells (red). GAPDH and tubulin were reprobed as a control for spurious correlations caused by loading. (D) EPR immunoreactivity (green) is adjacent to and overlapping with the cis-Golgi marker GM130 (magenta). A single MCF10A-5E cell (dashed arrow) in 3D culture is highlighted. (E) Deglycosylation of MDA-MB-231 cellular extracts causes the EPR (green) and 1E6 (magenta) antibodies to colocalize on a ~42 kDa band consistent with the predicted mass of pro-GDF11. p38 and GAPDH were used as loading controls.

5.3.6 GDF11 secretory defect in TNBCs

To examine the conversation of proGDF11 in greater detail, we returned to in vitro studies using TNBC cell lines. Paraffin-embedded pellets of MDA-MB-231 cells showed the same 1E6-reactive foci as TNBC patients, and focal counts were significantly reduced to an extent proportional to GDF11 knockdown (Figure 5.10A and Figure 5.9B). In this setting, ectopic overexpression of V5-tagged GDF11 had no effect on invasive 3D growth of MDA-MB-231 cells, whereas recombinant GDF11 remained strongly repressive (Figure 5.10B,C). These results correlate 1E6 foci with a lack of cell-derived GDF11 bioactivity.

Recently, it was reported that TNBC cells widely exhibit a chronic unfolded protein response (UPR) (228), which could disrupt the trafficking of secreted factors with stringent folding requirements (229). Like other TGFβ-family members, GDF11 requires one intermolecular and multiple intramolecular disulfide bonds to bioactivity (230); improper folding of GDF11 should block secretion and promote its intracellular retention. Using conditioned medium from non-transformed MCF10A-5E cells, we confirmed normal GDF11-V5 secretion was eliminated by various UPR inducers that block secretion of other ligands (Figure 5.10D) (231). We next tested for cancer-associated defects in GDF11 release by expressing GDF11-V5 in nine cell lines from various TNBC subtypes and detecting GDF11 secretion by V5 immunoprecipitation. Compared to MCF10A-5E cells, seven of nine TNBC lines were clearly deficient in the release of GDF11-V5 (Figure 5.10E). Defective GDF11 secretion did not reflect a global blockage of cytokine release, as TGFβ1 was readily detected in the conditioned medium of most TNBC lines (Figure 5.10E). Furthermore, GDF11 secretion was not inversely correlated with the established EPR markers, GRP78 and ATF4 (232), suggesting a mechanism of regulation that was distinct from a canonical UPR, yet unique to TNBC. These data suggest that

TNBC cells sequester GDF11 intracellularly, which blocks the tumor suppressor ability of GDF11 (Figure 5.7C and Figure 5.10C).

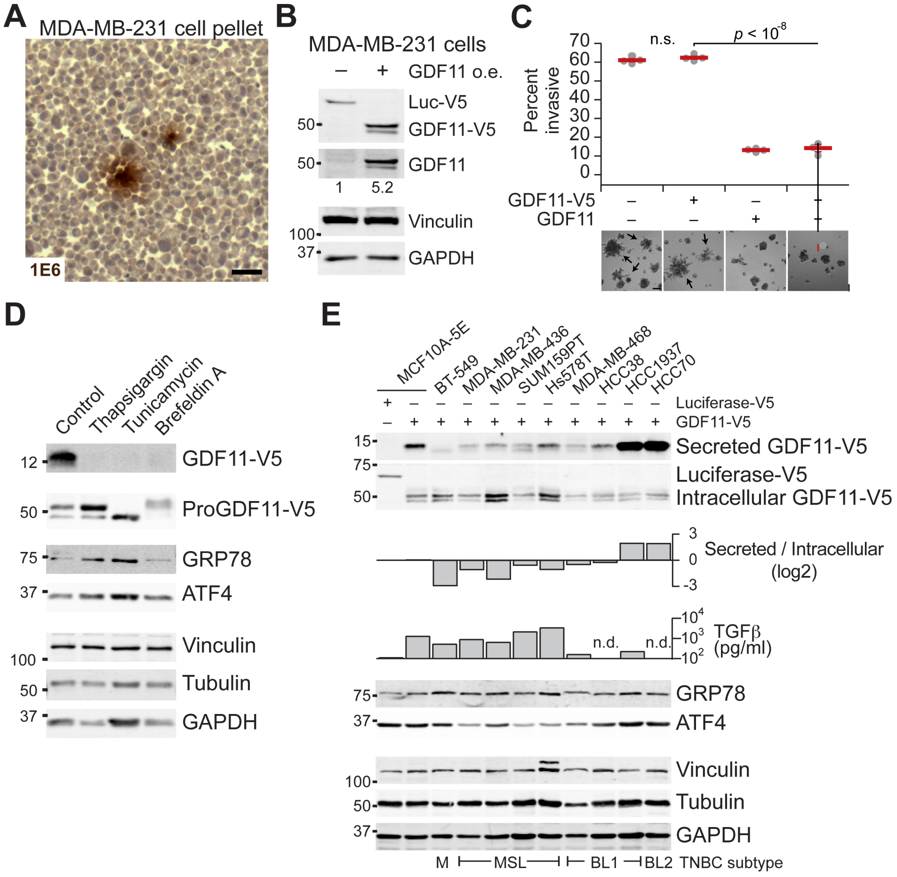


Figure 5.10 Misregulation of GDF11 secretion in TNBC.

(A) MDA-MB-231 claudin-low breast carcinoma cell pellet shows similar 1E6 immunoreactive foci as observed in clinical specimens of TNBC. (B) Constitutive overexpression of GDF11 in MDA-MB-231 cells. (C) Overexpression of GDF11 does not block invasion of MDA-MB-231 spheroids. (D) UPR inducers alter proGDF11 abundance and glycosylation, upregulate GRP78, and block mature GDF11 release. (E) Multiple triple-negative subtypes (72) show defects in secretion of ectopic GDF11. For (D) and (E), secreted GDF11 was measured by V5 immunoprecipitation from conditioned medium. Scale bar is 80 μm in (A) and 200 μm in (C).

5.4 Discussion

In this study, we identified a tumor suppressive role for GDF11 that is uniquely post-translationally misregulated in human basal-like breast cancers. We began by profiling expression heterogeneities in basal-like breast epithelial 3D spheroids and discovered that *GDF11* was heterogeneous amongst individual cells in these spheroids. Normalizing GDF11 expression led to improved lobular-like spheroid morphology, which was also observed in BLBC cell line spheroids. Consistent detection of GDF11 in human specimens of TNBC contradicted our in vitro and in vivo results. By carefully examining GDF11 processing in BLBC cells, we discovered that GDF11 is post-translationally sequestered, inactivating its tumor suppressive function. We propose intracellular sequestration as a new mechanism of how cancer cells escape from secreted tumor suppressors.

5.4.1 GDF11 as a tumor suppressor

The role of GDF11 in breast cancer has not been previously studied. GDF11 has been mostly implicated in development (141, 207, 233, 234) and, controversially, in regeneration of cardiac, muscle, and brain tissue (206, 209, 235-239). While mechanism of GDF11's regenerative ability is not currently known, GDF11 plays an important role in tissue patterning during development (207, 234). Additionally, GDF11 controls cell fate decisions and tissue homeostasis in various contexts (240-242). Interestingly, intracellular GDF11 displayed highly regular, polarized expression in normal breast epithelial cells (Figure 5.8A). GDF11 may function to define lobular epithelial cell patterning in mammary tissue, supported by the increased protein expression seen in the lobular regions of normal human breast tissue (Figure 5.8A, C). Heterogeneous expression of GDF11 in our spheroid model may be the individual cells in the spheroid asynchronously setting a lobular pattern during morphogenesis. Exogenous GDF11 homogenously forces the pattern across all the cells in the spheroid, leading to an improved lobular phenotype (Figure 5.1G,H). Loss of functional GDF11 in early DCIS lesions may remove an important instructive cue for the transformed epithelial cells to remain lobular-like and allow them to continue along a malignant trajectory (Figure 5.7A and Figure 5.8E,H) (242).

5.4.2 GDF11 as a clinical biomarker

GDF11 secretion is misregulated in TNBC and a subset of DCIS (Figure 5.8F and Figure 5.10). The correlation of improper GDF11 secretion with tumor progression suggests that this cell biology phenotype in the cancer cells could be used as a biomarker of prognosis. DCIS lesions are commonly identified by mammography (243). Interestingly, only a small subset of DCIS lesions will progress to invasive ductal carcinoma. There are no known biomarkers that differentiate the relatively benign and aggressive DCIS (243). Thus, all identified DCIS are treated aggressively, overtreating a large subset of patients that would otherwise progress very slowly (244, 245). Our data suggests that GDF11 secretion, as measured with the 1E6 antibody, could be an informative biomarker to predict aggressive DCIS lesions. Those DCIS tissues with misregulated GDF11 secretion (increased 1E6 foci) would be predicted to be more aggressive as their tissue staining is more similar to the advanced tumors than normal tissue.

GDF11 levels have been reported to increase in colorectal cancer (246), contradicting the role we identify for GDF11 in basal-like breast cancer. This work and our study can be reconciled, as we also observed an increase in intracellular GDF11 in our clinical specimens. GDF11 may also be post-translationally misregulated in colorectal cancers, and the increase in *GDF11* transcripts being an indicator of the cancer cells trying to express GDF11. The seeming contradiction in the results further points to the need to understand the bioactivity and bioavailability of genes and proteins in the context of cancer.

5.4.3 Selective misregulation of secreted tumor suppressors

GDF11 misregulation aligns with cancers disrupting development pathways (247-249). These pathways are typically activated or silenced by genetic mutation, as many other oncogenes or tumor suppressors (27). GDF11, by contrast, is misregulated post-translationally via intracellular sequestration. Other tumor suppressive secreted ligands may be inactivated in a similar manner. For example, growth-differentiation factor 15 (GDF15) has been shown to inhibit tumorigenesis in many contexts (250-254). The effects of GDF15 have been noted to be more potent at early stages of tumor progress and not reported to be silenced (251, 255). GDF15 may be post-translationally inactivated, allowing the cancer cells to escape growth inhibition.

The inability of TNBC cell lines to secrete GDF11 was surprisingly anticorrelated with canonical UPR markers (Figure 5.10D), suggesting GDF11 inactivation is regulated. Identifying the genes involved in GDF11 inactivation could reveal important and surprising oncogenes. Using conventional global approaches, GDF11 misregulation would not be identified on the genomic, transcriptomic, or proteomic level. Our work broadly suggests that there may be many tumor suppressor proteins that are misregulated post-translationally.

6 Chapter 6 – Conclusions and future directions

6.1 Summary of Dissertation

In this Dissertation, we used multiple approaches to demonstrate that transcriptional regulatory heterogeneities play a critical role in breast epithelial morphogenesis. Computational analyses (Part I) were used to extract parameters that were difficult to obtain (Chapters 2 and 4), and we use these parameters to quantitatively examine heterogeneity, leading to nonobvious roles for certain genes (Chapters 3 and 4). Additionally, modeling was used to synthesize multiple sets of experimental data to develop and support hypotheses by exploring the emergent phenotypic outcomes of the models (Chapter 3). Computational approaches can highlight or suggest biological function, but these predictions must be experimentally validated (Part II). We started with the prediction that *GDF11* was heterogeneously regulated in our 3D spheroid model. Through careful, mechanistic experimentation, we identified a pivotal role for GDF11 in breast cancer morphogenesis. Additionally, GDF11 treatment blocked the ability of a basal-like breast cancer cell line from seeding tumors, suggesting GDF11 plays a role in tumorigenesis. By examining human specimens of triple-negative breast cancer, we discovered that advanced tumors accumulated intracellular GDF11, seemingly contradicting both the in vitro and in vivo roles of GDF11 we had identified. However, by looking at multiple forms of GDF11, namely the processed, secreted form, we observed that the intracellular accumulation of GDF11 correlated with a lack of GDF11 secretion. Strikingly, we found that triple-negative breast post-translationally silence GDF11 through a defect in GDF11 secretion. This unusual mechanism of tumor suppressor silencing could not be detected with genomic, transcriptomic, or proteomic approaches (Chapter 5).

6.2 Future directions

While the work in this Dissertation shed light into the function of heterogeneity in basal-like breast epithelia, the approaches and results presented can still be further extended to uncover additional roles of heterogeneity. In this Section, we discuss future directions of the work presented in each Chapter.

6.2.1 Chapter 2 future work

In Chapter 2, we developed an automatic, computational approach to quantify spheroid metrics. We applied this approach to basal-like breast cancer spheroids and were able to use the quantified metrics to distinguish spheroid phenotypes in different cell lines. Additionally, we used the metrics to quantify spheroid phenotypic heterogeneity. This particular spheroid culture system was chosen for its relevance to the work in the Dissertation and for its potential utility in the lab (53). However, there are many 3D spheroid models, and the approach is poised to be applied to any of them (107, 114, 128, 136, 137, 256-259).

3D spheroids can be generated from patient derived material. Recently, efforts have gone towards identifying the isolation and culturing conditions needed to propagate both normal and tumor epithelial tissues, most notably in the colon and pancreas, in an organoid model (107, 114, 128, 129). These organoids have the potential to be powerful tools for personalized medicine as they recapitulate key characteristics of the primary tumor (107), yet can be perturbed in an experimentally tractable and timely manner (129, 260). Additionally, these organoids have been primarily characterized molecularly and qualitatively through histology or immunofluorescence (107, 114, 128). The morphometric properties of these organoids have not been examined.

We obtained a small set of images of tumor organoids derived from different primary tumors by Hans Clevers' group. The Clevers group split the primary tumors and generated organoids from different clones from the same patient. Through these efforts, they have observed inter- and intra-tumor heterogeneity in organoids phenotypes. We confirmed these observations quantitatively by segmenting individual organoids from two different patients and different subclones. We identified a cluster of organoids for each patient. Within each cluster, the clones varied. Additionally, each cluster contained spheroids from the other patient. There were also smaller clusters of organoids that were comprised of organoids from both patients across multiple clones (Figure 6.1). These results suggest that irrespective of patient (intertumor heterogeneity) and subclone (intratumor heterogeneity), there are categories of similar phenotypic behavior. Understanding the regulators of these phenotypes could yield important therapeutic targets that push organoids from aggressive to non-aggressive phenotypes.

Testing the algorithm on another spheroid system raised challenges not presented by our breast epithelial spheroid model. In the patient derived organoids, the density of cells is higher and the distance between spheroids is smaller. We observed many spheroids that were overlapping or had the appearance of being fused to one another. The analysis presented ignores all such spheroids. One future improvement we are currently working on is a means to fit the correct spheroids within a clump or overlapped region. While our eyes can do this segmentation quite easily, the computer cannot (261). Improvements to the algorithm will further enhance the utility and generality of digital morphometry.

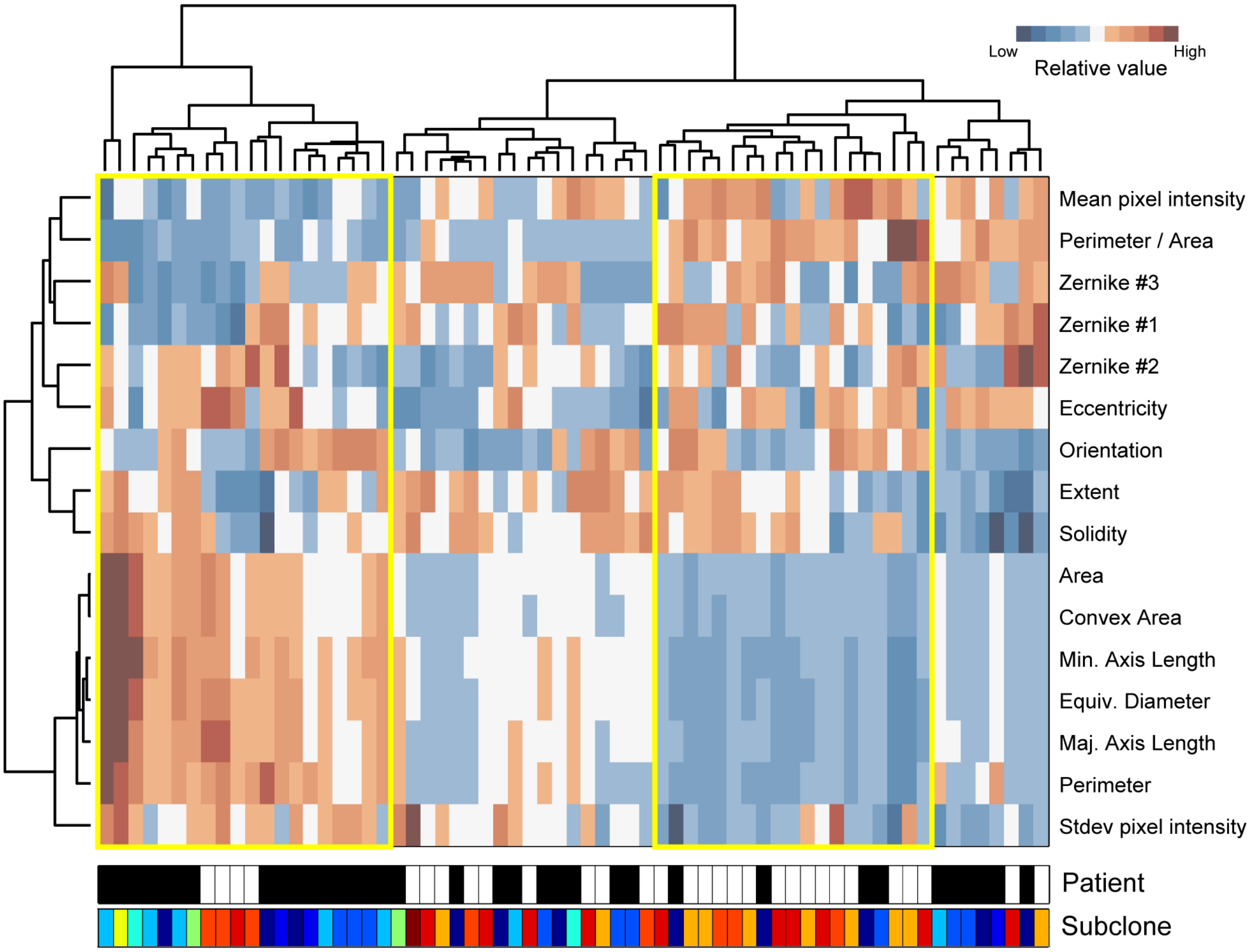


Figure 6.1 Digital morphometry quantifies tumor organoid heterogeneity.
Digital morphometry (see Chapter 2) was applied to tumor organoids derived and cultured in the Clevers group. Extracted signatures of tumor organoids were hierarchically clustered. Organoid metrics are displayed along the rows. Organoids from two patients were analyzed; the patient is shown below the clustergram, with white and black designating the two patients. Multiple subclones per patient were isolated, and the subclone status is shown by the colorbar underneath patient status. Yellow boxes highlight pockets of organoids from the same patient but different subclones.

6.2.2 Chapter 3 future work

In Chapter 3, we used an ODE model to demonstrate how connections between *JUND* and *TGFBR3* regulation led to damped, anticorrelated oscillations. We next used an agent-based model of the matrix molecule TNC to show how its regulation could explain the expression patterns of JUND and KRT5 in human basal-like breast cancers (53). Perturbing the models showed the robustness of the behaviors, and in the case of the ODE model, help motivate other work in this Dissertation by identifying GDF11 as a potential trigger of the oscillations (see Chapter 5). These two results show the power of using the emergent behavior of computational modeling to explain and simplify cellular heterogeneity.

Each of these models was constructed to explain distinctly different scales of cellular heterogeneity. The ODE model captured the dynamics that characterize JUND intercellular heterogeneity at the molecular level. The agent-based model captured the dynamics that characterize JUND intratumor heterogeneity at the multicellular level. Given the common molecular component between the two models (JUND), future effort could go toward fusing and reconciling these models into a singular, multi-scale model (Figure 6.2) (262, 263). The aggregated could be used to further understand the regulation of TNC in basal-like breast cancer and clinging carcinomas, as we have observed that TNC expression is required for basal-like breast tumor cell seeding. Perturbations of the model could elucidate potential therapeutics to mitigate the effects of TNC and prevent breast cancer cells from locally spreading throughout the ductal network.

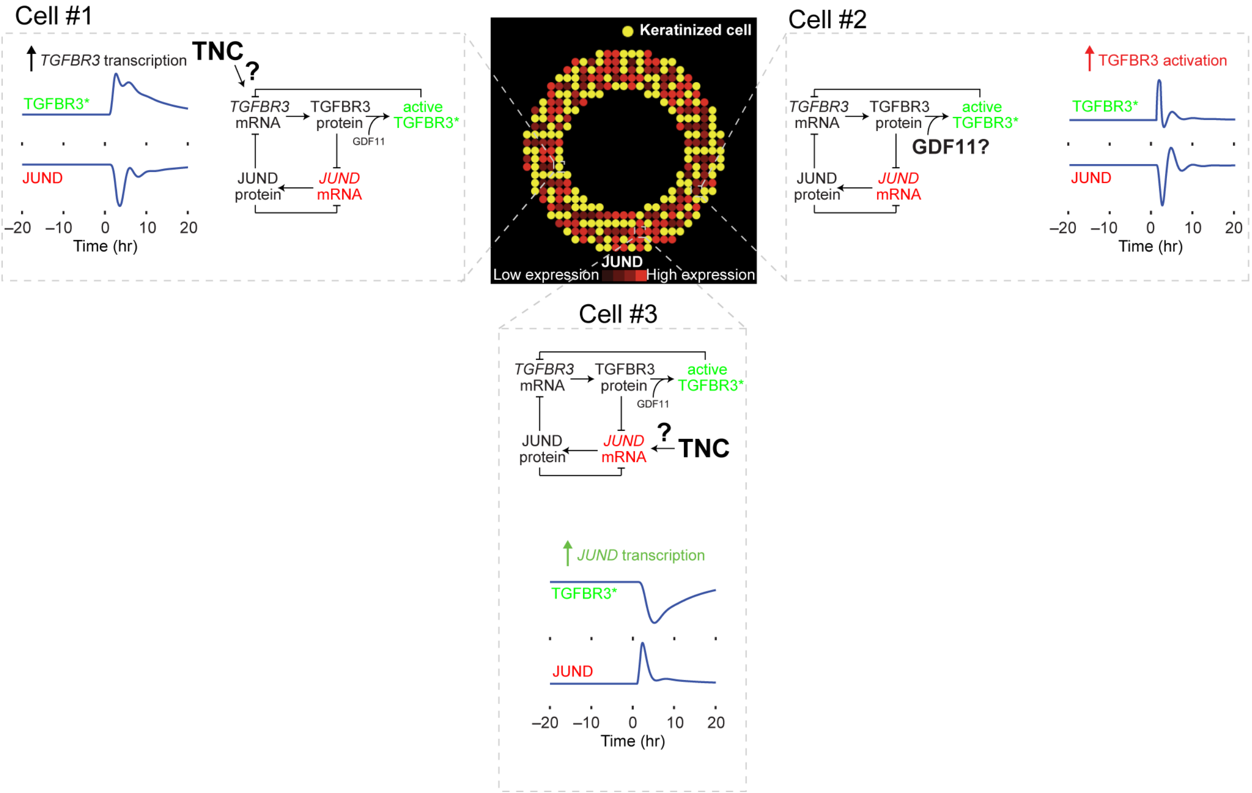


Figure 6.2 A multiscale model of JUND and TNC regulation in human basal-like breast cancer. Computational models were used in Chapter 3 to characterize JUND heterogeneity at the single cell (ODE) and multicellular (agent-based) levels. Future extensions of these models could focus on integrating the two scales of models into one cogent story of TNC and JUND regulation and the effect of keratinization in basal-like breast cancer biology.

6.2.3 Chapter 4 future work

In Chapter 4, we developed a maximum-likelihood approach to infer parameters of single-cell expression from the 10-cell stochastic transcriptional profiles. We validated this approach by inferring parameters from the original stochastic profiling dataset and experimentally measuring a test parameter with RNA FISH. We provisionally applied this approach to identify a rare transcriptional program. Remarkably, one gene from this program, *PIK3CD*, was important to the normal proliferation arrest of the MCF10A-5E breast spheroids, despite its rare expression (96).

Quantifying the single-cell expression distribution has many downstream applications. For example, drug resistance continues to be a challenge in cancer research (264). Cellular heterogeneity has been heavily implicated in contributing to drug resistance (54, 265). A prevailing question regarding drug resistance is where do these resistant clones come from? The resistant clones could either be already present in the tumor population before treatment, or treatment is acquired as a result of selective pressure (266). Intriguingly, the cellular population post treatment is sparse, implying that, regardless of the hypothesis, resistance is a rare characteristic in the tumor. Rare events are particularly challenging to capture with conventional techniques (see Section 1.3), which makes stochastic profiling a uniquely powerful tool to identify the regulatory mechanisms behind drug resistance.

Stochastic profiling can identify regulatory heterogeneities present in as few as ~1% of the cellular population (72, 96), the same order of magnitude as the percentage of drug resistant cells (134). To test which hypothesis of drug resistance is correct, we can stochastically sample cancer cells before and *during* treatment before the cells start dying (Figure 6.3). Since the percentage of resistant cells post-treatment is known for a given drug and cell line (or is easily measured), we can use the maximum-likelihood approach to identify gene regulatory programs with a matching expression frequency. We can scour the entire transcriptome for genes specifically upregulated in ~1-5% of the cells. These genes would be likely candidates that mediate drug resistance. If drug resistance were innate within the cell line, we would detect rarely expressed genes both before and during drug treatment. Additionally, the overlap between the genes identified would be the strongest candidate genes involved in resistance. Conventional techniques could not provide the same candidates, as the measurements would average out the rare behavior. Conversely, if resistance were acquired, we would expect to detect genes that correlate with resistance frequency in the samples collecting during drug treatment. Importantly, conventional techniques could not provide these candidates as well because we would have to rely on comparing molecular signatures pre- and post-treatment. The post-treatment signature will be a consequence of acquiring resistance and not necessarily the signature required to acquire resistance in the first place.

Identifying regulators of incompletely penetrant phenotypes is a challenging problem (22). To use conventional approaches, a pre- and post-phenotype strategy would be required, leading to identifying consequences rather than the causes of the phenotype. Stochastic profiling coupled with maximum-likelihood inference can uniquely address this problem, and can do so without the need to develop new techniques.

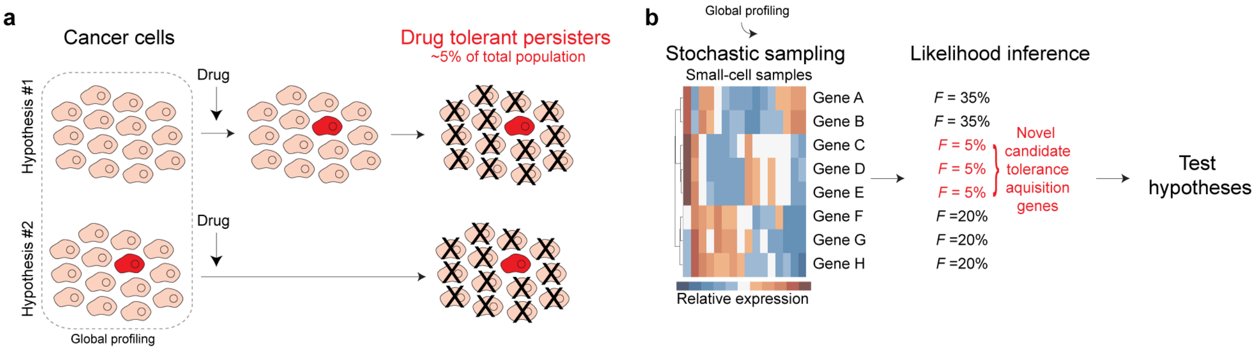


Figure 6.3 Using maximum-likelihood estimation to identify regulators of drug resistance. The acquisition of drug resistance remains a challenge in cancer research. However, the drug-resistant population may exist in the population before the drug is even added, suggesting intrinsic heterogeneity in cancers is responsible for drug resistance. Stochastic profiling coupled with maximum-likelihood estimation is uniquely poised to address these hypotheses. Cancer cells are profiled before and during drug treatment. Maximum-likelihood inference identifies gene expression programs with an expression frequency equivalent to the proportion of resistant cells. Identifying these programs before and during treatment could provide a stringent candidate list to test with further experimental follow up.

6.2.4 Chapter 5 future work

In Chapter 5, we identified a role of GDF11 in normal and abnormal breast epithelial morphogenesis in promoting compacted tissue architecture. Strikingly, GDF11 was posttranslationally misregulated in basal-like breast cancer cell lines and clinical specimens. The work in this Chapter suggested that GDF11 could be used as a therapy or biomarker for basal-like breast cancers.

We established a pilot cohort of patients to screen GDF11 immunoreactivity in human specimens of normal and triple-negative breast cancers. We observed an increase in 1E6 monoclonal antibody staining in a subset of premalignancies and virtually all advanced triple-negative tumors. These data correlated defects in GDF11 secretion, as measured by increased 1E6 foci, to tumor progression. Importantly, in parallel with accessing patient specimens, we can gain access to the clinical records for each patient. To test if the presence and amount of 1E6 foci are truly indicators of tumor behavior, we correlate focal counts with tumor grade, stage, and ultimately, patient outcome. The prediction would be that advanced triple-negative tumors have more 1E6 reactive foci and have worse clinical outcomes. Additionally, we would expect the subset of premalignancies that display 1E6 foci to eventually present with invasive ductal carcinoma if they did not receive surgery. If the results follow the hypothesis, GDF11 secretion would present the first specific immunohistochemical biomarker for assessing triple-negative breast cancer prognosis, particularly for early lesions.

GDF11 may also have therapeutic benefit in treating human basal-like breast cancers. Our data suggests that GDF11 prevents tumor seeding and decreases proliferation. GDF11 may be a potent tumor static treatment and prevent tumor progression and metastases (tumor seeding). However, GDF11 has a fast half-life in the circulation (206), suggesting that the ligand would be very difficult to dose clinically. Moving forward, modifications to GDF11 could increase the half-life and increase the bioavailability of GDF11 (267). Alternatively, small molecular libraries could be screened to identify activin receptor type IIB agonists. This same set of libraries could be screened to identify molecules that upregulate *ID2*. Increasing GDF11 expression through drug treatment has been reported (268), but our data suggests this approach would not work, as the tumor cells would still not be able to secrete the ligand. Drugs, however, could be identified that correct the secretion defect of GDF11 in breast cancer cells. Identifying a clinically viable option for inducing GDF11 phenotypes could present the first targeted therapy for basal-like breast cancer.

GDF11 function may not be limited to basal-like breast cancers. For example, we have anecdotally observed that EPR immunoreactivity is high in normal human pancreas. Interestingly, the pattern of expression is highly analogous to the normal breast tissue, where lobular tissue stained stronger for EPR than the ductal tissue (Figure 6.4). Additionally, pancreatic cancer cell lines display a similar heterogeneity in *GDF11* expression (105). These correlations may suggest that GDF11 acts globally to promote lobular-like phenotypes of epithelial cells in multiple organs.

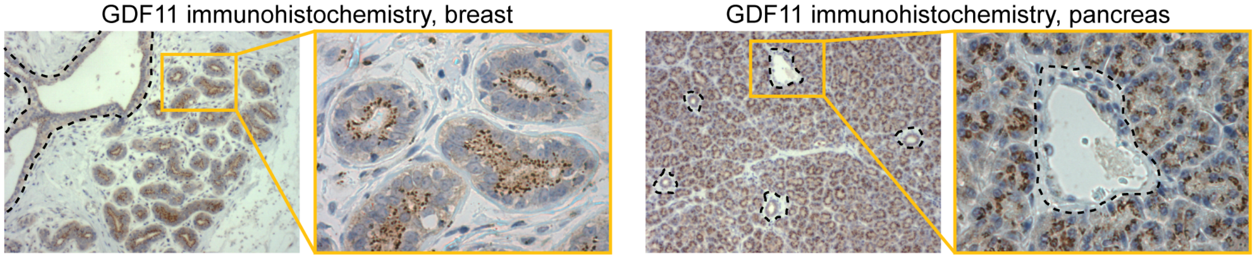

GDF11 immunohistochemistry, breast
GDF11 immunohistochemistry, pancreas

Figure 6.4 Expression of GDF11 in normal human breast and pancreas.

GDF11 expression measured by EPR immunoreactivity is analogous in human breast and human pancreas. Lobular units show high intracellular GDF11, while ductal tissue shows comparatively lower expression of GDF11.

6.3 Future outlook

The work in this Dissertation demonstrates the functional importance of the heterogeneous regulation of several genes to spheroid morphogenesis and human basal-like breast cancers. With stochastic profiling, we can globally identify heterogeneously regulated transcripts, but significant experimental follow up is required to identify function and importance. The workflow used in this Dissertation limited the functional analysis to a handful of genes and only in-depth analysis of one, *GDF11*. Developing methods to rapidly and reliably screen candidates from a stochastic profiling experiment will greatly enhance our understanding of heterogeneity.

Screening methods are strongly positioned to uncover functional importance to the panel of candidate regulatory heterogeneities genes during spheroid morphogenesis. Extending previously designed high throughput in vivo assays with single-cell overexpression could rapidly identify genes that alter morphogenesis when heterogeneously upregulated (269, 270). Coupling these experimental approaches with the work presented in Chapter 2, we can quickly arrive at the key set of genes whose heterogeneity has functional importance in spheroid morphogenesis. Intriguingly, these analyses may identify pathways or interacting genes that could begin to assemble the picture of how heterogeneities influence each other towards a biological outcome.

While genomic diversity and heterogeneity has been catalogued in human cancers, the extent of transcriptional heterogeneity is unknown. To exclude genomic differences, stochastic profiling should be first applied to xenograft models. Importantly, heterogeneity in xenograft models has been shown to have functional importance in the progression of the tumors (271). Stochastic profiling can catalogue regulatory differences in this "simple" context as proof of concept for exploring regulatory heterogeneity in human cancers. Before exerting the effort to stochastically profile human tumors, the technique should be extended to allow for genomic sequencing of the small-cell sample. A recent study profiled clinical material for both DNA and RNA (272); however, these analyses were performed on bulk samples. If stochastic profiling can be optimized to successfully purify the DNA and RNA from the same small cell samples, we can comprehensively catalogue the genomic (e.g., mutational) and transcriptional states within tumors. Importantly, the paired analysis would allow us to separate or correlate genomic heterogeneity from transcriptional regulatory heterogeneity with high fidelity.

6.4 Concluding remarks

By embracing the difficulties heterogeneity presents in biological research, we have developed a suite of tools that can address a multitude of questions that conventional approaches cannot. We started with a simple in vitro model of heterogeneity, identified a set of heterogeneously regulated transcripts, and discovered important roles for those transcripts in normal breast epithelial biology. Unexpectedly, those roles propagated from in vitro morphogenesis to the development of human basal-like breast cancers. Additionally, when the roles of these genes were different in cancers, they pointed to new ways in which cancer cells can seed tumors or avoid endogenous tumor suppressors. This Dissertation demonstrates that with the proper toolbox, we can begin to understand how human cancers work down to the single cell.

7 Methods

7.1 Chapter 2 Methods

7.1.1 Cell lines and 3D culture

Cell lines were obtained from the ATCC and maintained according to their specifications. 3D culture was performed as previously described (98). Cells were seeded at either 5,000 or 7,500 cells per well. 3D cultures were cultured in MCF10A-5E assay media or the growth media of the cell line. 3D cultures were reefed every four days.

7.1.2 Digital image acquisition

Digital images were acquired at 4x magnification with QCapture software using a QColor camera attached to an Olympus CKX41 inverted microscope. Chamber slides were placed onto a slide adapter for the microscope stage. The phase ring is removed from the optical path and the aperture stop is increased to provide contrast for the spheroids. Acquired images were imported into MATLAB segmentation routine (below).

7.1.3 Spheroid segmentation software

```
function varargout = adaptiveSegmentation(varargin)
% ADAPTIVESEGMENTATION MATLAB code for adaptiveSegmentation.fig
%      ADAPTIVESEGMENTATION, by itself, creates a new ADAPTIVESEGMENTATION or
raises the existing
%      singleton*.
%
%      H = ADAPTIVESEGMENTATION returns the handle to a new
ADAPTIVESEGMENTATION or the handle to
%      the existing singleton*.
%
%      ADAPTIVESEGMENTATION('CALLBACK',hObject,eventData,handles,...) calls
the local
```

```
%      function named CALLBACK in ADAPTIVESEGMENTATION.M with the given input
arguments.
%
%              ADAPTIVESEGMENTATION('Property','Value',...)  creates  a  new
ADAPTIVESEGMENTATION or raises the
%      existing singleton*.  Starting from the left, property value pairs are
%      applied to the GUI before adaptiveSegmentation_OpeningFcn gets called.
An
%      unrecognized property name or invalid value makes property application
%       stop.  All inputs are passed to adaptiveSegmentation_OpeningFcn via
varargin.
%
%      *See GUI Options on GUIDE's Tools menu.  Choose "GUI allows only one
%      instance to run (singleton)".
%
% See also: GUIDE, GUIDATA, GUIHANDLES

% Edit the above text to modify the response to help adaptiveSegmentation
global globalStruct
% Last Modified by GUIDE v2.5 23-Dec-2015 22:38:50

% Begin initialization code - DO NOT EDIT
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
    'gui_Singleton',  gui_Singleton, ...
    'gui_OpeningFcn', @adaptiveSegmentation_OpeningFcn, ...
    'gui_OutputFcn',  @adaptiveSegmentation_OutputFcn, ...
    'gui_LayoutFcn',  [] , ...
    'gui_Callback',   []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end

if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end
% End initialization code - DO NOT EDIT


% --- Executes just before adaptiveSegmentation is made visible.
function    adaptiveSegmentation_OpeningFcn(hObject,    eventdata,    handles,
varargin)
% This function has no output args, see OutputFcn.
% hObject    handle to figure
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
% varargin   command line arguments to adaptiveSegmentation (see VARARGIN)

% Choose default command line output for adaptiveSegmentation
handles.output = hObject;

% Update handles structure
guidata(hObject, handles);

% UIWAIT makes adaptiveSegmentation wait for user response (see UIRESUME)
% uiwait(handles.figure1);
% --- Outputs from this function are returned to the command line.
function   varargout   =   adaptiveSegmentation_OutputFcn(hObject,   eventdata,
handles)
% varargout  cell array for returning output args (see VARARGOUT);
% hObject    handle to figure
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Get default command line output from handles structure

varargout{1} = handles.output;

function edit1_Callback(hObject, eventdata, handles)
% hObject    handle to edit1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Hints: get(hObject,'String') returns contents of edit1 as text
%        str2double(get(hObject,'String')) returns contents of edit1 as a
double


% --- Executes during object creation, after setting all properties.
function edit1_CreateFcn(hObject, eventdata, handles)
% hObject    handle to edit1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns called

% Hint: edit controls usually have a white background on Windows.
%       See ISPC and COMPUTER.
if         ispc         &&         isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end



function edit2_Callback(hObject, eventdata, handles)
% hObject    handle to edit2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Hints: get(hObject,'String') returns contents of edit2 as text
%        str2double(get(hObject,'String')) returns contents of edit2 as a
double


% --- Executes during object creation, after setting all properties.
function edit2_CreateFcn(hObject, eventdata, handles)
% hObject    handle to edit2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns called

% Hint: edit controls usually have a white background on Windows.
%       See ISPC and COMPUTER.
```

```
if          ispc          &&          isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

% --- Executes on selection change in popupmenu1.
function popupmenu1_Callback(hObject, eventdata, handles)
% hObject    handle to popupmenu1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
%Read image
contents = get(hObject,'Value');
switch contents
    case 1
        rgb = imread('5E_D8_GDF11_0303.tiff');
    case 2
        rgb = imread('231_D16_notx_0303.tiff');
    case 3
        rgb = imread('BT20_D12_assay_notx_2.tiff');
    case 4
        rgb = imread('HCC1937_D12_growth_notx_1.tiff');
    case 5
        rgb = imread('HCC_D16_growth_notx_4.tiff');
    case 6
        rgb = imread('Hs578T_D12_assay_notx_2.tiff');
    case 7
        rgb = imread('shID2v2_D8_dox_gdf11_2.tiff');
    otherwise
        disp('error');
end
I = rgb2gray(rgb);
handles.I = I;
guidata(hObject,handles);
%Display raw image in GUI
axes(handles.axes1);
imshow(I,[])

%  Hints:  contents  =  cellstr(get(hObject,'String'))  returns  popupmenu1
contents as cell array
%        contents{get(hObject,'Value')} returns selected item from popupmenu1

% --- Executes during object creation, after setting all properties.
function popupmenu1_CreateFcn(hObject, eventdata, handles)
% hObject    handle to popupmenu1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns called

% Hint: popupmenu controls usually have a white background on Windows.
%       See ISPC and COMPUTER.
if          ispc          &&          isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end



function edit3_Callback(hObject, eventdata, handles)
% hObject    handle to edit3 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Hints: get(hObject,'String') returns contents of edit3 as text
%        str2double(get(hObject,'String')) returns contents of edit3 as a
double


% --- Executes during object creation, after setting all properties.
function edit3_CreateFcn(hObject, eventdata, handles)
% hObject    handle to edit3 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns called

% Hint: edit controls usually have a white background on Windows.
%       See ISPC and COMPUTER.
if      ispc      &&      isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end


% --- Executes on slider movement.
function slider1_Callback(hObject, eventdata, handles)
% hObject    handle to slider1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
% Hints: get(hObject,'Value') returns position of slider
%        get(hObject,'Min') and get(hObject,'Max') to determine range of
slider

selected = get(handles.listbox2,'Value');
if isfield(handles,'centersDeleted')
    handles                                                             =
rmfield(handles,{'centersDeleted','diametersDeleted','signatureDeleted','ccDe
leted'});
end
%Check for image load
checkImage = isfield(handles,'bulkI');
%Check for radio button
radioCheck1 = get(handles.radiobutton1,'Value');
radioCheck2 = get(handles.radiobutton2,'Value');
radioCheck3 = get(handles.radiobutton3,'Value');

if checkImage == 0
    set(hObject,'Value',0);
    mb1 = msgbox('Please select an image file');
    t = timer;
    t.StartDelay = 3;
    t.TimerFcn = @(~,~) delete(mb1);
    start(t)
    % elseif radioCheck1 == 0 && radioCheck2 == 0 && radioCheck3 == 0
    %     mb2 = msgbox('Please select a cell size');
    %     t = timer;
    %     t.StartDelay = 3;
    %     t.TimerFcn = @(~,~) delete(mb2);
    %     start(t)
elseif isfield(handles,'coloredLabels')
```

```
    fudgeFactor = get(hObject, 'Value');

    I = handles.I;

    %Opening and closing by reconstruction

    %Create structure element
    se = strel('disk', 5);

    %Erode and open
    Ie = imerode(I, se);
    Iobr = imreconstruct(Ie, I);

    %Opening-closing by reconstruction(Iobrcbr)
    Iobrd = imdilate(Iobr, se);
    Iobrcbr = imreconstruct(imcomplement(Iobrd), imcomplement(Iobr));
    Iobrcbr = imcomplement(Iobrcbr);

    % %border
    Iobrcbr(1:end,1) = Iobrcbr([2 2:end],2);
    Iobrcbr(1:end,end) = Iobrcbr([2 2:end],end-1);
    Iobrcbr(1,1:end) = Iobrcbr(2,[2 2:end]);
    Iobrcbr(end,1:end) = Iobrcbr(end-1,[2 2:end]);

    %Threshold
    ws = handles.ws;
    small = handles.small;
    bwEdge = adaptivethreshold(Iobrcbr,ws,fudgeFactor);
    %bwEdge = adaptivethreshold(Iobrcbr,[50 50]);
    edgeBetter = bwareaopen(bwEdge, small);
    bw = imclearborder(edgeBetter,4);
    %figure, imshow(bwnobord,[])
    % props = regionprops(bwnobord,'all');
    % boxProps = {props.BoundingBox};
    % bw = zeros(size(I,1),size(I,2));

    % for i = 1:length(boxProps)
    %     w = round(boxProps{i}(3));
    %     h = round(boxProps{i}(4));
    %     x = round(boxProps{i}(1));
    %     y = round(boxProps{i}(2));
    %
    %                                            bw(y:(y+h),x:(x+w))         =
imcomplement(im2bw(imsharpen(I(y:(y+h),x:(x+w))),graythresh(I(y:(y+h),x:(x+w)
))));
    %
    % end

    %remove small nonsense again
    bwsmall = bwareaopen(bw, small);

    %fill any remaining holes
    bwfinal = bwfill(bwsmall,'holes');
    %figure, imshow(bwfinal,[])

    %Add color labels
    CC = handles.CC;
    CC{selected} = bwconncomp(bwfinal);
    handles.CC = CC;
    CClabel = labelmatrix(CC{selected});
    coloredLabels = handles.coloredLabels;
    coloredLabels{selected} = label2rgb(CClabel, 'hsv', 'k', 'shuffle');
    handles.coloredLabels = coloredLabels;
    %Display segmented image in GUI
    axes(handles.axes2);
    imshow(coloredLabels{selected},[])

    recenters = handles.centers;
    diameters = handles.diameter;
    %Label numbers
    statsBW = regionprops(bwfinal, 'all');
    centers = [statsBW.Centroid];
    recenters{selected} = reshape(centers, 2, length(statsBW))';
    handles.centers = recenters;

    %     for i = 1:length(statsBW)
    %                text(recenters{selected}(i,1),  recenters{selected}(i,2),
num2str(i), 'color', 'w', 'FontSize', 10);
    %     end


    Area = [statsBW.Area];
    MajorAxisLength = [statsBW.MajorAxisLength];
    MinorAxisLength = [statsBW.MinorAxisLength];
    Eccentricity = [statsBW.Eccentricity];
    Orientation = [statsBW.Orientation];
    ConvexArea = [statsBW.ConvexArea];
    EquivDiameter = [statsBW.EquivDiameter];
    diameters{selected} = EquivDiameter;
    handles.diameter = diameters;
    Solidity = [statsBW.Solidity];
    Extent = [statsBW.Extent];
    Perimeter = [statsBW.Perimeter];
    p2a = Perimeter./Area;
    %zernike moments
    boundingBox = {statsBW.BoundingBox};
    Z = zeros(1,length(boundingBox));
    A = zeros(1,length(boundingBox));
    Phi = zeros(1,length(boundingBox));
    meanTotalPixels = zeros(1,length(boundingBox));
    stdTotalPixels = zeros(1,length(boundingBox));
    for i = 1:length(boundingBox)
        p = imcrop(bwfinal,boundingBox{i});
        if boundingBox{i}(3) > boundingBox{i}(4)
            p2                    =                   zeros(boundingBox{i}(3)-
boundingBox{i}(4),boundingBox{i}(3)+1);
            p = [p2;p];
        else
            p2          =         zeros(boundingBox{i}(4)+1,boundingBox{i}(4)-
boundingBox{i}(3));
            p = [p2 p];
        end
        [Z(i),A(i),Phi(i)] = Zernikmoment(p,4,2);
        %pixel characteristics
        mask = double(CClabel == i);
        totalCell = double(I).*mask;
        totalCellPixels = totalCell(mask == 1);
        meanTotalPixels(i) = mean(totalCellPixels);
        stdTotalPixels(i) = std(totalCellPixels);
    end

    signature = handles.signature;

    signature{selected}    =    {Area;    MajorAxisLength;    MinorAxisLength;
Eccentricity;  Orientation;  ConvexArea;  EquivDiameter;  Solidity;  Extent;
Perimeter; Z; A; Phi; meanTotalPixels; stdTotalPixels; p2a};
    handles.signature = signature;

    imgMasked = handles.imgMasked;
    imgMasked{selected} = imoverlay(I,bwfinal);
    handles.imgMasked = imgMasked;
    axes(handles.axes3)
    hold on
    axesChild = imshow(imgMasked{selected},[]);
    handles.axesChild = axesChild;
    set(axesChild,                                          'ButtonDownFcn',
{@MyCustomAxesButtonDownFunction,handles});

    %Store FudgeFactor
    thresh = handles.thresh;
    thresh{selected} = fudgeFactor;
    handles.thresh = thresh;

    guidata(hObject,handles);
else
    set(hObject,'Value',0);
end

% --- Executes during object creation, after setting all properties.
function slider1_CreateFcn(hObject, eventdata, handles)
% hObject    handle to slider1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns called

% Hint: slider controls usually have a light gray background.
if                                  isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor',[.9 .9 .9]);
end


% --- Executes on button press in pushbutton4.
function pushbutton4_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton4 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

%Check for image load
checkImage = isfield(handles,'bulkI');
%Check for radio button
radioCheck1 = get(handles.radiobutton1,'Value');
```

```
radioCheck2 = get(handles.radiobutton2,'Value');
radioCheck3 = get(handles.radiobutton3,'Value');
if checkImage == 0
    mb1 = msgbox('Please select an image file');
    t = timer;
    t.StartDelay = 3;
    t.TimerFcn = @(~,~) delete(mb1);
    start(t)
    % elseif radioCheck1 == 0 && radioCheck2 == 0 && radioCheck3 == 0
    %     mb2 = msgbox('Please select a cell size');
    %     t = timer;
    %     t.StartDelay = 3;
    %     t.TimerFcn = @(~,~) delete(mb2);
    %     start(t)
else

    bulkI = handles.bulkI;
    ws = handles.ws;
    small = handles.small;
    %Start waitbar
    h = waitbar(0,'Please wait...Segmenting Cells');
    for i = 1:length(bulkI)
        % slider position
        tempThresh = get(hObject, 'Value');

        I = bulkI{i};

        %Opening and closing by reconstruction

        %Create structure element
        se = strel('disk', 5);

        %Erode and open
        Ie = imerode(I, se);
        Iobr = imreconstruct(Ie, I);

        %Opening-closing by reconstruction(Iobrcbr)
        Iobrd = imdilate(Iobr, se);
        Iobrcbr = imreconstruct(imcomplement(Iobrd), imcomplement(Iobr));
        Iobrcbr = imcomplement(Iobrcbr);

        % %border
        Iobrcbr(1:end,1) = Iobrcbr([2 2:end],2);
        Iobrcbr(1:end,end) = Iobrcbr([2 2:end],end-1);
        Iobrcbr(1,1:end) = Iobrcbr(2,[2 2:end]);
        Iobrcbr(end,1:end) = Iobrcbr(end-1,[2 2:end]);

        %Threshold
        [bwEdgeFalse, thresh] = adaptivethreshold(Iobrcbr,ws,tempThresh);
        threshAll{i} = thresh/2;
        handles.thresh = threshAll;
        guidata(hObject,handles);
        bwEdge = adaptivethreshold(Iobrcbr,ws,thresh/2);
        set(handles.slider1, 'Value', thresh/2);

        %bwEdge = adaptivethreshold(Iobrcbr,[50 50]);
        edgeBetter = bwareaopen(bwEdge, small);
        bw = imclearborder(edgeBetter,4);
        %figure, imshow(bwnobord,[])
        % props = regionprops(bwnobord,'all');
        % boxProps = {props.BoundingBox};
        % bw = zeros(size(I,1),size(I,2));

        % for i = 1:length(boxProps)
        %     w = round(boxProps{i}(3));
        %     h = round(boxProps{i}(4));
        %     x = round(boxProps{i}(1));
        %     y = round(boxProps{i}(2));
        %
        %                                         bw(y:(y+h),x:(x+w))        =
imcomplement(im2bw(imsharpen(I(y:(y+h),x:(x+w))),graythresh(I(y:(y+h),x:(x+w)
))));
        %
        % end

        %remove small nonsense again
        bwsmall = bwareaopen(bw, small);

        %fill any remaining holes
        bwfinal = bwfill(bwsmall,'holes');
        %figure, imshow(bwfinal,[])

        %Add color labels
        CC{i} = bwconncomp(bwfinal);
        handles.CC = CC;
        CClabel = labelmatrix(CC{i});
        coloredLabels{i} = label2rgb(CClabel, 'hsv', 'k', 'shuffle');


        %Label numbers
        statsBW = regionprops(bwfinal, 'all');
        centers = [statsBW.Centroid];
        recenters{i} = reshape(centers, 2, length(statsBW))';
        handles.centers = recenters;
        % centers{i} = reshape(centers{i}, 2, length(statsBW))';
        %
        % for j = 1:length(statsBW)
        %        text(centers{i}(j,1),  centers{i}(j,2),  num2str(j),  'color',
'w', 'FontSize', 8);
        % end

        Area = [statsBW.Area];
        MajorAxisLength = [statsBW.MajorAxisLength];
        MinorAxisLength = [statsBW.MinorAxisLength];
        Eccentricity = [statsBW.Eccentricity];
        Orientation = [statsBW.Orientation];
        ConvexArea = [statsBW.ConvexArea];
        EquivDiameter = [statsBW.EquivDiameter];
        handles.diameter{i} = EquivDiameter;
        Solidity = [statsBW.Solidity];
        Extent = [statsBW.Extent];
        Perimeter = [statsBW.Perimeter];
        p2a = Perimeter./Area;


        %zernike moments
        boundingBox = {statsBW.BoundingBox};
        Z = zeros(1,length(boundingBox));
        A = zeros(1,length(boundingBox));
        Phi = zeros(1,length(boundingBox));
        meanTotalPixels = zeros(1,length(boundingBox));
        stdTotalPixels = zeros(1,length(boundingBox));

        for j = 1:length(boundingBox)
            p = imcrop(bwfinal,boundingBox{j});
            if boundingBox{j}(3) > boundingBox{j}(4)
                p2                 =                  zeros(boundingBox{j}(3)-
boundingBox{j}(4),boundingBox{j}(3)+1);
                p = [p2;p];
            else
                p2       =       zeros(boundingBox{j}(4)+1,boundingBox{j}(4)-
boundingBox{j}(3));
                p = [p2 p];
            end
            [Z(j),A(j),Phi(j)] = Zernikmoment(p,4,2);

            %pixel characteristics
            mask = double(CClabel == j);
            totalCell = double(I).*mask;
            totalCellPixels = totalCell(mask == 1);
            meanTotalPixels(j) = mean(totalCellPixels);
            stdTotalPixels(j) = std(totalCellPixels);

        end

        signature{i} = {Area; MajorAxisLength; MinorAxisLength; Eccentricity;
Orientation;  ConvexArea;  EquivDiameter;  Solidity;  Extent;  Perimeter;  Z;  A;
Phi; meanTotalPixels; stdTotalPixels; p2a};

        imgMasked{i} = imoverlay(I,bwfinal);

        guidata(hObject,handles);

        %update waitbar
        waitbar(i/length(bulkI))
    end

    %Store signature cell
    handles.signature = signature;
    %Store segmented images
    handles.coloredLabels = coloredLabels;
    %Store overlayed images
    handles.imgMasked = imgMasked;

    %Store last image in set
    handles.I = I;


    set(handles.listbox2,'Value',length(bulkI))

    %Display original image in GUI
    axes(handles.axes1);
    imshow(bulkI{end},[]);
    %Display segmented image in GUI
    axes(handles.axes2);
    imshow(coloredLabels{end},[]);

    %Display overlayed image in GUI
    axes(handles.axes3);
    hold on
    axesChild = imshow(imgMasked{end},[]);
    handles.axesChild = axesChild;
    set(axesChild,                                            'ButtonDownFcn',
{@MyCustomAxesButtonDownFunction,handles});
    %Close waitbar
    close(h);
    guidata(hObject,handles)
end


% --- Executes during object creation, after setting all properties.
function uipanel2_CreateFcn(hObject, eventdata, handles)
% hObject    handle to uipanel2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns called

% --- Executes when selected object is changed in uipanel2.
function uipanel2_SelectionChangeFcn(hObject, eventdata, handles)
% hObject    handle to the selected object in uipanel2
% eventdata  structure with the following fields (see UIBUTTONGROUP)
%     EventName: string 'SelectionChanged' (read only)
%     OldValue: handle of the previously selected object or empty if none was
selected
%     NewValue: handle of the currently selected object
% handles    structure with handles and user data (see GUIDATA)

%Check for image load
checkImage = isfield(handles,'bulkI');
if checkImage == 0
    mb = msgbox('Please select an image file');
    set(handles.radiobutton1,'Value',0);
    set(handles.radiobutton2,'Value',0);
    set(handles.radiobutton3,'Value',0);
    t = timer;
    t.StartDelay = 3;
    t.TimerFcn = @(~,~) delete(mb);
    start(t)
else

    switch get(eventdata.NewValue,'Tag') % Get Tag of selected object.
        case 'radiobutton1'
            % Code for when radiobutton1 is selected.
            handles.ws = [100 100];
            handles.small = 100;
            guidata(hObject,handles);
            set(handles.slider2,'Value',100);
            %             handles.ws = [150 150];
            %             handles.small = 300;
        case 'radiobutton2'
            % Code for when radiobutton2 is selected.
            handles.ws = [250 250];
            handles.small = 250;
            guidata(hObject,handles);
            set(handles.slider2,'Value',250);
        case 'radiobutton3'
            % Code for when radiobutton3 is selected.
            handles.ws = [400 400];
            handles.small = 400;
            guidata(hObject,handles);
            set(handles.slider2,'Value',400);
        otherwise
    end
    %Check for image load
    checkImage = isfield(handles,'bulkI');
    %Check for radio button
    radioCheck1 = get(handles.radiobutton1,'Value');
    radioCheck2 = get(handles.radiobutton2,'Value');
    radioCheck3 = get(handles.radiobutton3,'Value');

    if checkImage == 0
        set(hObject,'Value',250);
        mb1 = msgbox('Please select an image file');
        t = timer;
        t.StartDelay = 3;
        t.TimerFcn = @(~,~) delete(mb1);
        start(t)
        % elseif radioCheck1 == 0 && radioCheck2 == 0 && radioCheck3 == 0
        %     set(hObject,'Value',250);
        %     mb2 = msgbox('Please select a cell size');
        %     t = timer;
        %     t.StartDelay = 3;
        %     t.TimerFcn = @(~,~) delete(mb2);
        %     start(t)
    elseif isfield(handles,'thresh')

        selected = get(handles.listbox2,'Value');

        I = handles.I;

        %Opening and closing by reconstruction

        %Create structure element
        se = strel('disk', 5);

        %Erode and open
        Ie = imerode(I, se);
        Iobr = imreconstruct(Ie, I);

        %Opening-closing by reconstruction(Iobrcbr)
        Iobrd = imdilate(Iobr, se);
        Iobrcbr = imreconstruct(imcomplement(Iobrd), imcomplement(Iobr));
        Iobrcbr = imcomplement(Iobrcbr);

        % %border
        Iobrcbr(1:end,1) = Iobrcbr([2 2:end],2);
        Iobrcbr(1:end,end) = Iobrcbr([2 2:end],end-1);
        Iobrcbr(1,1:end) = Iobrcbr(2,[2 2:end]);
        Iobrcbr(end,1:end) = Iobrcbr(end-1,[2 2:end]);
```

```
        %Threshold
        ws = handles.ws;
        small = handles.small;
        thresh = handles.thresh;
        bwEdge = adaptivethreshold(Iobrcbr,ws,thresh{selected});
        %bwEdge = adaptivethreshold(Iobrcbr,[50 50]);
        edgeBetter = bwareaopen(bwEdge, small);
        bw = imclearborder(edgeBetter,4);
        %figure, imshow(bwnobord,[])
        % props = regionprops(bwnobord,'all');
        % boxProps = {props.BoundingBox};
        % bw = zeros(size(I,1),size(I,2));

        % for i = 1:length(boxProps)
        %     w = round(boxProps{i}(3));
        %     h = round(boxProps{i}(4));
        %     x = round(boxProps{i}(1));
        %     y = round(boxProps{i}(2));
        %
        %                                         bw(y:(y+h),x:(x+w))        =
imcomplement(im2bw(imsharpen(I(y:(y+h),x:(x+w))),graythresh(I(y:(y+h),x:(x+w)
))));
        %
        % end

        %remove small nonsense again
        bwsmall = bwareaopen(bw, small);

        %fill any remaining holes
        bwfinal = bwfill(bwsmall,'holes');
        %figure, imshow(bwfinal,[])

        %Add color labels
        CC{selected} = bwconncomp(bwfinal);
        handles.CC = CC;
        CClabel = labelmatrix(CC{selected});
        coloredLabels = handles.coloredLabels;
        coloredLabels{selected} = label2rgb(CClabel, 'hsv', 'k', 'shuffle');
        handles.coloredLabels = coloredLabels;
        %Display segmented image in GUI
        axes(handles.axes2);
        imshow(coloredLabels{selected},[])

        recenters = handles.centers;
        diameters = handles.diameter;

        %Label numbers
        statsBW = regionprops(bwfinal, 'all');
        centers = [statsBW.Centroid];
        recenters{selected} = reshape(centers, 2, length(statsBW))';
        handles.centers = recenters;

        for i = 1:length(statsBW)
            text(recenters{selected}(i,1),           recenters{selected}(i,2),
num2str(i), 'color', 'w', 'FontSize', 8);
        end

        Area = [statsBW.Area];
        MajorAxisLength = [statsBW.MajorAxisLength];
        MinorAxisLength = [statsBW.MinorAxisLength];
        Eccentricity = [statsBW.Eccentricity];
        Orientation = [statsBW.Orientation];
        ConvexArea = [statsBW.ConvexArea];
        EquivDiameter = [statsBW.EquivDiameter];
        diameters{selected} = EquivDiameter;
        handles.diameter = diameters;
        Solidity = [statsBW.Solidity];
        Extent = [statsBW.Extent];
        Perimeter = [statsBW.Perimeter];
        p2a = Perimeter./Area;

        %zernike moments
        boundingBox = {statsBW.BoundingBox};
        Z = zeros(1,length(boundingBox));
        A = zeros(1,length(boundingBox));
        Phi = zeros(1,length(boundingBox));
        meanTotalPixels = zeros(1,length(boundingBox));
        stdTotalPixels = zeros(1,length(boundingBox));

        for j = 1:length(boundingBox)
            p = imcrop(bwfinal,boundingBox{j});
            if boundingBox{j}(3) > boundingBox{j}(4)
                p2                 =                 zeros(boundingBox{j}(3)-
boundingBox{j}(4),boundingBox{j}(3)+1);
                p = [p2;p];
            else
                p2       =       zeros(boundingBox{j}(4)+1,boundingBox{j}(4)-
boundingBox{j}(3));
                p = [p2 p];
            end
            [Z(j),A(j),Phi(j)] = Zernikmoment(p,4,2);

            %pixel characteristics
            mask = double(CClabel == j);
            totalCell = double(I).*mask;
            totalCellPixels = totalCell(mask == 1);
            meanTotalPixels(j) = mean(totalCellPixels);
            stdTotalPixels(j) = std(totalCellPixels);

        end


        signature = handles.signature;
        signature{selected}   =   {Area;   MajorAxisLength;   MinorAxisLength;
Eccentricity;  Orientation;  ConvexArea;  EquivDiameter;  Solidity;  Extent;
Perimeter; Z; A; Phi; meanTotalPixels; stdTotalPixels; p2a};
        handles.signature = signature;

        imgMasked = handles.imgMasked;
        imgMasked{selected} = imoverlay(I,bwfinal);
        handles.imgMasked = imgMasked;
        axes(handles.axes3)
        hold on
        axesChild = imshow(imgMasked{selected},[]);
        handles.axesChild = axesChild;
        set(axesChild,                                          'ButtonDownFcn',
{@MyCustomAxesButtonDownFunction,handles});


        guidata(hObject,handles);
    else
    end
end

% --------------------------------------------------------------------
function dropdown1_Callback(hObject, eventdata, handles)
% hObject    handle to dropdown1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)


% --------------------------------------------------------------------
function open_Callback(hObject, eventdata, handles)
% hObject    handle to open (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% --- Executes on selection change in listbox2.
function listbox2_Callback(hObject, eventdata, handles)
% hObject    handle to listbox2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Hints: contents = cellstr(get(hObject,'String')) returns listbox2 contents
as cell array
%        contents{get(hObject,'Value')} returns selected item from listbox2
set(handles.radiobutton1,'Value',0);
set(handles.radiobutton2,'Value',0);
set(handles.radiobutton3,'Value',0);
selected = get(hObject,'Value');

bulkI = handles.bulkI;


%Store temporary image for manual thresholding
handles.I = bulkI{selected};
handles.num = selected;
guidata(hObject,handles);

if isfield(handles,'coloredLabels') == 0
    axes(handles.axes1);
    imshow(bulkI{selected},[]);
else
    %Set slider bar to correct thresh
    thresh = handles.thresh;
    set(handles.slider1, 'Value', thresh{selected});
    %Load data
    coloredLabels = handles.coloredLabels;
    imgMasked = handles.imgMasked;
    %Display original image in GUI
    axes(handles.axes1);
    imshow(bulkI{selected},[]);
    %Display segmented image in GUI
    axes(handles.axes2);
    imshow(coloredLabels{selected},[])

    %                                handles                                =
rmfield(handles,{'CC','ccDeleted','centersDeleted','signatureDeleted','diamet
ersDeleted'});

    %Display overlayed image in GUI
    axes(handles.axes3)
    hold on
    axesChild = imshow(imgMasked{selected},[]);
    handles.axesChild = axesChild;
    set(axesChild,                                            'ButtonDownFcn',
{@MyCustomAxesButtonDownFunction,handles});


end
% --- Executes during object creation, after setting all properties.
function listbox2_CreateFcn(hObject, eventdata, handles)
% hObject    handle to listbox2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns called

% Hint: listbox controls usually have a white background on Windows.
%       See ISPC and COMPUTER.
if          ispc          &&          isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end


% --- Executes on slider movement.
function slider2_Callback(hObject, eventdata, handles)
% hObject    handle to slider2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Hints: get(hObject,'Value') returns position of slider
%           get(hObject,'Min')  and  get(hObject,'Max')  to  determine  range  of
slider
%Check for image load
checkImage = isfield(handles,'bulkI');
%Check for radio button
radioCheck1 = get(handles.radiobutton1,'Value');
radioCheck2 = get(handles.radiobutton2,'Value');
radioCheck3 = get(handles.radiobutton3,'Value');

if checkImage == 0
    set(hObject,'Value',250);
    mb1 = msgbox('Please select an image file');
    t = timer;
    t.StartDelay = 3;
    t.TimerFcn = @(~,~) delete(mb1);
    start(t)
    % elseif radioCheck1 == 0 && radioCheck2 == 0 && radioCheck3 == 0
    %     set(hObject,'Value',250);
    %     mb2 = msgbox('Please select a cell size');
    %     t = timer;
    %     t.StartDelay = 3;
    %     t.TimerFcn = @(~,~) delete(mb2);
    %     start(t)
elseif isfield(handles,'thresh')
    %           handles.ws           =           [round(get(hObject,'Value')),
round(get(hObject,'Value'))];
    handles.small = round(get(hObject,'Value'));
    guidata(hObject,handles);

    selected = get(handles.listbox2,'Value');

    I = handles.I;

    %Opening and closing by reconstruction

    %Create structure element
    se = strel('disk', 5);

    %Erode and open
    Ie = imerode(I, se);
    Iobr = imreconstruct(Ie, I);

    %Opening-closing by reconstruction(Iobrcbr)
    Iobrd = imdilate(Iobr, se);
    Iobrcbr = imreconstruct(imcomplement(Iobrd), imcomplement(Iobr));
    Iobrcbr = imcomplement(Iobrcbr);

    % %border
    Iobrcbr(1:end,1) = Iobrcbr([2 2:end],2);
    Iobrcbr(1:end,end) = Iobrcbr([2 2:end],end-1);
    Iobrcbr(1,1:end) = Iobrcbr(2,[2 2:end]);
    Iobrcbr(end,1:end) = Iobrcbr(end-1,[2 2:end]);

    %Threshold
    ws = handles.ws;
    small = handles.small;
    thresh = handles.thresh;
    bwEdge = adaptivethreshold(Iobrcbr,ws,thresh{selected});
    %bwEdge = adaptivethreshold(Iobrcbr,[50 50]);
    edgeBetter = bwareaopen(bwEdge, small);
    bw = imclearborder(edgeBetter,4);
    %figure, imshow(bwnobord,[])
    % props = regionprops(bwnobord,'all');
    % boxProps = {props.BoundingBox};
    % bw = zeros(size(I,1),size(I,2));

    % for i = 1:length(boxProps)
    %     w = round(boxProps{i}(3));
    %     h = round(boxProps{i}(4));
    %     x = round(boxProps{i}(1));
    %     y = round(boxProps{i}(2));
    %
    %                                        bw(y:(y+h),x:(x+w))         =
imcomplement(im2bw(imsharpen(I(y:(y+h),x:(x+w))),graythresh(I(y:(y+h),x:(x+w)
))));
    %
    % end
    %remove small nonsense again
    bwsmall = bwareaopen(bw, small);

    %fill any remaining holes
    bwfinal = bwfill(bwsmall,'holes');
    %figure, imshow(bwfinal,[])

    %Add color labels
    CC = handles.CC;
    CC{selected} = bwconncomp(bwfinal);
    handles.CC = CC;
    CClabel = labelmatrix(CC{selected});
    coloredLabels = handles.coloredLabels;
    coloredLabels{selected} = label2rgb(CClabel, 'hsv', 'k', 'shuffle');
    handles.coloredLabels = coloredLabels;
    %Display segmented image in GUI
    axes(handles.axes2);
    imshow(coloredLabels{selected},[])

    recenters = handles.centers;
    diameters = handles.diameter;
    %Label numbers
    statsBW = regionprops(bwfinal, 'all');
    centers = [statsBW.Centroid];
    recenters{selected} = reshape(centers, 2, length(statsBW))';
    handles.centers = recenters;

    for i = 1:length(statsBW)
        text(recenters{selected}(i,1),  recenters{selected}(i,2),  num2str(i),
'color', 'w', 'FontSize', 8);
    end

    Area = [statsBW.Area];
    MajorAxisLength = [statsBW.MajorAxisLength];
    MinorAxisLength = [statsBW.MinorAxisLength];
    Eccentricity = [statsBW.Eccentricity];
    Orientation = [statsBW.Orientation];
    ConvexArea = [statsBW.ConvexArea];
    EquivDiameter = [statsBW.EquivDiameter];
    diameters{selected} = EquivDiameter;
    handles.diameter = diameters;
    Solidity = [statsBW.Solidity];
    Extent = [statsBW.Extent];
    Perimeter = [statsBW.Perimeter];
    p2a = Perimeter./Area;

    %zernike moments
    boundingBox = {statsBW.BoundingBox};
    Z = zeros(1,length(boundingBox));
    A = zeros(1,length(boundingBox));
    Phi = zeros(1,length(boundingBox));
    meanTotalPixels = zeros(1,length(boundingBox));
    stdTotalPixels = zeros(1,length(boundingBox));

    for j = 1:length(boundingBox)
        p = imcrop(bwfinal,boundingBox{j});
        if boundingBox{j}(3) > boundingBox{j}(4)
```

```
            p2                    =                    zeros(boundingBox{j}(3)-
boundingBox{j}(4),boundingBox{j}(3)+1);
            p = [p2;p];
        else
            p2          =         zeros(boundingBox{j}(4)+1,boundingBox{j}(4)-
boundingBox{j}(3));
            p = [p2 p];
        end
        [Z(j),A(j),Phi(j)] = Zernikmoment(p,4,2);

        %pixel characteristics
        mask = double(CClabel == j);
        totalCell = double(I).*mask;
        totalCellPixels = totalCell(mask == 1);
        meanTotalPixels(j) = mean(totalCellPixels);
        stdTotalPixels(j) = std(totalCellPixels);

    end


    signature = handles.signature;
    signature{selected}   =   {Area;   MajorAxisLength;   MinorAxisLength;
Eccentricity;  Orientation;  ConvexArea;  EquivDiameter;  Solidity;  Extent;
Perimeter; Z; A; Phi; meanTotalPixels; stdTotalPixels; p2a};
    handles.signature = signature;


    imgMasked = handles.imgMasked;
    imgMasked{selected} = imoverlay(I,bwfinal);
    handles.imgMasked = imgMasked;
    axes(handles.axes3)
    hold on
    axesChild = imshow(imgMasked{selected},[]);
    handles.axesChild = axesChild;
    set(axesChild,                                          'ButtonDownFcn',
{@MyCustomAxesButtonDownFunction,handles});


    guidata(hObject,handles);
else
    set(hObject,'Value',handles.small);
end


% --- Executes during object creation, after setting all properties.
function slider2_CreateFcn(hObject, eventdata, handles)
% hObject    handle to slider2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns called
set(hObject,'Value',250);
handles.ws = [250,250];
handles.small = 250;
guidata(hObject,handles);
% Hint: slider controls usually have a light gray background.
if                                isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor',[.9 .9 .9]);
end
```

```
function edit4_Callback(hObject, eventdata, handles)
% hObject    handle to edit4 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Hints: get(hObject,'String') returns contents of edit4 as text
%        str2double(get(hObject,'String')) returns contents of edit4 as a
double


% --- Executes during object creation, after setting all properties.
function edit4_CreateFcn(hObject, eventdata, handles)
% hObject    handle to edit4 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns called

% Hint: edit controls usually have a white background on Windows.
%       See ISPC and COMPUTER.
if          ispc          &&          isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end


% --- Executes on key press with focus on figure1 and none of its controls.
function figure1_KeyPressFcn(hObject, eventdata, handles)
% hObject    handle to figure1 (see GCBO)
% eventdata  structure with the following fields (see FIGURE)
%     Key: name of the key that was pressed, in lower case
%     Character: character interpretation of the key(s) that was pressed
%     Modifier: name(s) of the modifier key(s) (i.e., control, shift) pressed
% handles    structure with handles and user data (see GUIDATA)
modifier = eventdata.Modifier;
key = eventdata.Key;
modCheck = strcmp(modifier,'control');
keyCheck = strcmp(key,'o');
if  modCheck && keyCheck
    newOpen_Callback(hObject,eventdata,handles)
end


% --- Executes on key press with focus on slider2 and none of its controls.
function slider2_KeyPressFcn(hObject, eventdata, handles)
% hObject    handle to slider2 (see GCBO)
% eventdata  structure with the following fields (see UICONTROL)
%     Key: name of the key that was pressed, in lower case
%     Character: character interpretation of the key(s) that was pressed
%     Modifier: name(s) of the modifier key(s) (i.e., control, shift) pressed
% handles    structure with handles and user data (see GUIDATA)
% key = double(get(gcf,'CurrentCharacter'));
% if key == 28
%     current = get(slider2,'Value');
%     step = get(slider2,'SliderStep');
%     set(slider2,'Value',current-step);
% end
```

```
function MyCustomAxesButtonDownFunction(hObject, eventData, handles)
if get(handles.togglebutton1,'Value') == 1

    selected = get(handles.listbox2,'Value');
    [coords] = get(handles.axes3,'CurrentPoint');

    centers = handles.centers{selected};
    diameters = handles.diameter{selected};

    coloredLabels = handles.coloredLabels{selected};
    imgMasked = handles.imgMasked{selected};
    CC = handles.CC;
    signature = handles.signature;

    if isfield(handles,'centersDeleted') == 0
        % centersDeleted = zeros(length(CC{selected}.PixelIdxList),2);
        % signatureDeleted = [];
        % diametersDeleted = zeros(1,length(CC{selected}.PixelIdxList));
        % ccDeleted = [];
        handles.centersDeleted = cell(1,length(CC));
        handles.signatureDeleted = cell(1,length(CC));
        handles.diametersDeleted = cell(1,length(CC));
        handles.ccDeleted = cell(1,length(CC));
        centersDeleted = zeros(length(CC{selected}.PixelIdxList),2);
        signatureDeleted = [];
        diametersDeleted = zeros(1,length(CC{selected}.PixelIdxList));
        ccDeleted = [];
    elseif isempty(handles.centersDeleted{selected}) == 1
        centersDeleted = zeros(length(CC{selected}.PixelIdxList),2);
        signatureDeleted = [];
        diametersDeleted = zeros(1,length(CC{selected}.PixelIdxList));
        ccDeleted = [];
    else
        centersDeleted = handles.centersDeleted{selected};
        diametersDeleted = handles.diametersDeleted{selected};
        signatureDeleted = handles.signatureDeleted{selected};
        ccDeleted = handles.ccDeleted{selected};
    end
    for i = 1:length(centers)

        if    ((coords(1,1)-centers(i,1))^2+(coords(1,2)-centers(i,2))^2)    <=
(diameters(i)/2)^2
            ccDeleted{i} = CC{selected}.PixelIdxList{i};
            CC{selected}.PixelIdxList{i} = [];
            CClabel = labelmatrix(CC{selected});
            coloredLabels = label2rgb(CClabel, 'hsv', 'k', 'shuffle');
            grayLabel = rgb2gray(coloredLabels);
            grayLabel(grayLabel > 0) = 1;
            %bwMock = im2bw(CClabel);
            bwMock = grayLabel;
            imgMasked = imoverlay(handles.I,bwMock);
            centersDeleted(i,:) = centers(i,:);
            centers(i,:) = NaN;

            for j = 1:length(signature{selected})
                signatureDeleted{j}{i} = signature{selected}{j}(i);
                signature{selected}{j}(i) = NaN;
            end
            diametersDeleted(i)= handles.diameter{selected}(i);
            handles.diameter{selected}(i) = NaN;
            break
        elseif              ((coords(1,1)-centersDeleted(i,1))^2+(coords(1,2)-
centersDeleted(i,2))^2) <= (diametersDeleted(i)/2)^2
            CC{selected}.PixelIdxList{i} = ccDeleted{i};
            CClabel = labelmatrix(CC{selected});
            coloredLabels = label2rgb(CClabel,'hsv','k','shuffle');
            grayLabel = rgb2gray(coloredLabels);
            grayLabel(grayLabel > 0) = 1;
            bwMock = grayLabel;
            imgMasked = imoverlay(handles.I,bwMock);
            centers(i,:) = centersDeleted(i,:);
            handles.diameter{selected}(i) = diametersDeleted(i);
            for j = 1:length(signature{selected})
                signature{selected}{j}(i) = signatureDeleted{j}{i};
            end
        end
    end
    handles.signature = signature;
    handles.ccDeleted{selected} = ccDeleted;
    handles.centersDeleted{selected} = centersDeleted;
    handles.signatureDeleted{selected} = signatureDeleted;
    handles.diametersDeleted{selected} = diametersDeleted;
    handles.CC = CC;
    handles.centers{selected} = centers;
    handles.coloredLabels{selected} = coloredLabels;
    handles.imgMasked{selected} = imgMasked;

    guidata(hObject,handles);

    axes(handles.axes2);
    imshow(coloredLabels,[]);
    axes(handles.axes3);
    hold on
    axesChild = imshow(imgMasked,[]);
    handles.axesChild = axesChild;
    set(axesChild,                                    'ButtonDownFcn',
{@MyCustomAxesButtonDownFunction,handles});
else
end

% --------------------------------------------------------------------
function newOpen_Callback(hObject, eventdata, handles)
% hObject    handle to newOpen (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

set(handles.slider2,'Value',250);
handles.ws = [250,250];
handles.small = 250;
guidata(hObject,handles);

checkImage = isfield(handles,'bulkI');
checkSegment = isfield(handles,'coloredLabels');
checkDeleted = isfield(handles,'ccDeleted');
```

```
if checkDeleted == 1
    handles                                                                 =
rmfield(handles,{'signature','ccDeleted','centersDeleted','signatureDeleted',
'diametersDeleted','CC','ws','small','bulkI','coloredLabels','imgMasked','thr
esh','centers','diameter'});
    guidata(hObject,handles);
elseif checkSegment == 1
    handles                                                                 =
rmfield(handles,{'signature','CC','ws','small','bulkI','coloredLabels','imgMa
sked','thresh'});
    guidata(hObject,handles);
elseif  checkImage == 1
    handles = rmfield(handles,'bulkI');
    guidata(hObject,handles);

end

%Clear all axes and listbox
cla(handles.axes1);
cla(handles.axes2);
cla(handles.axes3);
set(handles.listbox2,'String','');
set(handles.radiobutton1,'Value',0);
set(handles.radiobutton2,'Value',0);
set(handles.radiobutton3,'Value',0);



[filename,    baseName]    =    uigetfile('*.jpg;*.tiff;*.tif','Select    an
Image','Multiselect', 'on');
if isequal(filename,0)
elseif iscellstr(filename) == 0
    handles.filename = filename;
    handles.baseName = baseName;
    set(handles.listbox2,'String',filename);
    rgb = imread(strcat(baseName,filename));
    bulkI{1} = rgb2gray(rgb);
    axes(handles.axes1);
    set(handles.listbox2,'Value',1);
    imshow(bulkI{1},[]);
    handles.bulkI = bulkI;
    handles.I = bulkI{1};
    guidata(hObject,handles);
else
    handles.filename = filename;
    handles.baseName = baseName;
    set(handles.listbox2,'String',filename);
    for i = 1:length(filename)
        rgb = imread(strcat(baseName,filename{i}));
        bulkI{i} = rgb2gray(rgb);
    end
    axes(handles.axes1);
    set(handles.listbox2,'Value',1);
    imshow(bulkI{1},[]);
    handles.bulkI = bulkI;
    handles.I = bulkI{1};
    guidata(hObject,handles);
end
```

```
% --------------------------------------------------------------------
function prevOpen_Callback(hObject, eventdata, handles)
% hObject    handle to prevOpen (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

set(handles.slider2,'Value',250);
handles.ws = [250,250];
handles.small = 250;
guidata(hObject,handles);

checkImage = isfield(handles,'bulkI');
checkSegment = isfield(handles,'coloredLabels');
checkDeleted = isfield(handles,'ccDeleted');
if checkDeleted == 1
    handles                                                                  =
rmfield(handles,{'signature','ccDeleted','centersDeleted','signatureDeleted',
'diametersDeleted','CC','ws','small','bulkI','coloredLabels','imgMasked','thr
esh','centers','diameter'});
    guidata(hObject,handles);
elseif checkSegment == 1
    handles                                                                  =
rmfield(handles,{'signature','CC','ws','small','bulkI','coloredLabels','imgMa
sked','thresh'});
    guidata(hObject,handles);
elseif  checkImage == 1
    handles = rmfield(handles,'bulkI');
    guidata(hObject,handles);

end
%Clear all axes and listbox
cla(handles.axes1);
cla(handles.axes2);
cla(handles.axes3);
set(handles.listbox2,'String','');
set(handles.radiobutton1,'Value',0);
set(handles.radiobutton2,'Value',0);
set(handles.radiobutton3,'Value',0);


[filenameNew, baseName] = uigetfile('*.mat' ,'Select an image set');
if isequal(filenameNew,0)
else
    load(strcat(baseName,filenameNew));
    handles.baseName = baseName;
    handles.filename = filename;
    handles.signature = signatureAllMetrics;
    handles.bulkI = bulkI;
    handles.coloredLabels = coloredLabels;
    handles.imgMasked = imgMasked;
    handles.CC = CC;
    handles.ccDeleted = ccDeleted;
    handles.centersDeleted = centersDeleted;
    handles.signatureDeleted = signatureDeleted;
    handles.diametersDeleted = diametersDeleted;
    handles.centers = centers;
    handles.diameter = diameters;
    handles.thresh = thresh;
    % handles.ws = ws;
    % handles.small = small;
    % set(handles.slider2,'Value',small);

    guidata(hObject,handles);

    set(handles.listbox2,'String',handles.filename);
    % rgb = imread(filename);
    % bulkI{1} = rgb2gray(rgb);
    for i = 1:length(filename)
        rgb = imread(strcat(baseName,filename{i}));
        bulkI{i} = rgb2gray(rgb);
    end
    axes(handles.axes1);
    set(handles.listbox2,'Value',1);
    imshow(bulkI{1},[]);
    handles.bulkI = bulkI;
    handles.I = bulkI{1};

    axes(handles.axes2);
    imshow(coloredLabels{1},[])

    axes(handles.axes3);
    imshow(imgMasked{1},[])
    guidata(hObject,handles);
end


% --- Executes on button press in togglebutton1.
function togglebutton1_Callback(hObject, eventdata, handles)
% hObject    handle to togglebutton1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Hint: get(hObject,'Value') returns toggle state of togglebutton1
selected = get(handles.listbox2,'Value');
I = handles.bulkI{selected};
CC = handles.CC{selected};
imgMasked = handles.imgMasked{selected};
CClabel = labelmatrix(CC);
% % CClabel(CClabel == 1) = 2;
% % CClabel(CClabel == 3) = 1;
% % CClabel(CClabel == 2) = 3;
% % CClabel(CClabel == 4) = 3;
% % CClabel(CClabel > 4) = 1;
% CClabel(CClabel < 13 & CClabel > 0) = 1;
% CClabel(CClabel == 13) = 3;
coloredLabels = label2rgb(CClabel, 'hsv', 'k','shuffle');
% %
% %
% figure, imshow(I,[]);
% figure, imshow(coloredLabels,[]);
% % imwrite(coloredLabels,'coloredSUM159_presentation.jpeg');
% imwrite(I,'SUM159_presentation_2.jpeg');

imwrite(coloredLabels,'labels_SUM159_Iobrcbr.tiff');
```

```
% % save('image_5E','I','-jpeg');
% % save('labels_5E','coloredLabels','jpeg');


% --------------------------------------------------------------------
function dropdown2_Callback(hObject, eventdata, handles)
% hObject    handle to dropdown2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)


% --------------------------------------------------------------------
function metrics_Callback(hObject, eventdata, handles)
% hObject    handle to metrics (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
global globalStruct
standaloneCheckBox


% --- Executes on button press in pushbutton8.
function pushbutton8_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton8 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
global globalStruct

%check if metrics have been selected or if default is necessary
if isfield(globalStruct,'default') == 0
    globalStruct.area = 1;
    globalStruct.majorAxisLength = 0;
    globalStruct.minorAxisLength = 0;
    globalStruct.eccentricity = 1;
    globalStruct.orientation = 0;
    globalStruct.convexArea = 0;
    globalStruct.diameter = 0;
    globalStruct.solidity = 0;
    globalStruct.extent = 0;
    globalStruct.perimeter = 1;
    globalStruct.Z = 0;
    globalStruct.A = 0;
    globalStruct.Phi = 0;
    globalStruct.meanTotalPixels = 0;
    globalStruct.stdTotalPixels = 0;
    globalStruct.p2a = 0;
    globalStruct.default = [1 4 10];
else
end


%Check for image load
checkImage = isfield(handles,'bulkI');
%Check for radio button
radioCheck1 = get(handles.radiobutton1,'Value');
radioCheck2 = get(handles.radiobutton2,'Value');
radioCheck3 = get(handles.radiobutton3,'Value');
%Check for segmentation
segmentCheck = isfield(handles,'coloredLabels');
```

```
if checkImage == 0
    mb1 = msgbox('Please select an image file');
    t = timer;
    t.StartDelay = 3;
    t.TimerFcn = @(~,~) delete(mb1);
    start(t)
    % elseif radioCheck1 == 0 && radioCheck2 == 0 && radioCheck3 == 0
    %     mb2 = msgbox('Please select a cell size');
    %     t = timer;
    %     t.StartDelay = 3;
    %     t.TimerFcn = @(~,~) delete(mb2);
    %     start(t)
elseif segmentCheck == 0
    mb3 = msgbox('Please choose automatic threshold');
    t = timer;
    t.StartDelay = 3;
    t.TimerFcn = @(~,~) delete(mb3);
    start(t)
else
    if isfield(handles,'ccDeleted') == 0
        handles.ccDeleted = [];
        handles.centersDeleted = [];
        handles.signatureDeleted = [];
        handles.diametersDeleted = [];
    end

    [answer,PathName]  =  uiputfile({'*.mat','MAT-files  (*.mat)';'*xls','Excel
Files (*.xls)'},'Export Metrics As');

    if answer == 0

    elseif isempty(strfind(answer,'.mat')) == 1
        keepMetrics    =    [globalStruct.area,    globalStruct.majorAxisLength
globalStruct.minorAxisLength ...
            globalStruct.eccentricity                globalStruct.orientation
globalStruct.convexArea globalStruct.diameter ...
            globalStruct.solidity  globalStruct.extent  globalStruct.perimeter
...
            globalStruct.Z           globalStruct.A           globalStruct.Phi
globalStruct.meanTotalPixels ...
            globalStruct.stdTotalPixels globalStruct.p2a];
        keepMetrics = find(keepMetrics);
        signature = handles.signature;
        metrics = {'Area', 'MajorAxisLength', 'MinorAxisLength', ...
            'Eccentricity', 'Orientation', 'ConvexArea', 'EquivDiameter', ...
            'Solidity', 'Extent', 'Perimeter', 'Z', 'A', 'Phi', ...
            'meanTotalPixels', 'stdTotalPixels','p2a'};
        metrics = metrics(keepMetrics);
        h = waitbar(0,'Please wait...Exporting to Excel');

        for j = 1:length(signature)
            signature{j} = signature{j}(keepMetrics,:);
            sheet = j;
            xlswrite(char(strcat(PathName,answer,'.xls')),{'Spheroid
#'},sheet,'A1');

xlswrite(char(strcat(PathName,answer,'.xls')),[1:100]',sheet,'A2');
            for i = 1:length(metrics)
```

```
                xlRange1 = strcat(char('A'+i),'1');
                xlRange2 = strcat(char('A'+i),'2');

xlswrite(char(strcat(PathName,answer,'.xls')),metrics(i),sheet,xlRange1);

xlswrite(char(strcat(PathName,answer,'.xls')),signature{j}{i}',sheet,xlRange2
);
                waitbar(((i/length(metrics))*(1/length(signature)))+((j-
1)/length(signature)));
            end
            waitbar(j/length(signature))
        end
        close(h);
        %Display message box
        mb = msgbox('Signature Export Successful!');
        t = timer;
        t.StartDelay = 2;
        t.TimerFcn = @(~,~) delete(mb);
        start(t)



    else

        keepMetrics    =    [globalStruct.area,    globalStruct.majorAxisLength
globalStruct.minorAxisLength ...
            globalStruct.eccentricity                 globalStruct.orientation
globalStruct.convexArea globalStruct.diameter ...
            globalStruct.solidity  globalStruct.extent  globalStruct.perimeter
...
            globalStruct.Z           globalStruct.A           globalStruct.Phi
globalStruct.meanTotalPixels ...
            globalStruct.stdTotalPixels globalStruct.p2a];
        keepMetrics = find(keepMetrics);
        filename = handles.filename;
        signatureAllMetrics = handles.signature;
        for i = 1:length(signatureAllMetrics)
            signatureTrimmed{i} = signatureAllMetrics{i}(keepMetrics,:);
        end
        bulkI = handles.bulkI;
        coloredLabels = handles.coloredLabels;
        imgMasked = handles.imgMasked;
        CC = handles.CC;
        ccDeleted = handles.ccDeleted;
        centersDeleted = handles.centersDeleted;
        signatureDeleted = handles.signatureDeleted;
        diametersDeleted = handles.diametersDeleted;
        centers = handles.centers;
        diameters = handles.diameter;
        thresh = handles.thresh;
        ws = handles.ws;
        small = handles.small;
        baseName = handles.baseName;

        save(char(strcat(PathName,answer)),'filename','signatureTrimmed',
'signatureAllMetrics','bulkI',...
            'coloredLabels','imgMasked','thresh',
'CC','ccDeleted','centersDeleted',...
```

```
            'signatureDeleted','diametersDeleted','centers','diameters','ws',
'small','baseName');
        %Uncheck radio buttons
        set(handles.radiobutton1,'Value',0);
        set(handles.radiobutton2,'Value',0);
        set(handles.radiobutton3,'Value',0);
        %Clear all axes and data
        cla(handles.axes1);
        cla(handles.axes2);
        cla(handles.axes3);
        handles                                                            =
rmfield(handles,{'ws','small','bulkI','coloredLabels','imgMasked','thresh',
'CC','ccDeleted','centersDeleted','signatureDeleted','diametersDeleted'});
        guidata(hObject,handles);
        set(handles.listbox2,'String','');
        clearvars -global

        %Display message box
        mb = msgbox('Signature Export Successful!');
        t = timer;
        t.StartDelay = 2;
        t.TimerFcn = @(~,~) delete(mb);
        start(t)

    end
end
% --------------------------------------------------------------------
function calibration_Callback(hObject, eventdata, handles)
% hObject    handle to calibration (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
prompt = {'Pixels:','Microns:'};
dlg_title = 'Calibration';
num_lines = 1;
% defaultans = {'20','hsv'};
answer = inputdlg(prompt,dlg_title,num_lines);
m2p = str2num(answer{2})./str2num(answer{1});
signature = handles.signature;
for i = 1:length(signature)
    signatureMicrons{i} = cellfun(@(x) x.*m2p,signature{i},'un',0);
end
handles.signature = signatureMicrons;
guidata(hObject,handles);
```

7.1.4 Image spheroid metrics

Spheroid area, perimeter, major axis length, minor axis length, equivalent diameter, convex area, solidity, extent, orientation, and eccentricity were calculated using the bwconncomp function in MATLAB. The perimeter to area ratio was calculated from the extracted parameters. The Zernike moments were calculated using publicly available functions.

7.1.5 Image signature analysis

Cell lines spheroid heterogeneity was analyzed by collecting the profiles across dozens to hundreds of spheroids. The median value for each metric was collected to make an aggregate signature. These aggregate signatures were clustered by Euclidean distance and ward linkage. Spheroid heterogeneity was quantified by gating the principle component values of the spheroid signatures to be less than 0.1 along principle component #1, and between 0 and -0.2 along principle component #2.

7.2 Chapter 3 Methods

7.2.1 Cell lines and 3D culture

The MCF10A-5E clone was previously reported and was grown in organotypic 3D culture as described for MCF10A cells (98, 273). MCF10DCIS.com cells (121) were originally obtained from Wayne State University and cultured in DMEM/F-12 medium (Invitrogen) plus 5% horse serum (Invitrogen). MDA-MB-468 cells were obtained from ATCC and cultured in L-15 medium (Invitrogen) plus 10% fetal bovine serum (Hyclone) without supplemental CO_2.

7.2.2 Plasmids

pLKO.1 shGFP puro (Addgene #12273), pLKO.1 shTGFBR3 puro (TRCN0000033430), and pLKO.1 shJUND puro (TRCN0000014974) were obtained from The RNAi Consortium (274) or Addgene. pBabe JunD-HA neo, pBabe JUND-HA neo, pBabe RFP1-Smad2 neo, pBabe JunD-Venus puro, and pBabe RFP1-KRT5 hygro were constructed by PCR cloning from plasmid templates (Open Biosystems) into the retroviral vector pBabe neo, pBabe puro, or pBabe hygro. Doxycycline-inducible Tgfbr3-HA, JunD-HA and expression vectors were constructed by PCR cloning or subcloning into the entry vector pEN_TTmiRc2, followed by LR recombination into the lentiviral vector pSLIK neo (275). The human TRIPZ lentiviral inducible shTNC construct V2THS_133229 (Ref. (276)) were obtained from Open Biosystems. pLenti PGK Blast V5-LUC (w528-1) was obtained from Addgene.

pTRF.1 udsVenus (P_{JUND}) was constructed starting with the commercial lentiviral vector, pTRF1-mCMV-dscGFP (System Biosciences). First, cGFP was excised from pTRF1-mCMV-dscGFP by restriction digest with *HindIII* and *EcoRV*, and the vector was ligated with a similarly digested Venus (149) prepared by PCR cloning designed to contain the appropriate motif for N-end rule degradation (149, 151). The resulting pTRF.1-mCMV-dsVenus was then digested with *HindIII*, dephosphorylated, and ligated with a similarly digested ubiquitin C monomer prepared by PCR cloning from oligo(dT)-primed MCF10A-5E cDNA to produce pTRF.1-mCMV-udsVenus. Last, this vector was digested with *EcoRI* and *SpeI* and ligated with a similarly digested P_{JUND} prepared by PCR from MCF10A-5E genomic DNA to produce pTRF.1 udsVenus (P_{JUND}). All constructs were verified by sequencing.

7.2.3 Viral transduction

Lentiviruses were prepared in 293T cells (ATCC) by triple transfection of the lentiviral vector together with psPAX2 + pMD.2G (Addgene) and transduced into MCF10A-5E, MDA-MB-468, and MCF10DCIS.com cells as described previously (277). Retroviruses were similarly prepared by double transfection of the pBabe construct together with pCL ampho (Addgene) and transduced into MCF10A-5E cells as described previously (277). For viral vectors carrying selectable markers, transduced cells were selected in growth medium containing 2 μg/ml puromycin, 300 μg/ml G418, 100 μg/ml hygromycin, or 4–6 μg/ml blasticidin until control plates had cleared. For addback experiments, viral titers were adjusted to match the endogenous protein expression as closely as possible. For live-cell reporters, we used the minimum viral titer that gave sufficient signal above background for long-term imaging. For pTRF.1 udsVenus (P_{JUND}), which lacks a selectable marker, transduced cells were flow sorted for baseline Venus fluorescence at the University of Virginia Flow Cytometry Core Facility.

7.2.4 Quantitative PCR

Quantitative PCR was performed as described elsewhere (278). Primer sequences are as follows: *TGFBR3*, 5’- tgtcacctggcacattcatt -3’ (forward), 5’- acaggatttgccatgcattt -3’ (reverse); *TGFBI*, 5’- ctatgccaagtccctggaaa -3’ (forward), 5’- cctccaagccacgtgtagat -3’ (reverse); *JUND*, 5’- cgttggttgtgtgtgtgtgt -3’ (forward), 5’- ggcgaaccaaggattacaaa -3’ (reverse); *KRT5*, 5’- tttgtctccaccacctcctc -3’ (forward), 5’-cctgggaaccaaagaatgtg -3’ (reverse); *RPS6*, 5’- ccccaaaagagctagcagaa -3’ (forward), 5’- ctgcaggacacgtggagtaa -3’ (reverse); *TNC*, .5’- aaccccaggagtttgagacc –3’ (forward), 5’- gggctccagtgattttccta -3’ (reverse).

7.2.5 Immunoblotting

MCF10A-5E cells expressing the indicated constructs were lysed in 50 mM Tris (pH 8.0), 150 mM NaCl, 5 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.5% sodium deoxycholate. 20–30 μg of clarified extract was separated on an 8 or 10% SDS-PAGE gel and transferred to PVDF (Immobilon-FL, Millipore). Membranes were blocked with 0.5× Odyssey blocking buffer (LI-COR, 1:1 in PBS) and incubated overnight at 4ºC in 0.5× Odyssey blocking buffer

(LI-COR) + 0.1% Tween, containing one of the following primary antibodies: TGF-β Receptor III (1:1000, Cell Signaling, #2519), HSP 90α/β (H-114) (1:1000, Santa Cruz, sc-7947), Jun D (329) (1:1000, Santa Cruz, sc-74), S6 Ribosomal Protein (54D2) (1:1000, Cell Signaling, #2317), Phospho-S6 Ribosomal Protein (Ser240/244) (D68F8) (1:1000, Cell Signaling, #5364), α/β-tubulin (1:1000, Cell Signaling, #2148), α-tubulin (1:20,000, Abcam, ab89984), β-actin (1:1000, Ambion, #AM4302), Keratin 5 (1:1000, Covance, SIG-3475), Keratin 14 (1:1000, Covance, PRB-155P), Smad2 (L16D3) (1:1000, Cell Signaling, #3103), GFP (1:1000, Invitrogen, A-11122), Tenascin (BC-24) (1:1000, Sigma, T2551), vimentin (SP20) (1:100, Abcam, ab16700), E-cadherin (36) (1:1000, BD Biosciences, #610182), caspase-3 (1:1000, Cell Signaling, #9662), and KRT15 (1:1000, Thermo, MA5-15567). Membranes were washed 4 × 5 minutes in PBS-T (PBS + 0.1% Tween) and incubated for 1 hr at room temperature in secondary antibody solution (0.5× Odyssey blocking buffer + 0.01% SDS + 0.1% Tween) containing IrDye 800 or IrDye 680LT-conjugated secondary antibody (1:20,000, LI-COR). Membranes were washed 4 × 5 minutes in PBS-T, rinsed with PBS, and imaged by infrared fluorescence on a LI-COR Odyssey instrument. Relative band intensities were quantified by densitometry with ImageJ.

7.2.6 Clinical samples

The pathology database at the University of Virginia from 2004-2012 was searched for all cases of high-grade DCIS, since this is the cohort that contains the basaloid subgroup. The set was then searched for estrogen receptor status and only those that were ER negative were selected. The search was confined to 2004 and later because 2004 was the year pathologists began reflexively testing DCIS for estrogen receptor status. All cases with an invasive carcinoma component were excluded. This resulted in 5–7 cases per year. The cases were deidentified for any patient demographics and used for immunohistochemical analysis of cytokeratin 5/6. Samples that were positive for cytokeratin 5/6 were followed up with a panel of six immunohistochemical markers: p53 (11/22 positive), E-cadherin (22/22 positive), KRT18 (21/22 positive), p63 (1/22 positive), smooth-muscle actin (21/22 positive), and vimentin (6/22 positive) (Supplementary Table 1). All clinical work was done according to a protocol under IRB-HSR approval #14176 and PRC approval #1363 (502-09).

7.2.7 Immunofluorescence.

Immunofluorescence in frozen sections or on coverslips was performed as described previously (277) using the following primary antibodies: Cytokeratin 5/6 (D5/16 B4) (1:200, Dako, M7237), Keratin 5 (1:5000, Covance, SIG-3475), Tenascin (BC-24) (1:2000, Sigma, T2551), or Jun D (329) (1:500, Santa Cruz, sc-74). For paraffin sections, slides underwent antigen retrieval before immunostaining as described for TGFBR3 below. Whole-mount immunofluorescence of day 10 acini was performed as described previously (277) using the following primary antibodies: E-cadherin (36) (1:500, BD Biosciences, #610182) or HA (3F10) (1:200, Roche, #11815016001).

7.2.8 Two-color time-lapse confocal imaging.

Live-cell experiments involved MCF10A-5E cells stably transduced with pTRF.1 udsVenus (P_{JUND}) and pBabe RFP1-Smad2 neo as described above. For long-term live-cell imaging of 3D acini, a plastic coverslip was cut to size and placed at the base of an 8-well chamber slide (BD Biosciences) before starting. Coverslipped chamber slides were then coated with Matrigel (BD Biosciences), and cells were grown in organotypic 3D culture as described for MCF10A cells (98). At day 10, the Matrigel-coated plastic coverslip containing adherent 3D acini was removed and flipped upside down into a culture dish with a fused glass coverslip (MatTek) filled with conditioned medium from the 3D culture. A second glass coverslip was placed on top of the inverted plastic coverslip, and the air-tight reservoir was sealed by applying high vacuum grease (Dow Corning) followed by a mixture of Vaseline, lanolin, and paraffin. The sealed reservoir was then covered with light mineral oil to prevent evaporation during imaging.

Live acinar cultures were maintained at 37°C with a heated blower. Time-lapse imaging was performed using a laser scanning confocal microscope (LSM 700, Carl Zeiss) equipped with an EC Plan-Neofluar 40x/1.30 oil-immersion objective and four diode lasers (5–10 mW) centered at 405, 488, 555, and 639 nm. udsVenus was excited at 488 nm and its emission detected between 488 and 585nm. RFP1 was excited at 555 nm and its emission detected above 582 nm. The confocal pinhole was kept at 1 Airy unit, and laser powers were typically set at 2–5% to minimize photobleaching. Time-lapse images were acquired every 15 minutes for 15–20 hr. Image assembly and processing were performed using MetaMorph (Molecular Devices).

7.2.9 Image segmentation and quantification.

Single cells and nuclei from live-cell images were manually segmented and applied to the RFP1 and udsVenus fluorescence stacks to calculate median nuclear, cytoplasmic, and total fluorescence intensities. The RFP1-Smad2 signal was evaluated as the ratio of median nuclear-to-cytoplasmic fluorescence. The median total udsVenus fluorescence per cell was normalized to the overall fluorescence intensity of cells in the same frame to account for photobleaching.

For RNA FISH scoring, single ECM-attached cells were manually binned into high expression and low-no expression based on DNP-labeled riboprobe staining intensity. A minimum of 150 cells was scored per hybridization across four independent hybridizations.

7.2.10 Time course alignment.

Data from live cell imaging time courses were spectrally decomposed with the fft function in MATLAB, smoothed by low-pass filtering at 8.6e^{-5} Hz, and then reconstructed with the ifft function in MATLAB. After spectral filtering, individual time courses were standardized and clustered hierarchically based on the joint alignment of the RFP1-Smad2 and udsVenus traces between experiments. The alignment algorithm is based on a full sliding window of both traces with zero gaps and a cost function that uses the sum-of-squared difference between the two experiments to be aligned, scaled by the extent of overlap between them. All possible experiment pairs and alignments within the dataset were considered, and the experiments with the best pairwise alignment were combined by average linkage. The exhaustive pairwise comparisons and linkages were repeated until all of the independent experiments were aligned. In the final pairwise comparison, one of two nearly equivalent alignments was visually selected.

7.2.11 Computational modeling

The TGFBR3–JUND circuit was modeled as a system of coupled ordinary differential equations. mRNA and protein species were assigned basal synthesis and degradation rates as described in Chapter 3. Transcriptional inhibition steps were modeled using the Hill equation without cooperativity, and feedback strengths were adjusted manually as free parameters to reproduce the damped periodicity observed experimentally. Sensitivity analysis on the manually adjusted parameters is described in Chapter 3. The model was simulated with ode15s in MATLAB and allowed to reach steady state before exciting the system with a 50% increase in the appropriate reaction rate for 1 hr.

The agent-based model of CC was constructed using NetLogo v4.1.1 (http://ccl.northwestern.edu/netlogo/). Single cells were seeded at a predefined geometry and initialized with the same basal level of KRT5 and JUND. During each time step of the simulation, KRT5 and JUND were incremented by a uniform pseudorandom number between zero and one [U(0,1)] that was inversely scaled by the number of neighboring TNC-positive cells (TNC$^+$) as follows: $\frac{U(0,1)}{2^{TNC^+}}$. Since JUND inhibits late keratinization, the fluctuating JUND–KRT5 difference was used as a proxy for keratinization. The JUND–KRT5 difference was evaluated after each time step, and keratinization occurred when the difference reached a critical negative threshold. Keratinized cells then expressed TNC and were no longer incremented for KRT5 or JUND expression. The model was run until steady state, and the final display was used as the model output.

7.2.12 Ordinary differential equation model code

```
function dJT_dt = JR3ode(t,JT, params)
% nH;                  %Hill coefficient for transcription steps
% IC50_TGFbetaRIII;    %IC50 for TGFbetaRIII inhibition of transcription
% IC50_junD;           %IC50 for junD inhibition of transcription
% k_basaltxn;          %baseline transcription rate
% k_translation;       %constant translation rate
% k_deg_RNA;           %RNA turnover rate
% k_activation;        %Activation of ligand bound R3
% k_deg_junD;          %Degradation of JUND protein
% k_deg_R3;            %Degradataion of TGFBR3 protein
% k_deg_udsVenus;      %Degradation of Venus protein

%unpackage parameters
paramsCell=mat2cell(params,ones(size(params,1),1),ones(size(params,2),1));
[nH, IC50_TGFbetaRIII, IC50_junD, k_basaltxn_R, k_basaltxn_J, k_translation,
...
    k_deg_RNA, k_deg_junD, k_deg_R3, k_activation, k_deg_udsVenus, f1, f2,
f3] = paramsCell{:};


%Species
JUND=JT(1);
junD=JT(2);
TGFBR3=JT(3);
```

```
TGFbetaRIII=JT(4);
R3star = JT(5);
Venus = JT(6);


%Remove negative values
if JUND<0
    JUND=0;
end
if junD<0
    junD=0;
end
if TGFBR3<0
    TGFBR3=0;
end
if TGFbetaRIII<0
    TGFbetaRIII=0;
end
if R3star < 0
    R3star = 0;
end
if Venus < 0;
    Venus = 0;
end

%Differential equations determining change of species at each time step - see
Supplemental Note for details
dTGFBR3_dt=k_basaltxn_R-f1*R3star^nH/(R3star^nH+IC50_TGFbetaRIII^nH)- ...
    f3*junD^nH/(junD^nH+IC50_junD^nH)-k_deg_RNA*TGFBR3;

dJUND_dt=k_basaltxn_J-f2*R3star^nH/(R3star^nH+IC50_TGFbetaRIII^nH)- ...
   f3*junD^nH/(junD^nH+IC50_junD^nH)-k_deg_RNA*JUND;

dTGFbetaRIII_dt=k_translation*TGFBR3-k_deg_R3*TGFbetaRIII-
k_activation*TGFbetaRIII;

djunD_dt=k_translation*JUND-k_deg_junD*junD;

dR3star_dt = k_activation*TGFbetaRIII-k_deg_R3*R3star;

dVenus_dt = k_translation*JUND-k_deg_udsVenus*Venus;

%Do not allow species to go below 0 units
if JUND==0 && dJUND_dt<0
    dJUND_dt=0;

end
if junD==0 && djunD_dt<0
    djunD_dt=0;

end
if TGFBR3==0 && dTGFBR3_dt<0
    dTGFBR3_dt=0;
end
if TGFbetaRIII==0 && dTGFbetaRIII_dt<0
    dTGFbetaRIII_dt=0;
end
if R3star==0 && dR3star_dt<0
```

```
    dR3star_dt=0;
end
if Venus==0 && dVenus_dt<0
    dVenus_dt=0;
end

%Repackage outputs
dJT_dt=[dJUND_dt;   djunD_dt;   dTGFBR3_dt;   dTGFbetaRIII_dt;   dR3star_dt;
dVenus_dt];

return

% Run JUND-R3 module

clear all
close all
clc

%Initialize parameters
%Hill parameters
nH = 1;
IC50_TGFBRIII = 100;
IC50_junD = 100;

%Synthesis and degradation
k_basaltxn_J = 4;        %hr^-1 based on Schwanhausser et al. Nature 473:337-
42 (2011)
k_basaltxn_R = 4;        %hr^-1 based on Schwanhausser et al. Nature 473:337-
42 (2011)
k_translation = 100;     %mRNA^-1 hr^-1 based on Schwanhausser et al. Nature
473:337-42 (2011)

%Degradation rates (see Supplemental Note 1)
k_deg_RNA = 0.23;        %hr^-1 based on Zou et al. Mol Cell Biol 30:5021-32
(2010) and
                        %Hempel et al. Carcinogenesis 29:905-12 (2008)
k_deg_junD = 0.37;      %hr^-1 based on Fig. SN1 in Supplementary Note 1
k_deg_R3 = 3.0;         %hr^-1 based on Fig. SN1 in Supplementary Note 1

%R3 -> R3*
k_activation = 1;     %hr^-1 based on steady-state distribution of RFP1-Smad2
reporter
                        %in Supplementary Fig. S3b (see Eqn. 8 in
Supplementary Note 1)

%Venus degradation
k_deg_udsVenus = 2.8;   %hr^-1 based on Supplementary Fig. S3d

%Feedback strengths
f1 = 7;
f2 = 9;
f3 = 5;

%Assign parameters
params = [nH, IC50_TGFBRIII, IC50_junD, k_basaltxn_R, k_basaltxn_J,
k_translation, ...
    k_deg_RNA, k_deg_junD, k_deg_R3, k_activation, k_deg_udsVenus, f1, f2,
f3];
```

```
%Initial conditions for JUND mRNA, junD protein, TGFBR3 mRNA, ...
%TGFbetaRIII protein, active R3* (SMAD reporter), Venus
init_cond = 1000*ones(1,6);

tspan1 = [0 501];
tspan2 = [501 502];
tspan3 = [502 800];

%% Excite the system by increasing k_activation by 1.5x

%Run model to steady state
options = [];
[time,JT_soln]=ode15s(@JR3ode, tspan1, init_cond, options, params);

% Increase k_activation 1.5 fold for one hour
params(10) = 1.5*params(10);
init_cond_2 = JT_soln(end,:);
[time2,JT_soln2]=ode15s(@JR3ode,tspan2,init_cond_2, options, params);

%Allow model to relax after activation
params(10) = k_activation;
init_cond_3 = JT_soln2(end,:);
[time3,JT_soln3]=ode15s(@JR3ode,tspan3,init_cond_3, options, params);

%Collect total traces for each species
time=[time' time2' time3']';

JUND_soln=[JT_soln(:,1)' JT_soln2(:,1)' JT_soln3(:,1)']';
junD_soln=[JT_soln(:,2)' JT_soln2(:,2)' JT_soln3(:,2)']';
TGFBR3_soln=[JT_soln(:,3)' JT_soln2(:,3)' JT_soln3(:,3)']';
TGFbetaRIII_soln=[JT_soln(:,4)' JT_soln2(:,4)' JT_soln3(:,4)']';
R3_star=[JT_soln(:,5)' JT_soln2(:,5)' JT_soln3(:,5)']';
Venus =[JT_soln(:,6)' JT_soln2(:,6)' JT_soln3(:,6)']';

R3_star_o =[JT_soln3(:,5)']';
Venus_o =[JT_soln3(:,6)']';

%Fig. 3H-i plot
figure(1)

a = subplot(2,3,1);

plot(time, R3_star, 'r')
xlim([490 550])
xlabel('Time')
ylabel('TGFBR3* reporter strength');
title(a, 'Increase k_a_c_t of TGFBR3 as system stimulus');

figure(1)
subplot(2,3,4)
plot(time, Venus, 'g')
xlim([490 550])
xlabel('Time')
ylabel('JUND reporter strength');

%% Excite the system by increasing k_txn_R3 by 1.5x
```

```
% Run model to steady state
options = [];
[time,JT_soln]=ode15s(@JR3ode, tspan1, init_cond, options, params);

%Increase k_txn_JUND 1.5 fold for one hour
params(4) = 1.5*params(4);
init_cond_2 = JT_soln(end,:);
[time2,JT_soln2]=ode15s(@JR3ode,tspan2,init_cond_2, options, params);

%Allow model to relax after activation
params(4) = k_basaltxn_R;
init_cond_3 = JT_soln2(end,:);
[time3,JT_soln3]=ode15s(@JR3ode,tspan3,init_cond_3, options, params);

%Collect total traces for each species
time=[time' time2' time3']';

JUND_soln=[JT_soln(:,1)' JT_soln2(:,1)' JT_soln3(:,1)']';
junD_soln=[JT_soln(:,2)' JT_soln2(:,2)' JT_soln3(:,2)']';
TGFBR3_soln=[JT_soln(:,3)' JT_soln2(:,3)' JT_soln3(:,3)']';
TGFbetaRIII_soln=[JT_soln(:,4)' JT_soln2(:,4)' JT_soln3(:,4)']';
R3_star=[JT_soln(:,5)' JT_soln2(:,5)' JT_soln3(:,5)']';
Venus =[JT_soln(:,6)' JT_soln2(:,6)' JT_soln3(:,6)']';

% %Fig. 3H-ii plot
figure(1)

a = subplot(2,3,2);
plot(time, R3_star, 'r')
xlim([490 550])
xlabel('Time')
ylabel('TGFBR3* reporter strength');
title(a, 'Increase k_t_x_n of TGFRB3 as system stimulus');

figure(1)
subplot(2,3,5)
plot(time, Venus, 'g')
xlim([490 550])
xlabel('Time')
ylabel('JUND reporter strength');


%% Excite the system by increasing k_txn_JUND by 1.5x

% Run model to steady state
options = [];
[time,JT_soln]=ode15s(@JR3ode, tspan1, init_cond, options, params);

%Increase k_txn_JUND 1.5 fold for one hour
params(5) = 1.5*params(5);
init_cond_2 = JT_soln(end,:);
[time2,JT_soln2]=ode15s(@JR3ode,tspan2,init_cond_2, options, params);

%Allow model to relax after activation
params(5) = k_basaltxn_J;
init_cond_3 = JT_soln2(end,:);
[time3,JT_soln3]=ode15s(@JR3ode,tspan3,init_cond_3, options, params);
```

```
%Collect total traces for each species
time=[time' time2' time3']';

JUND_soln=[JT_soln(:,1)' JT_soln2(:,1)' JT_soln3(:,1)']';
junD_soln=[JT_soln(:,2)' JT_soln2(:,2)' JT_soln3(:,2)']';
TGFBR3_soln=[JT_soln(:,3)' JT_soln2(:,3)' JT_soln3(:,3)']';
TGFbetaRIII_soln=[JT_soln(:,4)' JT_soln2(:,4)' JT_soln3(:,4)']';
R3_star=[JT_soln(:,5)' JT_soln2(:,5)' JT_soln3(:,5)']';
Venus =[JT_soln(:,6)' JT_soln2(:,6)' JT_soln3(:,6)']';

R3_star_o =[JT_soln3(:,5)']';
Venus_o =[JT_soln3(:,6)']';

%Fig. 3H-iii plot
figure(1)

a = subplot(2,3,3);
plot(time, R3_star, 'r')
xlim([490 550])
%axis([450 801 10 40])

xlabel('Time')
ylabel('TGFBR3* reporter strength');
title(a, 'Increase k_t_x_n of JUND as system stimulus');

figure(1)
subplot(2,3,6)
plot(time, Venus, 'g')
xlim([490 550])
%axis([450 801 80 120])

xlabel('Time')
ylabel('JUND reporter strength');
```

7.2.13 Agent-based model code

```
; Global variable
breed [cells cell]
breed [ghosts ghost]

turtles-own [ KRT5 JUND locked TNC_prob]

to setup
; This flag is for if you want to draw a custom geometry or if you want to
load a predetermined geometry
  ifelse drawornot [import-world "cc_4.csv"]
  [

  ]
; Reset the random seed generator
  random-seed timer
; Set the default shapes for the cells
  set-default-shape cells "circle"
  set-default-shape ghosts "circle"
```

```
; Spawn a cell on any gray patch, and initialize the KRT5 and JUND basal
levels
ask patches [
   if pcolor = gray [
     sprout-cells 1 [
      set KRT5 1
      set JUND 1

      set color 11
      set locked 0
      set TNC_prob 1
    ]
  ]
  ]
end

to go
; Ask patches to disappear
  ask patches [set pcolor black]
; Progress the cells
  if ticks < 5000 [
    ask cells [progress]
    ask cells [isghost]
; Scale the color of the cells to the JUND level
    ask cells [
      ifelse ((11 + (JUND / 100)) < 15)
      [set color (11 + JUND / 100)]
      [set color 15]
]
; Progress
    tick
  ]

end

to progress
; If not a KRT+ cell, ramp up both KRT5 and JUND by a random amount scaled by
the local TNC
  if locked < 1 [
    set KRT5 KRT5 + ((random 100) / 100) / TNC_prob
    set JUND JUND + ((random 100) / 100) / TNC_prob
]
end

to isghost
; If the KRT5 level exceeds the threshold, set the cell to a keratinized cell
and secrete local TNC
if KRT5 - JUND > threshold [
    ask cells-here[
      set breed ghosts
      set color 45
      if TNC [
        ask cells-on neighbors [
       set TNC_prob 2 * TNC_prob
       ]
      ]
    ]
]
end

to clear
  clear-all
end

; Draw custom geometries
to draw-tumor

   if mouse-down?[
    ask patch (mouse-xcor) (mouse-ycor) [ set pcolor gray ]
   ]
end
```

7.2.14 Suspension assays

MCF10A-5E cells expressing the indicated constructs were trypsinized and plated at 400,000 cells/ml in assay medium containing 5 ng/ml EGF on poly-(2-hydroxyethyl methacrylate) (poly-HEMA) coated tissue culture plates. At the indicated time points, medium was removed by centrifugation at 150 rcf for 3 minutes. Cells were washed with 500 μl ice-cold PBS. For immunoblotting, cells were lysed in 62.5 mM Tris (pH 6.8), 2% SDS, 10% glycerol, 0.01% bromphenol blue, 2.5% EtOH (0.04%), 100 mM DTT, and whole-cell extracts were separated on 6% or 12% SDS-PAGE gel. For immunofluorescence, cells were fixed with 3.7% PFA at room temperature for 15 minutes, permeabilized with 0.3% Triton X-100 in PBS, and then processed for immunofluorescence as described above**.**

7.3 Chapter 4 Methods

7.3.1 Single-cell model of regulated gene expression

Distributions of transcripts per cell were generated under the steady-state approximation as previously described (279, 280). The basal lognormal regulatory state (Fig. 1*A*, blue) was defined with the following model parameters: $k_{binding}$ = 5, $k_{unbinding}$ = 10, $k_{elongation}$ = 50, $k_{degradation}$ = 1. The exponential regulatory state (Fig. 1*B*, blue) was defined with the following model parameters: $k_{binding}$ = 0.5, $k_{unbinding}$ = 10, $k_{elongation}$ = 50, $k_{degradation}$ = 1. Basal regulatory states were perturbed by increasing $k_{binding}$ by 10-fold (lognormal) or 20-fold (exponential), decreasing $k_{degradation}$ by 3.3-fold (lognormal) or 5-fold (exponential), or both. Probability densities were compared with the lognormal and exponential test distributions by integrating over integer copy numbers to generate a representative observation set. Observations and distributions were compared with the qqplot function in MATLAB (Mathworks).

7.3.2 Simulations of random 10-cell samples

Simulated 10-cell expression profiles were generated in MATLAB with the statistics toolbox or in R. The LN–LN model assumes a binomial distribution for the two regulatory states and a lognormal distribution of the transcripts within each state. For a random *n*-cell sampling (here n = 10), the number of cells drawn from the high regulatory state (h) was specified by a binomial distribution with parameters n and F. Next, h expression measurements were randomly drawn from a lognormal distribution with log-mean μ_1 and log-standard deviation σ. The remaining $n - h$ expression measurements were also drawn from a lognormal distribution with log-mean μ_2 and log-standard deviation σ. The sum of n measurements constituted one stochastic *n*-cell sample. In the EXP–LN model, transcripts from the basal regulatory state were drawn from an exponential distribution with rate parameter λ. This procedure was repeated for the indicated number of random samples.

7.3.3 Derivation of LN–LN maximum likelihood estimator

To derive the LN–LN maximum-likelihood estimator, we began with a mixed population of cells occupying one of two regulatory states. The basal regulatory state expresses a transcript (*g*) at a low level with log-mean $\mu_2^{(g)}$ and log-standard deviation σ. The induced regulatory state expresses the transcript at a higher level with log-mean $\mu_1^{(g)}$ and log-standard deviation σ. The probability of drawing a single cell from the high regulatory state is characterized by the parameter *F*.

According to the two-state model, the single-cell expression for transcript *g* follows the pdf:

$$f_{mixture}^{(g)} = F \cdot f_1^{(g)} + (1-F) \cdot f_2^{(g)} \qquad (1)\,,$$

where $f_2^{(g)}$ and $f_1^{(g)}$ are defined as:

$$f_v^{(g),LN-LN}(x) = \frac{1}{\sqrt{2\pi}\sigma x} \cdot \exp\left\{ -\frac{\left[\log(x) - \mu_v^{(g)}\right]^2}{2\sigma^2} \right\} \qquad \text{for } x > 0 \text{ and } v \in \{1,2\} \qquad (2)\,.$$

The i^{th} random sample of transcript *g*, $Y_i^{(g)}$, is the sum of *n* independent single-cell expression measurements (here, $n = 10$):

$$Y_i^{(g)} = \sum_{j=1}^{n} X_{ij}^{(g)} \qquad (3)\,,$$

where $X_{ij}^{(g)}$ is the expression of transcript *g* in the *j*th cell of the *i*th random sample. Together, the random sample $Y_i^{(g)}$ describing the *n*-cell mixture has the pdf:

$$f_n^{LN\text{-}LN}(y \mid F, \mu_1^{(g)}, \mu_2^{(g)}, \sigma) = \sum_{j=0}^{n} \binom{n}{j} F^j (1-F)^{n-j} f_{(j,n-j)}^{(g),LN-LN}(y) \qquad (4) .$$

$\binom{n}{j} F^j (1-F)^{n-j}$ $\binom{\mathrm{n}}{\mathrm{j}}\,\mathrm{F}^{\mathrm{j}} \cdot (1\text{-}\mathrm{F})^{\mathrm{n\text{-}j}}$ represents the binomial selection of cells from the basal or induced regulatory states with probabilities *F* and 1 – *F*, respectively. $f_{(j,n-j)}^{(g)}$ is the density of a sum $Z_1 + \ldots + Z_n$ of independent random variables representing the *n*-cell draw from the following mixture model:

$$Z_c^{LN-LN} = \begin{cases} LN(\mu_1^{(g)}, \sigma^2) & if \;\; 1 \le c \le j \\ LN(\mu_2^{(g)}, \sigma^2) & if \;\; j < c \le n \end{cases} \qquad (5) .$$

The pdf for the sum of lognormally distributed random variables was approximated as previously described (281).

When expanded to a cluster of *m* transcripts, the log-likelihood function for the model parameters given *k* random *n*-cell samples is:

$$\ell^{LN-LN}(F, \underline{\mu_1}, \underline{\mu_2}, \sigma) = \sum_{g=1}^{m} \sum_{i=1}^{k} \log[f_n^{LN-LN}(y_i^{(g)} \mid F, \mu_1^{(g)}, \mu_2^{(g)}, \sigma)] \qquad (6) ,$$

where $\underline{\mu_1}$ and $\underline{\mu_2}$ are vectors containing the transcript-specific log-means for the two regulatory states: $\underline{\mu_1} = (\mu_1^{(1)}, \ldots, \mu_1^{(m)})$ and $\underline{\mu_2} = (\mu_2^{(1)}, \ldots, \mu_2^{(m)})$. The log-likelihood functions assume that the expression levels of each transcript are independent as defined by the specific mixture model and *F*.

To derive the panel of maximum likelihood estimators, we began with a mixed population of cells occupying one of two regulatory states. The basal regulatory state expresses a transcript (g) at a low level with 1) log-mean $\mu_2^{(g)}$ and log-standard deviation σ for the LN–LN model, 2) log-mean $\mu_2^{(g)}$ and log-standard deviation σ_2 for the relaxed LN–LN model, or 3) mean and standard deviation $(\lambda^{(g)})^{-1}$ for the EXP–LN model. The induced regulatory state expresses the transcript at a higher level with 1) log-mean $\mu_1^{(g)}$ and log-standard deviation σ for the LN–LN model, 2) log-mean $\mu_1^{(g)}$ and log-standard deviation σ_1 for the relaxed LN–LN model, or 3) log-mean $\mu^{(g)}$ and log-standard deviation σ for the EXP–LN model. In all mixture models, the probability of drawing a single cell from the high regulatory state is characterized by the parameter F.

According to the two-state model, the single-cell expression for transcript g follows the pdf:

$$f_{mixture}^{(g)} = F \cdot f_1^{(g)} + (1-F) \cdot f_2^{(g)} \qquad (7)\ ,$$

where $f_2^{(g)}$ and $f_1^{(g)}$ for the LN–LN mixture model are defined as:

$$f_v^{(g)}(x) = \frac{1}{\sqrt{2\pi}\sigma x} \cdot \exp\left\{ -\frac{\left[\log(x) - \mu_v^{(g)}\right]^2}{2\sigma^2} \right\} \qquad \text{for } x > 0 \text{ and } v \in \{1,2\} \qquad (8)\ ,$$

$f_2^{(g)}$ and $f_1^{(g)}$ for the relaxed LN–LN mixture model are defined as:

$$f_v^{(g)}(x) = \frac{1}{\sqrt{2\pi}\sigma_v x} \cdot \exp\left\{ -\frac{\left[\log(x) - \mu_v^{(g)}\right]^2}{2\sigma_v^2} \right\} \qquad \text{for } x > 0 \text{ and } v \in \{1,2\} \qquad (9)\ ,$$

and $f_2^{(g)}$ and $f_1^{(g)}$ for the EXP–LN mixture model are defined as:

$$\begin{aligned} f_2^{(g)}(x) &= \lambda^{(g)} \cdot \exp(-\lambda^{(g)} \cdot x) \qquad \text{for } x \geq 0 \\ f_1^{(g)}(x) &= \frac{1}{\sqrt{2\pi}\sigma x} \cdot \exp\left\{ -\frac{\left[\log(x) - \mu^{(g)}\right]^2}{2\sigma^2} \right\} \qquad \text{for } x > 0 \end{aligned} \qquad (10)\ .$$

The i^{th} random sample of transcript g, $Y_i^{(g)}$, is the sum of n independent single-cell expression measurements (here, $n = 10$):

$$Y_i^{(g)} = \sum_{j=1}^{n} X_{ij}^{(g)} \qquad (11)\ ,$$

where $X_{ij}^{(g)}$ is the expression of transcript g in the j^{th} cell of the i^{th} random sample. Together, the random variable $Y_i^{(g)}$ describing the n-cell LN–LN mixture has the pdf:

$$f_n(y \mid F, \mu_1^{(g)}, \mu_2^{(g)}, \sigma) = \sum_{j=0}^{n} \binom{n}{j} F^j (1-F)^{n-j} f_{(j,n-j)}^{(g)}(y) \qquad (12)\ ,$$

the relaxed LN–LN mixture has the pdf:

$$f_n(y \mid F, \mu_1^{(g)}, \mu_2^{(g)}, \sigma_1, \sigma_2) = \sum_{j=0}^{n} \binom{n}{j} F^j (1-F)^{n-j} f_{(j,n-j)}^{(g)}(y) \qquad (13)\ ,$$

and the EXP–LN mixture has the pdf:

$$f_n(y \mid F, \mu^{(g)}, \lambda^{(g)}, \sigma) = \sum_{j=0}^{n} \binom{n}{j} F^j (1-F)^{n-j} f_{(j,n-j)}^{(g)}(y) \qquad (14)\ .$$

$\binom{n}{j} F^j (1-F)^{n-j}$ $\binom{\mathrm{n}}{\mathrm{j}} \mathrm{F}^{\mathrm{j}} \cdot (1\text{-}\mathrm{F})^{\mathrm{n\text{-}j}}$ represents the binomial selection of cells from the basal or induced regulatory states with probabilities F and $1 - F$, respectively. $f_{(j,n-j)}^{(g)}$ is the density of a sum $Z_1 + \ldots + Z_n$ of independent random variables representing the n-cell draw:

$$Z_c = \begin{cases} LN(\mu_1^{(g)}, \sigma^2) & \text{if } 1 \leq c \leq j \\ LN(\mu_2^{(g)}, \sigma^2) & \text{if } j < c \leq n \end{cases} \qquad \text{(LN–LN mixture)} \qquad (15)\ ,$$

$$Z_c = \begin{cases} LN(\mu_1^{(g)}, \sigma_1^2) & if \;\; 1 \le c \le j \\ LN(\mu_2^{(g)}, \sigma_2^2) & if \;\; j < c \le n \end{cases} \qquad \text{(relaxed LN–LN mixture)} \qquad (16)\ ,$$

$$Z_c = \begin{cases} LN(\mu^{(g)}, \sigma^2) & if \;\; 1 \le c \le j \\ EXP(\lambda^{(g)}) & if \;\; j < c \le n \end{cases} \qquad \text{(EXP–LN mixture)} \qquad (17)\ .$$

The pdf for the sum of log-normally distributed random variables was approximated as previously described(281) and applied to the LN–LN and relaxed LN–LN mixture models. The sum of independent exponentially distributed random variables follows an Erlang distribution(282). The pdf for the EXP–LN mixture model is the convolution of a log-normal and an Erlang density, whose integral was solved numerically.

When expanded to a cluster of *m* transcripts, the log-likelihood function for the model parameters given *k* random *n*-cell samples is:

$$\ell(F, \underline{\mu_1}, \underline{\mu_2}, \sigma) = \sum_{g=1}^{m} \sum_{i=1}^{k} \log[f_n(y_i^{(g)} \mid F, \mu_1^{(g)}, \mu_2^{(g)}, \sigma)] \qquad \text{(LN–LN mixture)} \qquad (18)$$

$$\ell(F, \underline{\mu_1}, \underline{\mu_2}, \sigma_1, \sigma_2) = \sum_{g=1}^{m} \sum_{i=1}^{k} \log[f_n(y_i^{(g)} \mid F, \mu_1^{(g)}, \mu_2^{(g)}, \sigma_1, \sigma_2)] \quad \text{(relaxed LN–LN mixture)} \qquad (19)$$

$$\ell(F, \underline{\mu}, \underline{\lambda}, \sigma) = \sum_{g=1}^{m} \sum_{i=1}^{k} \log[f_n(y_i^{(g)} \mid F, \mu^{(g)}, \lambda^{(g)}, \sigma)] \qquad \text{(EXP–LN mixture)} \qquad (20)$$

where $\underline{\mu_1}$, $\underline{\mu_2}$, $\underline{\mu}$, and $\underline{\lambda}$ are vectors containing the transcript-specific log-means (or inverse means for $\underline{\lambda}$ in EXP–LN mixture) for the two regulatory states: $\underline{\mu_1} = (\mu_1^{(1)}, \ \dots\ , \mu_1^{(m)})$, $\underline{\mu_2} = (\mu_2^{(1)}, \ \dots\ , \mu_2^{(m)})$, $\underline{\mu} = (\mu^{(1)}, \ \dots\ , \mu^{(m)})$, and $\underline{\lambda} = (\lambda^{(1)}, \dots, \lambda^{(m)})$. The log-likelihood functions assume that the expression levels of each transcript are independent within the two regulatory states defined by the specific mixture model and *F*.

7.3.4 Maximum-likelihood parameter estimation and model selection

The derived log-likelihood functions in equations 18–20 are maximized by the most likely combination of parameters for the data $Y_i^{(g)}$. To estimate the parameters for the LN–LN mixtures, we required that $\mu_1^{(1)} > \mu_2^{(1)} \mu_1^{(1)} > \mu_2^{(1)}$. This constraint ensured identifiability because $\ell(F, \underline{\mu_1}, \underline{\mu_2}, \sigma) = \ell(1 - F, \underline{\mu_2}, \underline{\mu_1}, \sigma)$. We also transformed F with the logit function and $\underline{\lambda}$ and σ with the logarithm function to enable the use of faster, unconstrained optimization algorithms.

Because the log-likelihood function was multimodal, it precluded the straightforward use of gradient-based approaches to find globally optimal parameter combinations. We solved the high-dimensional non-convex global optimization problem by combining genetic and simplex algorithms. First, the log-likelihood function was computed at randomly drawn parameter combinations to identify high-likelihood regions in parameter space at computationally low cost. In the regions of highest log-likelihood, we then used the Nelder-Mead algorithm (283) to identify local maxima of the likelihood function. We further localized the global optimum by repeating a random search of parameter space around the optimum identified by the Nelder-Mead algorithm. The resulting high-likelihood regions were used to seed another Nelder-Mead optimization. The iterations of random search and Nelder-Mead optimization continued until convergence.

For estimating model parameters from transcriptional clusters, we first considered smaller subgroups of the cluster of interest. The best balance of computational time and stability of the resulting parameter estimates was achieved with four-gene subgroups (See SI Appendix, fig. S6). The log-likelihood of each subgroup was optimized by the algorithm described above to identify the most-likely parameters for the transcripts in the subgroup. Based on the subgroup estimate, we then kept fixed $\underline{\mu_1}$ and $\underline{\mu_2}$ (for the LN–LN and relaxed LN–LN models) or $\underline{\mu}$ (for the EXP–LN model) and globally inferred F and σ (or F, σ_1, and σ_2 for the relaxed LN–LN model, or $\underline{\lambda}$, F, and σ for the EXP–LN model) by using the optimization algorithm described above. To confirm that the global optimum for each model had been identified, we pursued a constrained optimization in parallel, which required that the two regulatory states were sufficiently distinct from each other. Specifically, the density of the high regulatory state was constrained to be greater than the low regulatory state in the domain between the mode of the high state and the largest observation in the dataset. The likelihoods of the constrained and unconstrained optimizations were compared, and the higher likelihood inference was selected as the best parameterization for that mixture model. Last, the three mixture models were compared according to their BIC score:

$$BIC = -2\ell(\hat{\underline{\theta}}) + c\log(mk) \qquad (7)$$

where $\hat{\underline{\theta}}$ is the vector of inferred parameters, c is the number of inferred parameters in the model (including subgroup inferences), m is the number of transcripts in the cluster, and k is the number of n-cell random samples for each transcript. The best model predicted two distinct regulatory states with the lowest BIC score.

Approximate 95% confidence intervals for the best model were estimated by numerically computing the inverse Hessian matrix of the negative log-likelihood function evaluated at the optimal parameter combination. Each i^{th} diagonal element (d_i) of this matrix leads to the confidence in the i^{th} inferred parameter ($\hat{\theta}_i$) as follows:

$$95\%\ CI_i = \hat{\theta}_i \pm 1.96\sqrt{d_i} \qquad (8)$$

Source code for the maximum-likelihood parameter estimation can be found at http://hmgu.de/icb/StochasticProfiling_ML.

7.3.5 Inference comparisons of one- and ten-cell random samples

We simulated measurements for various gene clusters as described above with either $n = 1$ or $n = 10$, $m = 12$, and $k = 16$ with the mixture model and F specified in Table 1. Values of $\underline{\mu_1}$, $\underline{\mu_2}$, $\underline{\lambda}$, $\underline{\mu}$, and σ were drawn randomly from the individual transcripts comprising the inferences of Fig. 3 *F* and *G* and 4*A*. Model parameters were inferred as described above with the correct value of n in equations 12 and 14. The inference procedure was repeated 100 times, yielding estimates $\hat{\theta}_i^j$ (j = 1, 2, ... 100) for each true parameter θ_i. This gives the following Monte Carlo estimates of bias, variance, and mean-squared error:

$$\begin{aligned} Bias(\hat{\theta}_i) &= \frac{1}{100}\sum_{j=1}^{100}\hat{\theta}_i^j - \theta_i \\ Var(\hat{\theta}_i) &= \frac{1}{99}\sum_{j=1}^{100}\{\hat{\theta}_i^j - \theta_i\}^2 \\ MSE(\hat{\theta}_i) &= Bias(\hat{\theta}_i)^2 + Var(\hat{\theta}_i) \end{aligned} \tag{9}$$

7.3.6 Stochastic sampling

Stochastic samples of *SOD2* were collected as previously described (273, 277, 284). Briefly, 3D cultures were snap frozen and sectioned at day 10 of morphogenesis. Random 10-cell samples of ECM-attached acinar cells were achieved by laser-capture microdissection from cryosections.

The RNA collected from these samples was amplified with a custom small-sample mRNA amplification procedure and quantified by qPCR or microarray (273, 277, 284). Microarray-based expression clusters were identified based on correlated expression fluctuations as described (273, 277).

7.3.7 Digital scoring of expression-frequency index

Multicolor RNA FISH images were acquired with WGA, the riboprobe of interest, and the loading-control riboprobes as described above. Individual ECM-attached cells were manually segmented in ImageJ using the WGA, riboprobe, and loading-control stains to determine cell boundaries. The segmented regions of interest (ROIs) were saved as a single ZIP-file in ImageJ. The pixels within each ROI were extracted and compared against a null pixel distribution comprised of a random set of pixels from segmented cells within the same image. The 85th–95th percentiles of the cell ROI and the null distribution were compared after bootstrapping each distribution 300 times. A cell was scored in the high regulatory state if the 90% bootstrapped CI of the cell ROI was consistently greater than the 90% bootstrapped CI of the null distribution when evaluated from the 85th–95th percentile of pixels. Performing the same analysis on the loading-control riboprobes showed that less than 1% of all cells segmented showed detectable differences in total RNA expression. Therefore, the expression-frequency index for a field of view was quantified as the number of cells detected in the high regulatory state divided by the total number of cells segmented. At least 18 fields of view with at least 10 cells per field were acquired for each gene analyzed. Source code for image analysis can be found at http://hmgu.de/icb/StochasticProfiling_ML.

7.3.8 shRNA cloning and lentiviral RNAi

shRNA sequences against *PIK3CD* were cloned based on the targeting sequences suggested by the RNAi Consortium, except that the *XhoI* restriction site in the shRNA loop was changed to a *PstI* site for easier diagnosis during cloning. shGFP control was used as previously described (47). Primers were annealed at 95ºC in annealing buffer (10 mM Tris-HCl, 100 mM NaCl, 1 mM EDTA) for 5 min on a thermocycler and cooled slowly to room temperature by unplugging the instrument. Annealed primers were phosphorylated in vitro with T4 polynucleotide kinase (New England Biolabs) and then cloned into pLKO.1 puro (274) that had been digested with *EcoR1* and *AgeI*. Lentiviruses were packaged and transduced into MCF10A-5E cells as previously described (277). Stable lines were screened for knockdown efficiency by immunoblotting.

7.3.9 Cell lines

The MCF10A-5E clone was maintained as previously described (73). shGFP and shPIK3CD cell lines were derived by transducing MCF10A-5E cells with lentiviruses and selecting with 2 μg/ml puromycin as previously described (100).

7.3.10 3D culture

MCF10A-5E, shGFP, and shPIK3CD cell lines were cultured in 3D as previously described (98). For p110δ inhibitor experiments, 20 μM IC87114 (Calbiochem) was added to the assay medium and replaced every four days.

7.3.11 Riboprobe synthesis

Digoxigenin or dinitrophenol-labeled riboprobes for the genes of interest were cloned, synthesized, and validated as previously described (73, 100).

7.3.12 RNA FISH and expression-frequency scoring

MCF10A-5E 3D cultures were processed for RNA FISH as previously described (73, 100). Individual ECM-attached cells were scored manually by fluorescence intensity. Experimental estimates of *F* were measured by calculating the number of ECM-attached cells with strong fluorescent signal over basal levels of the probe divided by the entire ECM-attached population in a given tissue section. At least 150 ECM-attached cells were scored per slide.

7.3.13 Frozen section immunofluorescence and expression-frequency scoring

MCF10A-5E 3D cultures were processed for immunofluorescence as previously described (100). Sections were stained for IRF2 (1:200; Santa Cruz) and E-cadherin (1:500; BD Transduction Laboratories) and counterstained with DAPI to visualize nuclei. Experimental estimates of *F* were measured by calculating the number of ECM-attached cells with strong fluorescent signal over basal levels of the probe divided by the entire ECM-attached population in a given tissue section. At least 150 ECM-attached cells were scored per slide.

7.3.14 Confocal microscopy

RNA FISH sections or whole-mount stained 3D cultures were imaged with a 40°—, 1.3 NA oil objective on a Nikon TE-2002-E2 inverted confocal microscope with one 488 nm Ar laser and two HeNe lasers with excitation wavelengths of 543 and 632 nm (Melles Griot). Wheat germ agglutinin (WGA) conjugated to Alexa Fluor 488 (Molecular Probes) was imaged using a 488 nm excitation laser and bandpass emission filter of 515–530 nm.

RNA FISH probes and pRb immunofluorescence samples were imaged using a 543 nm excitation laser and bandpass emission filter of 565–615 nm. DRAQ5 nuclear counterstains and loading control riboprobes (100) were imaged using a 632 nm excitation filter and 650 nm long-pass emission filter. Laser gains were adjusted to avoid widespread saturation of the photomultiplier tubes during imaging.

7.3.15 Whole-mount immunofluorescence and pRb quantification

Whole-mount immunofluorescence of shGFP and shPIK3CD cultures was performed as previously described (100). Samples were stained for phospho-Rb Ser807/811 (1:200; Cell Signaling) and counterstained with DRAQ5 to visualize nuclei. Samples were scored for pRb-positive acini as before (100).

7.3.16 Image processing

For probe validation images, antisense and sense control images were exposure matched and scaled to identical linear ranges of intensity display. Representative images of riboprobe staining and pRb staining were scaled to highlight regions of strong staining.

7.3.17 qPCR

qPCR was performed on cDNA reverse transcribed from MCF10A-5E RNA isolated during 3D culture as previously described (278). Relative copy numbers of *PIK3CA*, *PIK3CB*, and *PIK3CD* were obtained using a common standard of MCF10A-5E genomic DNA.

7.3.18 IC87114 inhibitor validation

MCF10A-5E cells were plated in 6-well dishes and cultured in serum-free medium overnight. Wells were pre-treated with or without 20 µM IC87114 (Calbiochem) for one hour and then stimulated with 10 µM lysophosphatidic acid (Sigma-Aldrich) for five minutes. Cells were lysed as previously described (100), and lysates were probed for pAkt levels by immunoblotting.

7.3.19 Immunoblotting

30–50 µg of clarified cell extract was separated on an 8% or 10% SDS-PAGE gel and transferred to PVDF (Millipore). Membranes were blocked in 0.5°— Odyssey blocking buffer (Licor) for one hour and incubated overnight at 4°C in 0.5°— Odyssey blocking buffer with 0.1% Tween containing one of the following primary antibodies: anti-p110δ (1:500; Santa Cruz Biotechnology) or anti-pAkt Ser473 (1:1000; Cell Signaling). Additionally, chicken anti-tubulin (1:20000; Abcam) was added to each primary solution as a loading control. Membranes were washed four times in phosphate buffered saline + 0.1% Tween (PBS-T) for five minutes each, and incubated in 0.5°— Odyssey + 0.1% Tween + 0.01% SDS with the following secondary antibodies: goat anti-rabbit 800 (1:20000; Licor) and goat anti-chicken 680 (1:20000; Licor). Membranes were washed four times with PBS-T and once with PBS for five minutes each before imaging with an Odyssey detection system (Licor).

7.4 Chapter 5 Methods

7.4.1 Cell culture, 3D spheroid assay, phenotype scoring

Cells were cultured as previously described or by ATCC standards (53). 3D cultures were performed as previously described (53). Constitutive knockdown assays were performed by inducing hairpins three days in 2D before seeding in 3D. Overexpression assays were performed by inducing overexpression 24 hours in 2D before seeding in 3D. Breast cancer cell line spheroids were grown in the same conditions as the MCF10A-5E spheroids or in the ATCC growth media. Cultures were treated with the corresponding concentrations of ligands prepared according to datasheet (Peproech). Phenotypes were scored based on criteria for each phenotype. The premise was done as previous described (53).

7.4.2 Stochastic profiling transcriptional analysis

Stochastic transcriptional profiles were collected in a previous study (73). Measurements were analyzed by creating an average signature of the Z-score profiles of *GDF11*, *TGFBR3*, and *TGFBI*. Every reliably detected transcript was correlated to the aggregate profile by Pearson and Spearman correlation. The top candidates were manually curated and GO-term curated for TGFβ-related genes (285).

7.4.3 Cryosection RNA FISH and protein immunofluorescence

RNA FISH and protein immunofluorescence on spheroid cryosections were performed as previously described (53). Immunofluorescent slides underwent antigen retrieval as previously described. GDF11 antisense probe was used at 20 ng/mL with a 60 degrees Celsius hybridization temperature. EPR, R+D, and 1E6 monoclonal antibodies were used at 1:200 dilution.

7.4.4 Whole mount immunofluorescence

Whole mount immunofluorescence was performed as previous described (53).

7.4.5 RNA sequencing analysis

RNA sequencing data was downloaded from public repositories (27, 211). The reads per million were calculated for the analyzed transcripts.

7.4.6 Immunoblotting

Immunoblotting was peformed as previously described (286). All antibodies used at 1:1000 dilution except for p38, HSP90, V5, FLAG (1:5000) and GAPDH, vinculin, tubulin (1:20000).

7.4.7 Transcriptional profiling

MCF10A-5E or NMuMG spheroids were cultured for six days in 3D. Each cell line was stimulated with 250 ng/mL GDF11 or 50 ng/mL TGFβ1 for 4 hours. RNA was collected as previously described (100). RNA was profiled using Illumina BeadChip arrays and analyzed using the lumi R package.

7.4.8 Intraductal imaging and bioluminescence imaging

Intraductal injectiions and bioluminescence imaging were performed as previously described (53). Knockdowns were induced by placing the mice on doxycycline chow (625 mg/kg) after two weeks of standard chow. GDF11 inocluation was performed by diluting 1 mg/mL GDF11 (Peprotech) into the cell suspension (final concentration: 100 μg/mL).

7.4.9 Clinical immunohistochemistry and scoring

Clinical immunohistochemistry was performed as previously described (53). 1E6 foci were counted across every 4x field of view with normal or tumor tissue present.

7.4.10 Anitbody validation

Cell pellets from 293Ts were transfected with LacZ or GDF11 overexpression and GDF11 hairpins. These pellets were paraffin embedded. Slides were stained with EPR antibody similarly as the clinical specimens. Cell pellets from MDA-MB-231 cells were transfected with GDF11 hairpins, prepared as cell pellets, and stained for 1E6 antibody. Whole cell lysates were degylcosylated with the New England Biolabs Protein Deglycosylation Kit.

7.4.11 Conditioned medium analysis

0.5 micrograms of anti-V5 antibody was co-inbuated with 500 microliters of conditioned media overnight at 4 degrees Celsius. Protein A/G beads purify the antibody, and the sample is examined by Western blot. TGFβ1 ELISA was performed as described by the purchased kit (R+D systems).

7.4.12 Plasmids

Hairpin plasmids were cloned as previously described (96), with targeting sequences identified from the RNAi consortium. GDF11 overexpression constructs were obtained by recombining a GDF11 expression plasmid from Arizona State into the panel of destination vectors. SMAD4 and ID2 overexpression plasmids were obtained by recombining the coding sequnces of those two genes cloned into a 3xFLAG pEN_TT_mirC2 plasmid and pSLIK_NEO.

8 References

1. Altschuler SJ & Wu LF (2010) Cellular heterogeneity: do differences make a difference? *Cell* 141(4):559-563.
2. Elowitz MB, Levine AJ, Siggia ED, & Swain PS (2002) Stochastic gene expression in a single cell. *Science* 297(5584):1183-1186.
3. Raj A, Peskin CS, Tranchina D, Vargas DY, & Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. *Plos Biol* 4(10):e309.
4. Raj A & van Oudenaarden A (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* 135(2):216-226.
5. Dunlop MJ, Cox RS, 3rd, Levine JH, Murray RM, & Elowitz MB (2008) Regulatory activity revealed by dynamic correlations in gene expression noise. *Nat Genet* 40(12):1493-1498.
6. Rosenfeld N, Young JW, Alon U, Swain PS, & Elowitz MB (2005) Gene regulation at the single-cell level. *Science* 307(5717):1962-1965.
7. Paul F*, et al.* (2015) Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* 163(7):1663-1677.
8. Chang HH, Hemberg M, Barahona M, Ingber DE, & Huang S (2008) Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* 453(7194):544-U510.
9. Morris SA*, et al.* (2010) Origin and formation of the first two distinct cell types of the inner cell mass in the mouse embryo. *Proc Natl Acad Sci U S A* 107(14):6364-6369.
10. Guo G*, et al.* (2010) Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev Cell* 18(4):675-685.
11. Claveria C, Giovinazzo G, Sierra R, & Torres M (2013) Myc-driven endogenous cell competition in the early mammalian embryo. *Nature* 500(7460):39-44.
12. Bruce AW & Zernicka-Goetz M (2010) Developmental control of the early mammalian embryo: competition among heterogeneous cells that biases cell fate. *Current opinion in genetics & development* 20(5):485-491.
13. Tang F*, et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 6(5):377-382.
14. Hayashi K, Lopes SM, Tang F, & Surani MA (2008) Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. *Cell Stem Cell* 3(4):391-401.
15. Kumar RM*, et al.* (2014) Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* 516(7529):56-61.
16. Zunder ER, Lujan E, Goltsev Y, Wernig M, & Nolan GP (2015) A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. *Cell Stem Cell* 16(3):323-337.
17. Lujan E*, et al.* (2015) Early reprogramming regulators identified by prospective isolation and mass cytometry. *Nature* 521(7552):352-356.

18. Hercend T, Reinherz EL, Meuer S, Schlossman SF, & Ritz J (1983) Phenotypic and functional heterogeneity of human cloned natural killer cell lines. *Nature* 301(5896):158-160.
19. Shalek AK*, et al.* (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498(7453):236-240.
20. Bendall SC*, et al.* (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* 332(6030):687-696.
21. Shalek AK*, et al.* (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 510(7505):363-369.
22. Raj A, Rifkin SA, Andersen E, & van Oudenaarden A (2010) Variability in gene expression underlies incomplete penetrance. *Nature* 463(7283):913-918.
23. Spencer SL, Gaudet S, Albeck JG, Burke JM, & Sorger PK (2009) Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* 459(7245):428-432.
24. Burrell RA, McGranahan N, Bartek J, & Swanton C (2013) The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501(7467):338-345.
25. Network TCGA (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* 368(22):2059-2074.
26. Network TCGA (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499(7456):43-49.
27. Cancer Genome Atlas N (2012) Comprehensive molecular portraits of human breast tumours. *Nature* 490(7418):61-70.
28. Cancer Genome Atlas N (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487(7407):330-337.
29. Network TCGA (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487(7407):330-337.
30. Network TCGA (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489(7417):519-525.
31. Cancer Genome Atlas Research N (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* 474(7353):609-615.
32. Shah NP*, et al.* (2007) Sequential ABL kinase inhibitor therapy selects for compound drug-resistant BCR-ABL mutations with altered oncogenic potency. *J Clin Invest* 117(9):2562-2569.
33. Marusyk A, Almendro V, & Polyak K (2012) Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer* 12(5):323-334.
34. Andor N*, et al.* (2016) Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med* 22(1):105-113.
35. Gerlinger M*, et al.* (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 366(10):883-892.
36. Zhang J*, et al.* (2014) Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* 346(6206):256-259.
37. Navin N*, et al.* (2011) Tumour evolution inferred by single-cell sequencing. *Nature* 472(7341):90-94.
38. Wang Y*, et al.* (2014) Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 512(7513):155-160.
39. Bhang HE*, et al.* (2015) Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nat Med* 21(5):440-448.

40. Maley CC, *et al.* (2006) Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet* 38(4):468-473.
41. Shah SP, *et al.* (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* 486(7403):395-399.
42. Almendro V, *et al.* (2014) Inference of Tumor Evolution during Chemotherapy by Computational Modeling and In Situ Analysis of Genetic and Phenotypic Cellular Diversity. *Cell Rep* 6(3):514-527.
43. Thorlacius S, *et al.* (1996) A single BRCA2 mutation in male and female breast cancer families from Iceland with varied cancer phenotypes. *Nat Genet* 13(1):117-119.
44. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, & Kehrer-Sawatzki H (2013) Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet* 132(10):1077-1130.
45. Paszek MJ, *et al.* (2005) Tensional homeostasis and the malignant phenotype. *Cancer Cell* 8(3):241-254.
46. Fukumura D, *et al.* (1998) Tumor induction of VEGF promoter activity in stromal cells. *Cell* 94(6):715-725.
47. Orimo A, *et al.* (2005) Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated SDF-1/CXCL12 secretion. *Cell* 121(3):335-348.
48. Karnoub AE, *et al.* (2007) Mesenchymal stem cells within tumour stroma promote breast cancer metastasis. *Nature* 449(7162):557-563.
49. Dey SS, Kester L, Spanjaard B, Bienko M, & van Oudenaarden A (2015) Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol* 33(3):285-289.
50. Patel AP, *et al.* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344(6190):1396-1401.
51. Cleary AS, Leonard TL, Gestl SA, & Gunther EJ (2014) Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers. *Nature* 508(7494):113-117.
52. Gupta PB, *et al.* (2011) Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* 146(4):633-644.
53. Wang CC, Bajikar SS, Jamal L, Atkins KA, & Janes KA (2014) A time- and matrix-dependent TGFBR3-JUND-KRT5 regulatory circuit in single breast epithelial cells and basal-like premalignancies. *Nat Cell Biol* 16(4):345-356.
54. Zhao B, Pritchard JR, Lauffenburger DA, & Hemann MT (2014) Addressing Genetic Tumor Heterogeneity through Computationally Predictive Combination Therapy. *Cancer discovery* 4(2):166-174.
55. Heppner GH (1984) Tumor heterogeneity. *Cancer Res* 44(6):2259-2265.
56. Sorlie T, *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 98(19):10869-10874.
57. Perou CM, *et al.* (2000) Molecular portraits of human breast tumours. *Nature* 406(6797):747-752.
58. Nielsen TO, *et al.* (2004) Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin Cancer Res* 10(16):5367-5374.
59. van de Rijn M, *et al.* (2002) Expression of cytokeratins 17 and 5 identifies a group of breast carcinomas with poor clinical outcome. *Am J Pathol* 161(6):1991-1996.

60. Rakha EA, Reis-Filho JS, & Ellis IO (2008) Basal-like breast cancer: a critical review. *J Clin Oncol* 26(15):2568-2581.

61. Prat A*, et al.* (2010) Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast cancer research : BCR* 12(5):R68.

62. Dent R*, et al.* (2007) Triple-negative breast cancer: clinical features and patterns of recurrence. *Clin Cancer Res* 13:4429-4434.

63. Millikan RC*, et al.* (2008) Epidemiology of basal-like breast cancer. *Breast Cancer Res Tr* 109(1):123-139.

64. Higgins MJ & Baselga J (2011) Targeted therapies for breast cancer. *J Clin Invest* 121(10):3797-3803.

65. Lehmann BD*, et al.* (2011) Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest* 121(7):2750-2767.

66. Lane DP (1992) Cancer. p53, guardian of the genome. *Nature* 358(6381):15-16.

67. Ding L*, et al.* (2010) Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* 464(7291):999-1005.

68. Banerji S*, et al.* (2012) Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* 486(7403):405-409.

69. Guha M (2011) PARP inhibitors stumble in breast cancer. *Nat Biotechnol* 29(5):373-374.

70. Wang CC & Janes KA (2014) Non-genetic heterogeneity caused by differential single-cell adhesion. *Cell Cycle* 13(14):2149-2150.

71. Laakso M*, et al.* (2006) Basoluminal Carcinoma: A New Biologically and Prognostically Distinct Entity Between Basal and Luminal Breast Cancer. *Clin Cancer Res* 12(14 Pt 1):4185-4191.

72. Wang L & Janes KA (2013) Stochastic profiling of transcriptional regulatory heterogeneities in tissues, tumors and cultured cells. *Nature Protocols* 8(2):282-301.

73. Janes KA, Wang CC, Holmberg KJ, Cabral K, & Brugge JS (2010) Identifying single-cell molecular programs by stochastic profiling. *Nat Methods* 7(4):311-317.

74. Levsky JM & Singer RH (2003) Gene expression and the myth of the average cell. *Trends Cell Biol* 13(1):4-6.

75. Hulett HR, Bonner WA, Barrett J, & Herzenberg LA (1969) Cell sorting: automated separation of mammalian cells as a function of intracellular fluorescence. *Science* 166(3906):747-749.

76. De Rosa SC, Herzenberg LA, Herzenberg LA, & Roederer M (2001) 11-color, 13-parameter flow cytometry: identification of human naive T cells by phenotype, function, and T-cell receptor diversity. *Nat Med* 7(2):245-248.

77. Quan Y*, et al.* (2012) Impact of cell dissociation on identification of breast cancer stem cells. *Cancer Biomark* 12(3):125-133.

78. Hughes AJ*, et al.* (2014) Single-cell western blotting. *Nat Methods* 11(7):749-755.

79. Giesen C*, et al.* (2014) Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods* 11(4):417-422.

80. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, & Tyagi S (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* 5(10):877-879.

81. Levsky JM, Shenoy SM, Pezo RC, & Singer RH (2002) Single-cell gene expression profiling. *Science* 297(5582):836-840.

82. Battich N, Stoeger T, & Pelkmans L (2013) Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nat Methods* 10(11):1127-1133.

83. Battich N, Stoeger T, & Pelkmans L (2015) Control of Transcript Variability in Single Mammalian Cells. *Cell* 163(7):1596-1610.

84. Chen KH, Boettiger AN, Moffitt JR, Wang S, & Zhuang X (2015) RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348(6233):aaa6090.

85. Sanchez-Freire V, Ebert AD, Kalisky T, Quake SR, & Wu JC (2012) Microfluidic single-cell real-time PCR for comparative analysis of gene expression patterns. *Nat Protoc* 7(5):829-838.

86. Citri A, Pang ZP, Sudhof TC, Wernig M, & Malenka RC (2012) Comprehensive qPCR profiling of gene expression in single neuronal cells. *Nat Protoc* 7(1):118-127.

87. Macosko EZ*, et al.* (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161(5):1202-1214.

88. Jaitin DA*, et al.* (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343(6172):776-779.

89. Ramskold D*, et al.* (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* 30(8):777-782.

90. Hashimshony T, Wagner F, Sher N, & Yanai I (2012) CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* 2(3):666-673.

91. Tang F*, et al.* (2010) RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc* 5(3):516-535.

92. Picelli S*, et al.* (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* 10(11):1096-1098.

93. Sasagawa Y*, et al.* (2013) Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol* 14(4):R31.

94. Grün D*, et al.* (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*.

95. Reiter M*, et al.* (2011) Quantification noise in single cell experiments. *Nucleic acids research* 39(18):e124.

96. Bajikar SS, Fuchs C, Roller A, Theis FJ, & Janes KA (2014) Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles. *Proc Natl Acad Sci U S A* 111(5):E626-635.

97. Debnath J & Brugge JS (2005) Modelling glandular epithelial cancers in three-dimensional cultures. *Nat Rev Cancer* 5(9):675-688.

98. Debnath J, Muthuswamy SK, & Brugge JS (2003) Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods* 30(3):256-268.

99. Debnath J*, et al.* (2002) The role of apoptosis in creating and maintaining luminal space within normal and oncogene-expressing mammary acini. *Cell* 111(1):29-40.

100. Wang L, Brugge JS, & Janes KA (2011) 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* 108(40):E803-812.

101. Snijder B*, et al.* (2009) Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature* 461(7263):520-523.

102. Kenny PA*, et al.* (2007) The morphologies of breast cancer cell lines in three-dimensional assays correlate with their profiles of gene expression. *Mol Oncol* 1(1):84-96.

103. Lee GY, Kenny PA, Lee EH, & Bissell MJ (2007) Three-dimensional culture models of normal and malignant breast epithelial cells. *Nat Methods* 4(4):359-365.

104. Neve RM*, et al.* (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10(6):515-527.

105. Barretina J*, et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483(7391):603-607.

106. Slack MD, Martinez ED, Wu LF, & Altschuler SJ (2008) Characterizing heterogeneous cellular responses to perturbations. *Proc Natl Acad Sci U S A* 105(49):19306-19311.

107. Gao D*, et al.* (2014) Organoid cultures derived from patients with advanced prostate cancer. *Cell* 159(1):176-187.

108. Koo BK*, et al.* (2012) Controlled gene expression in primary Lgr5 organoid cultures. *Nat Methods* 9(1):81-83.

109. Petersen OW, Ronnov-Jessen L, Howlett AR, & Bissell MJ (1992) Interaction with basement membrane serves to rapidly distinguish growth and differentiation pattern of normal and malignant human breast epithelial cells. *Proc Natl Acad Sci U S A* 89(19):9064-9068.

110. Muthuswamy SK, Li DM, Lelievre S, Bissell MJ, & Brugge JS (2001) ErbB2, but not ErbB1, reinitiates proliferation and induces luminal repopulation in epithelial acini. *Nat Cell Biol* 3(9):785-792.

111. Martin KJ, Patrick DR, Bissell MJ, & Fournier MV (2008) Prognostic breast cancer signature identified from 3D culture model accurately predicts clinical outcome across independent datasets. *PLoS One* 3(8):e2994.

112. Han J*, et al.* (2010) Molecular predictors of 3D morphogenesis by breast cancer cell lines in 3D culture. *PLoS Comput Biol* 6(2):e1000684.

113. Kaipparettu BA*, et al.* (2008) Novel egg white-based 3-D cell culture system. *BioTechniques* 45(2):165-168, 170-161.

114. Huang L*, et al.* (2015) Ductal pancreatic cancer modeling and drug screening using human pluripotent stem cell- and patient-derived tumor organoids. *Nat Med* 21(11):1364-1371.

115. Shamir L, Delaney JD, Orlov N, Eckley DM, & Goldberg IG (2010) Pattern recognition software and techniques for biological image analysis. *PLoS Comput Biol* 6(11):e1000974.

116. Loo LH*, et al.* (2009) An approach for extensibly profiling the molecular states of cellular subpopulations. *Nat Methods* 6(10):759-765.

117. Perlman ZE*, et al.* (2004) Multidimensional drug profiling by automated microscopy. *Science* 306(5699):1194-1198.

118. Machacek M*, et al.* (2009) Coordination of Rho GTPase activities during cell protrusion. *Nature* 461(7260):99-103.

119. Stegmaier J*, et al.* (2016) Real-Time Three-Dimensional Cell Segmentation in Large-Scale Microscopy Data of Developing Embryos. *Dev Cell* 36(2):225-240.

120. Debnath J, Walker SJ, & Brugge JS (2003) Akt activation disrupts mammary acinar architecture and enhances proliferation in an mTOR-dependent manner. *J Cell Biol* 163(2):315-326.

121. Miller FR, Santner SJ, Tait L, & Dawson PJ (2000) MCF10DCIS.com xenograft model of human comedo ductal carcinoma in situ. *Journal of the National Cancer Institute* 92(14):1185-1186.

122. Jensen KJ & Janes KA (2012) Modeling the latent dimensions of multivariate signaling datasets. *Physical biology* 9(4):045004.

123. Volk-Draper LD, Rajput S, Hall KL, Wilber A, & Ran S (2012) Novel model for basaloid triple-negative breast cancer: behavior in vivo and response to therapy. *Neoplasia* 14(10):926-942.

124. Noble WS (2006) What is a support vector machine? *Nat Biotechnol* 24(12):1565-1567.

125. Nidhi, Glick M, Davies JW, & Jenkins JL (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* 46(3):1124-1133.

126. Danuser G (2011) Computer vision in cell biology. *Cell* 147(5):973-978.

127. Jaqaman K*, et al.* (2008) Robust single-particle tracking in live-cell time-lapse sequences. *Nat Methods* 5(8):695-702.

128. van de Wetering M*, et al.* (2015) Prospective derivation of a living organoid biobank of colorectal cancer patients. *Cell* 161(4):933-945.

129. Boj SF*, et al.* (2015) Organoid models of human and mouse ductal pancreatic cancer. *Cell* 160(1-2):324-338.

130. Wernet MF*, et al.* (2006) Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature* 440(7081):174-180.

131. Laslo P*, et al.* (2006) Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell* 126(4):755-766.

132. Shipitsin M*, et al.* (2007) Molecular definition of breast tumor heterogeneity. *Cancer Cell* 11(3):259-273.

133. Singh DK*, et al.* (2010) Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. *Mol Syst Biol* 6:369.

134. Sharma SV*, et al.* (2010) A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* 141(1):69-80.

135. Wang CC, Jamal L, & Janes KA (2012) Normal morphogenesis of epithelial tissues and progression of epithelial tumors. *Wiley Interdiscip Rev Syst Biol Med* 4(1):51-78.

136. Ewald AJ, Brenot A, Duong M, Chan BS, & Werb Z (2008) Collective epithelial migration and cell rearrangements drive mammary branching morphogenesis. *Dev Cell* 14(4):570-581.

137. Sato T*, et al.* (2009) Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature* 459(7244):262-265.

138. Shackleton M*, et al.* (2006) Generation of a functional mammary gland from a single stem cell. *Nature* 439(7072):84-88.

139. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, & Clarke MF (2003) Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci U S A* 100(7):3983-3988.

140. Wang XF*, et al.* (1991) Expression cloning and characterization of the TGF-beta type III receptor. *Cell* 67(4):797-805.

141. Nakashima M, Toyono T, Akamine A, & Joyner A (1999) Expression of growth/differentiation factor 11, a new member of the BMP/TGFbeta superfamily during mouse embryogenesis. *Mech Dev* 80(2):185-189.

142. Ahmed AA*,* *et al.* (2007) The extracellular matrix protein TGFBI induces microtubule stabilization and sensitizes ovarian cancers to paclitaxel. *Cancer Cell* 12(6):514-527.

143. Schmelzle T*,* *et al.* (2007) Functional role and oncogene-regulated expression of the BH3-only factor Bmf in mammary epithelial anoikis and morphogenesis. *Proc Natl Acad Sci U S A* 104(10):3787-3792.

144. Kang BH, Jensen KJ, Hatch JA, & Janes KA (2013) Simultaneous profiling of 194 distinct receptor transcripts in human cells. *Sci Signal* 6(287):rs13.

145. Tyson JJ, Chen KC, & Novak B (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr Opin Cell Biol* 15(2):221-231.

146. Berger I & Shaul Y (1991) Structure and function of human jun-D. *Oncogene* 6(4):561-566.

147. Hempel N*,* *et al.* (2008) Expression of the type III TGF-beta receptor is negatively regulated by TGF-beta. *Carcinogenesis* 29(5):905-912.

148. Tsai TY*,* *et al.* (2008) Robust, tunable biological oscillations from interlinked positive and negative feedback loops. *Science* 321(5885):126-129.

149. Nagai T*,* *et al.* (2002) A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nat Biotechnol* 20(1):87-90.

150. Li X*,* *et al.* (1998) Generation of destabilized green fluorescent protein as a transcription reporter. *J Biol Chem* 273(52):34970-34975.

151. Dantuma NP, Lindsten K, Glas R, Jellne M, & Masucci MG (2000) Short-lived green fluorescent proteins for quantifying ubiquitin/proteasome-dependent proteolysis in living cells. *Nat Biotechnol* 18(5):538-543.

152. Kearns JD, Basak S, Werner SL, Huang CS, & Hoffmann A (2006) IkappaBepsilon provides negative feedback to control NF-kappaB oscillations, signaling dynamics, and inflammatory gene expression. *J Cell Biol* 173(5):659-664.

153. Zou T*,* *et al.* (2010) Polyamines regulate the stability of JunD mRNA by modulating the competitive binding of its 3' untranslated region to HuR and AUF1. *Mol Cell Biol* 30(21):5021-5032.

154. Schwanhausser B*,* *et al.* (2011) Global quantification of mammalian gene expression control. *Nature* 473(7347):337-342.

155. Bryan BB, Schnitt SJ, & Collins LC (2006) Ductal carcinoma in situ with basal-like phenotype: a possible precursor to invasive basal-like breast cancer. *Modern Pathol* 19(5):617-621.

156. Dong M*,* *et al.* (2007) The type III TGF-beta receptor suppresses breast cancer progression. *J Clin Invest* 117(1):206-217.

157. Azzopardi JG, Ahmed A, & Millis RR (1979) Problems in breast pathology. *Major Probl Pathol* 11:i-xvi, 1-466.

158. Reginato MJ*,* *et al.* (2003) Integrins and EGFR coordinately regulate the pro-apoptotic protein Bim to prevent anoikis. *Nat Cell Biol* 5(8):733-740.

159. Muthuswamy SK, Li D, Lelievre S, Bissell MJ, & Brugge JS (2001) ErbB2, but not ErbB1, reinitiates proliferation and induces luminal repopulation in epithelial acini. *Nat Cell Biol* 3(9):785-792.

160. Bhowmick NA, Neilson EG, & Moses HL (2004) Stromal fibroblasts in cancer initiation and progression. *Nature* 432(7015):332-337.

161. Candi E, Schmidt R, & Melino G (2005) The cornified envelope: a model of cell death in the skin. *Nat Rev Mol Cell Biol* 6(4):328-340.

162. Uhlen M*, et al.* (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics* 4(12):1920-1932.
163. Uhlen M*, et al.* (2010) Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* 28(12):1248-1250.
164. Humphries JD, Byron A, & Humphries MJ (2006) Integrin ligands at a glance. *J Cell Sci* 119(Pt 19):3901-3903.
165. Jones PL & Jones FS (2000) Tenascin-C in development and disease: gene regulation and cell function. *Matrix Biol* 19(7):581-596.
166. Oskarsson T*, et al.* (2011) Breast cancer cells produce tenascin C as a metastatic niche component to colonize the lungs. *Nat Med* 17(7):867-874.
167. Schalkwijk J*, et al.* (1991) Tenascin expression in human dermis is related to epidermal proliferation. *Am J Pathol* 139(5):1143-1150.
168. Thorne BC, Bailey AM, & Peirce SM (2007) Combining experiments with multi-cell agent-based modeling to study biological tissue patterning. *Brief Bioinform* 8(4):245-257.
169. Schwartz MA & Assoian RK (2001) Integrins and cell proliferation: regulation of cyclin-dependent kinases via cytoplasmic signaling pathways. *J Cell Sci* 114(Pt 14):2553-2560.
170. Shi M*, et al.* (2011) Latent TGF-beta structure and activation. *Nature* 474(7351):343-349.
171. Mailleux AA*, et al.* (2007) BIM regulates apoptosis during mammary ductal morphogenesis, and its absence reveals alternative cell death mechanisms. *Dev Cell* 12(2):221-234.
172. Ishihara A, Yoshida T, Tamaki H, & Sakakura T (1995) Tenascin expression in cancer cells and stroma of human breast cancer and its prognostic significance. *Clin Cancer Res* 1(9):1035-1041.
173. Cheung KJ, Gabrielson E, Werb Z, & Ewald AJ (2013) Collective invasion in breast cancer requires a conserved Basal epithelial program. *Cell* 155(7):1639-1651.
174. Tyson DR, Garbett SP, Frick PL, & Quaranta V (2012) Fractional proliferation: a method to deconvolve cell population dynamics from single-cell data. *Nat Methods* 9(9):923-928.
175. Dalerba P*, et al.* (2011) Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* 29(12):1120-1127.
176. Taniguchi K, Kajiyama T, & Kambara H (2009) Quantitative analysis of gene expression in a single cell by qPCR. *Nat Methods* 6(7):503-506.
177. Lubeck E & Cai L (2012) Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat Methods*.
178. Kurimoto K*, et al.* (2006) An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res* 34(5):e42.
179. Tietjen I*, et al.* (2003) Single-cell transcriptional analysis of neuronal progenitors. *Neuron* 38(2):161-175.
180. Hansen KD, Wu Z, Irizarry RA, & Leek JT (2011) Sequencing technology does not eliminate biological variability. *Nat Biotechnol* 29(7):572-573.
181. Riedel N & Berg J (2013) Statistical mechanics approach to the sample deconvolution problem. *Phys Rev E* 87(4):042715.
182. Shen-Orr SS*, et al.* (2010) Cell type-specific gene expression differences in complex tissues. *Nat Methods* 7(4):287-289.

183. Gong T*, et al.* (2011) Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS One* 6(11):e27156.

184. Loo LH*, et al.* (2009) Heterogeneity in the physiological states and pharmacological responses of differentiating 3T3-L1 preadipocytes. *Journal of Cell Biology* 187(3):375-384.

185. Bengtsson M, Stahlberg A, Rorsman P, & Kubista M (2005) Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res* 15(10):1388-1392.

186. Warren L, Bryder D, Weissman IL, & Quake SR (2006) Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proc Natl Acad Sci U S A* 103(47):17807-17812.

187. Peccoud J & Ycart B (1995) Markovian Modeling of Gene-Product Synthesis. *Theor Popul Biol* 48(2):222-234.

188. Munsky B, Neuert G, & van Oudenaarden A (2012) Using gene expression noise to understand gene regulation. *Science* 336(6078):183-187.

189. Gaestel M (2006) MAPKAP kinases - MKs - two's company, three's a crowd. *Nat Rev Mol Cell Biol* 7(2):120-130.

190. Limpert E, Stahel WA, & Abbt M (2001) Log-normal distributions across the sciences: Keys and clues. *Bioscience* 51(5):341-352.

191. Newman JR*, et al.* (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* 441(7095):840-846.

192. Bar-Even A*, et al.* (2006) Noise in protein expression scales with natural protein abundance. *Nat Genet* 38(6):636-643.

193. Stewart-Ornstein J, Weissman JS, & El-Samad H (2012) Cellular noise regulons underlie fluctuations in saccharomyces cerevisiae. *Mol Cell* 45(4):483-493.

194. Seal S*, et al.* (2006) Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat Genet* 38(11):1239-1241.

195. Doherty GM, Boucher L, Sorenson K, & Lowney J (2001) Interferon regulatory factor expression in human breast cancer. *Ann Surg* 233(5):623-629.

196. Fujii H*, et al.* (2005) Frequent down-regulation of HIVEP2 in human breast cancer. *Breast Cancer Res Tr* 91(2):103-112.

197. Foukas LC, Berenjeno IM, Gray A, Khwaja A, & Vanhaesebroeck B (2010) Activity of any class IA PI3K isoform can sustain cell proliferation and survival. *Proc Natl Acad Sci U S A* 107(25):11381-11386.

198. Knight ZA*, et al.* (2006) A pharmacological map of the PI3-K family defines a role for p110alpha in insulin signaling. *Cell* 125(4):733-747.

199. Erkkila T*, et al.* (2010) Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics* 26(20):2571-2577.

200. Repsilber D*, et al.* (2010) Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics* 11:27.

201. Tolliver D, Tsourakakis C, Subramanian A, Shackney S, & Schwartz R (2010) Robust unmixing of tumor states in array comparative genomic hybridization data. *Bioinformatics* 26(12):i106-114.

202. Ciriello G*, et al.* (2015) Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* 163(2):506-519.

203. Kurbel S (2013) In search of triple-negative DCIS: tumor-type dependent model of breast cancer progression from DCIS to the invasive cancer. *Tumour Biol* 34(1):1-7.

204. Li B, Wen G, Zhao Y, Tong J, & Hei TK (2012) The role of TGFBI in mesothelioma and breast cancer: association with tumor suppression. *BMC cancer* 12:239.

205. Wen G*, et al.* (2011) TGFBI expression reduces in vitro and in vivo metastatic potential of lung and breast tumor cells. *Cancer Lett* 308(1):23-32.

206. Loffredo FS*, et al.* (2013) Growth differentiation factor 11 is a circulating factor that reverses age-related cardiac hypertrophy. *Cell* 153(4):828-839.

207. McPherron AC, Lawler AM, & Lee SJ (1999) Regulation of anterior/posterior patterning of the axial skeleton by growth/differentiation factor 11. *Nat Genet* 22(3):260-264.

208. Andersson O, Reissmann E, & Ibanez CF (2006) Growth differentiation factor 11 signals through the transforming growth factor-beta receptor ALK5 to regionalize the anterior-posterior axis. *EMBO reports* 7(8):831-837.

209. Egerman MA*, et al.* (2015) GDF11 Increases with Age and Inhibits Skeletal Muscle Regeneration. *Cell Metab* 22(1):164-174.

210. Heiser LM*, et al.* (2012) Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc Natl Acad Sci U S A* 109(8):2724-2729.

211. Daemen A*, et al.* (2013) Modeling precision treatment of breast cancer. *Genome Biol* 14(10):R110.

212. Piek E, Moustakas A, Kurisaki A, Heldin CH, & ten Dijke P (1999) TGF-(beta) type I receptor/ALK-5 and Smad proteins mediate epithelial to mesenchymal transdifferentiation in NMuMG breast epithelial cells. *J Cell Sci* 112 (Pt 24):4557-4568.

213. Miettinen PJ, Ebner R, Lopez AR, & Derynck R (1994) TGF-beta induced transdifferentiation of mammary epithelial cells to mesenchymal cells: involvement of type I receptors. *J Cell Biol* 127(6 Pt 2):2021-2036.

214. Liu JS, Farlow JT, Paulson AK, Labarge MA, & Gartner ZJ (2012) Programmed cell-to-cell variability in Ras activity triggers emergent behaviors during mammary epithelial morphogenesis. *Cell Rep* 2(5):1461-1470.

215. Derynck R & Miyazono K (2008) TGF-ß and the TGF-ß Family. *The TGF-ß Family,* Cold Spring Harbor Laboratory Monograph Series, eds Derynck R & Miyazono K (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), p 1114.

216. Shamir ER*, et al.* (2014) Twist1-induced dissemination preserves epithelial identity and requires E-cadherin. *J Cell Biol* 204(5):839-856.

217. Lagna G, Hata A, Hemmati-Brivanlou A, & Massague J (1996) Partnership between DPC4 and SMAD proteins in TGF-beta signalling pathways. *Nature* 383(6603):832-836.

218. Fink SP, Mikkola D, Willson JK, & Markowitz S (2003) TGF-beta-induced nuclear localization of Smad2 and Smad3 in Smad4 null cancer cell lines. *Oncogene* 22(9):1317-1323.

219. Basolo F*, et al.* (1994) Response of normal and oncogene-transformed human mammary epithelial cells to transforming growth factor beta 1 (TGF-beta 1): lack of growth-inhibitory effect on cells expressing the simian virus 40 large-T antigen. *Int J Cancer* 56(5):736-742.

220. Gupta GP*, et al.* (2007) ID genes mediate tumor reinitiation during breast cancer lung metastasis. *Proc Natl Acad Sci U S A* 104(49):19506-19511.

221. Stighall M, Manetopoulos C, Axelson H, & Landberg G (2005) High ID2 protein expression correlates with a favourable prognosis in patients with primary breast cancer and reduces cellular invasiveness of breast cancer cells. *Int J Cancer* 115(3):403-411.

222. Behbod F*, et al.* (2009) An intraductal human-in-mouse transplantation model mimics the subtypes of ductal carcinoma in situ. *Breast cancer research : BCR* 11(5):R66.

223. Nosek BA*, et al.* (2015) SCIENTIFIC STANDARDS. Promoting an open research culture. *Science* 348(6242):1422-1425.

224. Open Science C (2015) PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716.

225. Ge G, Hopkins DR, Ho WB, & Greenspan DS (2005) GDF11 forms a bone morphogenetic protein 1-activated latent complex that can modulate nerve growth factor-induced differentiation of PC12 cells. *Mol Cell Biol* 25(14):5846-5858.

226. Essalmani R*, et al.* (2008) In vivo functions of the proprotein convertase PC5/6 during mouse development: Gdf11 is a likely substrate. *Proc Natl Acad Sci U S A* 105(15):5750-5755.

227. De Bie I*, et al.* (1996) The isoforms of proprotein convertase PC5 are sorted to different subcellular compartments. *J Cell Biol* 135(5):1261-1275.

228. Chen X*, et al.* (2014) XBP1 promotes triple-negative breast cancer by controlling the HIF1alpha pathway. *Nature* 508(7494):103-107.

229. Hammond C & Helenius A (1995) Quality control in the secretory pathway. *Curr Opin Cell Biol* 7(4):523-529.

230. McPherron AC (2010) Metabolic Functions of Myostatin and Gdf11. *Immunol Endocr Metab Agents Med Chem* 10(4):217-231.

231. Higgins R*, et al.* (2015) The Unfolded Protein Response Triggers Site-Specific Regulatory Ubiquitylation of 40S Ribosomal Proteins. *Mol Cell* 59(1):35-49.

232. Luo S, Baumeister P, Yang S, Abcouwer SF, & Lee AS (2003) Induction of Grp78/BiP by translational block: activation of the Grp78 promoter by ATF4 through and upstream ATF/CRE site independent of the endoplasmic reticulum stress elements. *J Biol Chem* 278(39):37375-37385.

233. Gamer LW, Cox KA, Small C, & Rosen V (2001) Gdf11 is a negative regulator of chondrogenesis and myogenesis in the developing chick limb. *Dev Biol* 229(2):407-420.

234. Liu JP, Laufer E, & Jessell TM (2001) Assigning the positional identity of spinal motor neurons: rostrocaudal patterning of Hox-c expression by FGFs, Gdf11, and retinoids. *Neuron* 32(6):997-1012.

235. Sinha M*, et al.* (2014) Restoring systemic GDF11 levels reverses age-related dysfunction in mouse skeletal muscle. *Science* 344(6184):649-652.

236. Katsimpardi L*, et al.* (2014) Vascular and neurogenic rejuvenation of the aging mouse brain by young systemic factors. *Science* 344(6184):630-634.

237. Zhou Y*, et al.* (2016) Circulating Concentrations of Growth Differentiation Factor 11 Are Heritable and Correlate With Life Span. *J Gerontol A Biol Sci Med Sci*.

238. Poggioli T*, et al.* (2016) Circulating Growth Differentiation Factor 11/8 Levels Decline With Age. *Circ Res* 118(1):29-37.

239. Smith SC*, et al.* (2015) GDF11 does not rescue aging-related pathological hypertrophy. *Circ Res* 117(11):926-932.

240. Lander AD, Gokoffski KK, Wan FYM, Nie Q, & Calof AL (2009) Cell Lineages and the Logic of Proliferative Control. *Plos Biol* 7(1):84-100.

241. Kim J*, et al.* (2005) GDF11 controls the timing of progenitor cell competence in developing retina. *Science* 308(5730):1927-1930.

242. Wu HH*, et al.* (2003) Autoregulation of neurogenesis by GDF11. *Neuron* 37(2):197-207.

243. Worni M*, et al.* (2015) Trends in Treatment Patterns and Outcomes for Ductal Carcinoma In Situ. *Journal of the National Cancer Institute* 107(12):djv263.

244. Moss S (2005) Overdiagnosis and overtreatment of breast cancer: overdiagnosis in randomised controlled trials of breast cancer screening. *Breast cancer research : BCR* 7(5):230-234.

245. Ryser MD*, et al.* (2016) Outcomes of Active Surveillance for Ductal Carcinoma in Situ: A Computational Risk Analysis. *Journal of the National Cancer Institute* 108(5).

246. Yokoe T*, et al.* (2007) Clinical significance of growth differentiation factor 11 in colorectal cancer. *Int J Oncol* 31(5):1097-1101.

247. Taipale J & Beachy PA (2001) The Hedgehog and Wnt signalling pathways in cancer. *Nature* 411(6835):349-354.

248. Lee EY*, et al.* (1992) Mice deficient for Rb are nonviable and show defects in neurogenesis and haematopoiesis. *Nature* 359(6393):288-294.

249. Wiman KG (1993) The retinoblastoma gene: role in cell cycle control and cell differentiation. *FASEB J* 7(10):841-845.

250. Martinez JM*, et al.* (2006) Drug-induced expression of nonsteroidal anti-inflammatory drug-activated gene/macrophage inhibitory cytokine-1/prostate-derived factor, a putative tumor suppressor, inhibits tumor growth. *J Pharmacol Exp Ther* 318(2):899-906.

251. Cekanova M*, et al.* (2009) Nonsteroidal anti-inflammatory drug-activated gene-1 expression inhibits urethane-induced pulmonary tumorigenesis in transgenic mice. *Cancer prevention research* 2(5):450-458.

252. Lambert JR*, et al.* (2006) Prostate derived factor in human prostate cancer cells: gene induction by vitamin D via a p53-dependent mechanism and inhibition of prostate cancer cell growth. *J Cell Physiol* 208(3):566-574.

253. Albertoni M*, et al.* (2002) Anoxia induces macrophage inhibitory cytokine-1 (MIC-1) in glioblastoma cells independently of p53 and HIF-1. *Oncogene* 21(27):4212-4219.

254. Baek SJ*, et al.* (2006) Nonsteroidal anti-inflammatory drug-activated gene-1 over expression in transgenic mice suppresses intestinal neoplasia. *Gastroenterology* 131(5):1553-1560.

255. Amin AR*, et al.* (2015) Evasion of anti-growth signaling: A key step in tumorigenesis and potential target for treatment and prophylaxis by natural compounds. *Semin Cancer Biol* 35 Suppl:S55-77.

256. Guo W*, et al.* (2012) Slug and Sox9 cooperatively determine the mammary stem cell state. *Cell* 148(5):1015-1028.

257. Hiler D*, et al.* (2015) Quantification of Retinogenesis in 3D Cultures Reveals Epigenetic Memory and Higher Efficiency in iPSCs Derived from Rod Photoreceptors. *Cell Stem Cell* 17(1):101-115.

258. Lancaster MA & Knoblich JA (2014) Organogenesis in a dish: modeling development and disease using organoid technologies. *Science* 345(6194):1247125.

259. Mroue R & Bissell MJ (2013) Three-dimensional cultures of mouse mammary epithelial cells. *Methods Mol Biol* 945:221-250.

260. Freedman BS*, et al.* (2015) Modelling kidney disease with CRISPR-mutant kidney organoids derived from human pluripotent epiblast spheroids. *Nature communications* 6:8715.

261. Jung C, Kim C, Chae SW, & Oh S (2010) Unsupervised segmentation of overlapped nuclei using Bayesian classification. *IEEE Trans Biomed Eng* 57(12):2825-2832.

262. Bajikar SS & Janes KA (2012) Multiscale models of cell signaling. *Annals of biomedical engineering* 40(11):2319-2327.

263. Rejniak KA & Anderson AR (2011) Hybrid models of tumor growth. *Wiley Interdiscip Rev Syst Biol Med* 3(1):115-125.

264. Gottesman MM (2002) Mechanisms of cancer drug resistance. *Annu Rev Med* 53:615-627.

265. Rottenberg S*, et al.* (2012) Impact of intertumoral heterogeneity on predicting chemotherapy response of BRCA1-deficient mammary tumors. *Cancer Res* 72(9):2350-2361.

266. Dean M, Fojo T, & Bates S (2005) Tumour stem cells and drug resistance. *Nat Rev Cancer* 5(4):275-284.

267. Sarkar CA*, et al.* (2002) Rational cytokine design for increased lifetime and enhanced potency using pH-activated "histidine switching". *Nat Biotechnol* 20(9):908-913.

268. Zhang X*, et al.* (2004) Activation of the growth-differentiation factor 11 gene by the histone deacetylase (HDAC) inhibitor trichostatin A and repression by HDAC3. *Mol Cell Biol* 24(12):5106-5118.

269. Wood KC*, et al.* (2012) MicroSCALE screening reveals genetic modifiers of therapeutic response in melanoma. *Sci Signal* 5(224):rs4.

270. Leung CT & Brugge JS (2012) Outgrowth of single oncogene-expressing cells from suppressive epithelial environments. *Nature* 482(7385):410-413.

271. Marusyk A*, et al.* (2014) Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature* 514(7520):54-58.

272. He J*, et al.* (2016) Integrated genomic DNA/RNA profiling of hematologic malignancies in the clinical setting. *Blood*.

273. Janes KA, Wang CC, Holmberg KJ, Cabral K, & Brugge JS (2010) Identifying single-cell molecular programs by stochastic profiling. *Nat Methods* 7(4):311-317.

274. Moffat J*, et al.* (2006) A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* 124(6):1283-1298.

275. Shin KJ*, et al.* (2006) A single lentiviral vector platform for microRNA-based conditional RNA interference and coordinated transgene expression. *Proc Natl Acad Sci U S A* 103(37):13759-13764.

276. Oskarsson T*, et al.* (2011) Breast cancer cells produce tenascin C as a metastatic niche component to colonize the lungs. *Nat Med* 17(7):867-874.

277. Wang L, Brugge JS, & Janes KA (2011) 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* 108(40):E803-812.

278. Miller-Jensen K, Janes KA, Brugge JS, & Lauffenburger DA (2007) Common effector processing mediates cell-specific responses to stimuli. *Nature* 448(7153):604-608.

279. Peccoud J & Ycart B (1995) Markovian Modeling of Gene-Product Synthesis. *Theor Popul Biol* 48(2):222-234.

280. Raj A, Peskin CS, Tranchina D, Vargas DY, & Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 4(10):e309.

281. Fenton L (1960) The Sum of Log-Normal Probability Distributions in Scatter Transmission Systems. *Communication Systems, IRE Transactions on* 8(1):57-67.

282. Feldman RM, Valdez-Flores C, & SpringerLink (Online service) (2010) Applied probability and stochastic processes. (Springer-Verlag, Berlin ; Heidelberg).

283. Nelder JA & Mead R (1965) A Simplex Method for Function Minimization. *Comput J* 7(4):308-313.

284. Wang L & Janes KA (2013) Stochastic profiling of transcriptional regulatory heterogeneities in tissues, tumors and cultured cells. *Nat Protoc* 8(2):282-301.

285. Mi H, Poudel S, Muruganujan A, Casagrande JT, & Thomas PD (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res* 44(D1):D336-342.

286. Janes KA (2015) An analysis of critical factors for quantitative immunoblotting. *Sci Signal* 8(371):rs2.